A CRITICAL ANALYSIS OF THE TEACHING OF FUNDAMENTAL PHYSICAL CONCEPTS AND PRINCIPLES, WITH PARTICULAR REFERENCE TO SOUTH AFRICAN SCHOOLS

A thesis submitted for the degree of
Doctor of Philosophy at Rhodes University

by

Hugh Helm

November, 1969.

With the exception of certain arguments that will be recognized as "standard" by those concerned with elementary physics teaching, all parts of this thesis that I have knowingly derived from the work of others are clearly indicated as such. For the rest, this thesis is my own work.

A C K N O W L E D G E M E N T S

I am happy to place on record my sincere gratitude to a number of people who in one way or another have given me very considerable help. First I must thank my supervisor, Professor J.A. Gledhill, for suggesting that I should undertake this work and for reading and criticizing the first draft of this thesis. I have benefited immeasurably from his encouragement and constructive criticism. I wish to thank all my colleagues in the Department of Physics at Rhodes University for many fruitful discussions, and for their generosity in accepting a greater teaching load than mine during 1968. It is impossible to mention by name all the teachers and university education diploma students with whom I have discussed some of the ideas set forth in this thesis, but I am grateful to them all for helping me to shape these ideas. I am pleased to thank Mrs. A.M. van der Vyver for unflinchingly and competently coping with the formidable task of typing this thesis. Finally I must thank my wife and children for enduring with fortitude and patience the abnormal circumstances of the past few years. Indeed my wife did not merely "endure", but at all times sought to help me in whatever manner she could. I cannot find adequate words to express my gratitude for the support she has given me.

C O N T E N T S

		Page
Chapter 1:	General introduction	1
Chapter 2:	General issues related to the elementary teaching of physics	
2.1	Science teaching and educational theory	
	2.1.1 Concept formation and the pupil	5
	2.1.2 The objectives of physics teaching	8
	2.1.3 Summary	12
2.2	The problem of simplification: myths and half truths in physics teaching	12
2.3	The place of the atom, the molecule, the electron and other sub-atomic "particles" in elementary physics teaching	20
2.4	The use of "black boxes" in elementary teaching	27
2.5	The use of strict logic in the teaching of concepts and principles	37
2.6	The historical or "humanistic" approach to physics teaching	43
2.7	Physics as an "open" subject	50
2.8	The role of practical work in the elementary physics curriculum	
	2.8.1 Introduction	60
	2.8.2 The aims of practical work	61
	2.8.3 The need for practical work	70
	2.8.4 String and sealing-wax apparatus	70
2.9	The content of school physics syllabuses	
	2.9.1 Some general considerations	73
	2.9.2 Recent developments in syllabus making	74
	2.9.3 What should we exclude?	77
	2.9.4 Should technological material be included in physics syllabuses?	80
	2.9.5 Past and present syllabuses in South Africa	82

		page
Chapter 3:	Physics in the junior secondary school	
3.1	Introduction	101
3.2	The nature and importance of definitions	
	3.2.1 The role of definitions in the teaching of physics	101
	3.2.2 The nature of definitions	104
	3.2.3 Operational and constitutive definitions	105
	3.2.4 Is operationalism tenable?	109
	3.2.5 The role of operational definitions	113
	3.2.6 Definitions and the teacher	115
3.3	Physical laws and the teacher	123
3.4	Measurement and units	131
3.5	The concept of length and its units and standards	132
3.6	The concept of time and its measurement	133
3.7	Mass	
	3.7.1 Introduction	138
	3.7.2 Mass and "quantity of matter"	139
	3.7.3 How should mass be introduced?	148
	3.7.4 Mass and the teaching of chemistry	161
	3.7.5 Experimental work	163
3.8	Force	
	3.8.1 Introduction	166
	3.8.2 Introducing force	168
	3.8.3 The measurement of force	172
	3.8.4 Newton's Third Law	177
	3.8.5 Summary of the introductory treatment of force	182
	3.8.6 The "turning effect" of forces	183
3.9	Elementary notions of energy and work	
	3.9.1 Energy in present-day physics teaching	183
	3.9.2 The introductory treatment of energy	185
	3.9.3 The introduction of work	187
	3.9.4 "Derived" units, including those of work	188

		page
3.9.5	Towards conservation of energy	191
3.10	Pressure	
3.10.1	The need for the concept of pressure in elementary physics courses	194
3.10.2	The nature and definition of pressure	195
3.10.3	The measurement of pressure	197
3.10.4	Fluid pressure from the molecular point of view	201
3.11	Temperature	
3.11.1	Introduction	203
3.11.2	A critique of the introduction of temperature in terms of the kinetic theory	204
3.11.3	A suggested introduction to temperature- Stage One	207
3.11.4	The status of the "Zeroth Law"	210
3.11.5	A suggested introduction to temperature - Stages Two and Three	216
3.11.6	The "advantages" of mercury-in-glass thermometric scales	219
3.11.7	Conclusion	222
3.12	Other topics connected with "heat"	
3.12.1	Thermal expansion of solids and liquids	222
3.12.2	Heat and calorimetry	224
3.12.3	The transport of heat	233
3.13	Atoms and molecules as the basis of matter: the kinetic theory of matter	
3.13.1	Elementary physics and the particle structure of matter	237
3.13.2	Experimental evidence in support of the particle model of matter	238
3.13.3	The integration of the particle picture of matter with the teaching of the entire syllabus	245
3.14	Electricity	

		page
3.14.1	Electricity in the revised syllabuses	249
3.14.2	Electricity via electrostatics	251
3.14.3	Introductory electrostatics	253
3.15	Current electricity: elementary treatment	
3.15.1	Introduction	268
3.15.2	The introduction of the concepts of current and potential difference: some preliminary considerations	269
3.15.3	Current electricity in the classroom: where to start?	274
3.15.4	The measurement of current	279
3.15.5	Summary of requirements for a satisfactory introduction to current electricity	284
3.15.6	Elementary treatment of resistance	286
3.15.7	Potential difference	288
3.15.8	Digression: a sophisticated approach to potential difference via the water analogy	289
3.15.9	Potential difference as energy transferred per unit charge; the "voltmeter" approach	293
3.15.10	Potential difference as energy transferred per unit charge: direct approach	297
3.15.11	The "voltage" of cells	298
3.15.12	The use of analogy in teaching potential difference	300
3.15.13	The quantitative treatment of resistance	301
3.16	Elementary electromagnetism	
3.16.1	Elementary magnetism	305
3.16.2	The action of "keepers"	307
3.16.3	The magnetic effects of electric currents	312

		page
3.16.4	Electromagnetic induction	314
3.17	Wave motion	321
3.17.1	Waves in the school curriculum	321
3.17.2	Waves on springs	322
3.17.3	Waves on surfaces: the ripple tank	324
3.17.4	Graphical representation of transverse and longitudinal waves	328
3.18	Sound	
3.18.1	Sound as wave motion	329
3.18.2	The nature of sound waves	333
3.18.3	The waveforms of sound waves	337
3.18.4	Sound, the ear and hearing	339
3.18.5	Resonance	341
3.19	Light	
3.19.1	Why teach light?	342
3.19.2	Introduction to geometrical optics: P.S.S.C. and Nuffield approaches	342
3.19.3	Errors to be avoided in the teaching of geometrical optics	347
3.19.4	The laws of reflection and refraction	349
3.19.5	Treatment of the nature of light in the junior secondary school	352 end Book I

Chapter 4: Classical mechanics in the senior secondary school

4.1	Introduction	354
4.2	Vectors and scalars, with special reference to the vector nature of displacement and velocity	355
4.3	The concept of acceleration	358
4.4	Rectilinear, uniformly accelerated motion	
4.4.1	Some general considerations	359
4.4.2	Experimental work	362
4.5	Projectile motion	368
4.6	Frames of reference	369

		page
4.7	Newton's laws of motion	
	4.7.1 Newtonian mechanics and the teacher	386
	4.7.2 Newton's formulation of mechanics	388
	4.7.3 Mach's formulation of mechanics	400
	4.7.4 Critique of Mach's approach to mechanics	405
	4.7.5 Mass, force and relativity	417
	4.7.6 Other approaches to mechanics	419
	4.7.7 Attitudes to teaching Newton's Laws	422
	4.7.8 Newton's First Law	424
	4.7.9 Newton's Second Law: introduction	431
	4.7.10 Newton's Second Law: possible approaches to F = ma	433
	4.7.11 Units of mass and force	446
	4.7.12 The inertial balance	447
	4.7.13 Force as a vector	449
	4.7.14 Problems on "F = ma"	449
4.8	Momentum as a vector	450
4.9	Motion in a curved path as an introduction to circular motion; centripetal force; circular orbital motion	452
4.10	Newton's Law of Gravitation and falling bodies	459
4.11	Weight	
	4.11.1 The concept of weight	462
	4.11.2 Weight and gravitational and inertial mass	466
4.12	Newton's Third Law and Conservation of Momentum	471
4.13	Can one "derive" the Principles of Conservation of Linear and Angular Momentum?	476
4.14	Dynamics via momentum	480
4.15	Work, energy, conservation of energy and power	
	4.15.1 Background	484
	4.15.2 The teaching of the concept of energy	499
	4.15.3 Power	507

page

Chapter 5: Heat, kinetic theory and the physics of matter in bulk

(N.B. There is no page 543!)

5.1 Introduction 509

5.2 Temperature and the gas laws

5.2.1 The traditional approach 510

5.2.2 An improved approach 513

5.3 Relation between heat and work 524

5.4 The particle structure of matter

5.4.1 Introduction - intermolecular forces 527

5.4.2 Simple kinetic theory of gases 530

5.4.3 Some predictions of the kinetic theory of gases 540

5.4.4 "Heat" and temperature 544

5.4.5 Non-ideal gases: introduction 546

5.4.6 Advanced treatments of non-ideal gases 547

5.4.7 Elementary treatments of non-ideal gases 550

5.4.8 Critique of elementary treatments 553

5.4.9 Non-ideal gases and thermodynamics 555

5.4.10 The effect of the walls of the container on the behaviour of enclosed gases 560

5.4.11 Recommendations for an elementary treatment of non-ideal gases 564

5.4.12 Dalton's law of partial pressures 565

5.4.13 Conversion of "heat" to work 565

5.4.14 The solid state: crystallinity 571

5.4.15 The solid state: physical properties 577

5.4.16 The solid state: lattice energy 579

5.4.17 Changes of phase 580

5.4.18 The structure of liquids 586

5.4.19 Surface tension of liquids: introduction 591

page

5.4.20 A survey of views of surface tension 592

5.4.21 Surface energy and surface tension 597

5.4.22 The elementary teaching of surface tension 600

5.4.23 Surface tension and the floating of objects on liquid surfaces 604

Chapter 6: Vibrations and waves

6.1 Vibrations

6.1.1 Introduction 606

6.1.2 Examples of vibrating systems 607

6.1.3 Physical characteristics of vibration 608

6.1.4 The energy of vibrating systems 609

6.1.5 Forced vibration and resonance 610

6.2 Waves

6.2.1 Waves in the senior secondary school 610

6.2.2 Polarization 611

6.2.3 Principle of superposition 612

6.2.4 Diffraction 616

6.2.5 Reflection and refraction of surface waves in a ripple tank 620

6.2.6 Energy transmission by waves 623

6.2.7 Sound waves 625

Chapter 7: Electricity and magnetism in the senior secondary school

7.1 Some problems in the teaching of electricity and magnetism 629

7.2 Units 630

7.3 Some features of the proposed treatment 633

7.4 Coulomb's Law

7.4.1 Should quantitative electrostatics be based on Coulomb's Law? 634

		page
7.4.2	Introducing Coulomb's Law: (i) the inverse-square law	637
7.4.3	Introducing Coulomb's Law: (ii) dependence of the force on the magnitude of the charges.	638
7.4.4	Introducing Coulomb's Law: (iii) the proportionality constant and the coulomb	641
7.4.5	Coulomb's Law and conservation of charge: some relevant experiments.	646
7.4.6	Conservation of charge and the earthing of a charged conductor.	650
7.4.7	The Principle of Superposition	652
7.5	The electric field	
7.5.1	Introduction	652
7.5.2	Teaching the electric field concept: analogy and preliminary ideas	653
7.5.3	Can the electric field vector be defined?	654
7.5.4	The quantitative treatment of the electric field	656
7.5.5	Experimental support	658
7.5.6	Calculations	659
7.5.7	Uniform electric fields	660
7.6	Potential difference.	
7.6.1	Potential energy in gravitational and electric fields	661
7.6.2	Introduction of potential and potential difference	662
7.6.3	Potential difference and the electric field between charged parallel plates	665
7.6.4	The surface of an isolated charged conductor is equipotential	666

		page
7.6.5	The electroscope as a voltmeter	667
7.6.6	Calculations on potential difference; the choice of a potential "zero"	668
7.6.7	Further comment on the association of electrical potential energy with the electric field	669
7.7	The "graininess" of charge	
7.7.1	Introduction	672
7.7.2	Films of Millikan's experiment and the associated theory	673
7.7.3	Analogues of Millikan's experiment	676
7.7.4	Conclusions from Millikan's experiment	676
7.8	The distribution of charge on isolated conductors	677
7.9	Lightning	678
7.10	Energy in the electric field: capacitors	679
7.11	The principle of the Van de Graaff generator	680
7.12	The bridge between static and current electricity	
7.12.1	The pupil's background	682
7.12.2	An experimental approach	683
7.13	Elementary consideration of electrical conduction as a movement of charges	
7.13.1	An introductory experiment	687
7.13.2	A particle model of electrical conduction	688
7.13.3	The particle model and electrical heating	690
7.13.4	A useful analogy for electrical conduction	692
7.14	The nature of the electric current in metallic conductors, electrolytes, ionized gases and electron and positive ion beams	
7.14.1	Introduction	695
7.14.2	Metallic conductors	696
7.14.3	Electron beams	697

		page
7.14.4	Current in electrolytes	699
7.14.5	Current in gases	702
7.14.6	Positive ion beams	703
7.15	The magnetic field, treated quantitatively	
7.15.1	Reasons for a quantitative treatment of the magnetic field in a school syllabus	703
7.15.2	Should the approach to "$\vec{B}$" be "Amperian" or "Point-Pole"?	704
7.15.3	Preliminary ideas to the introduction of "$\vec{B}$"	712
7.15.4	Force between parallel current-carrying wires: definition of the ampere	714
7.15.5	Introduction of the $\vec{B}$ vector	720
7.15.6	Field of a current-carrying straight wire, flat coil and solenoid	728
7.15.7	Force on a moving charge in a magnetic field	734
7.15.8	Orbits of charged particles in magnetic fields: the charge-to-mass ratio of the electron	738
7.16	An historical approach to the electron	742
7.17	Electrical circuits - quantitative treatment	
7.17.1	Objectives	745
7.17.2	Further aspects of the particle model of electrical conduction	746
7.17.3	Potential difference in electrical circuits	748
7.17.4	EMF: the nature and definition of the concept	752
7.17.5	EMF in the elementary syllabus	763
7.17.6	Ohm's Law and resistance	767
7.18	The proportionality constant in Coulomb's Law	770
7.19	Definitions of current, resistance and resistivity	771
7.20	Electromagnetic induction	

		Page
7.20.1	Faraday's Law: "motional" and "variational" EMF's	773
7.20.2	Faraday's Law: derived or "experimental"?	780
7.20.3	Motional EMF's and "relativity"	784
7.20.4	The teaching of electromagnetic induction	792
7.20.5	Furtner discussion of "variational" or "transformer" EMF's	797
7.20.6	The field concept and electromagnetic induction	798
7.20.7	Self-induction	803
7.20.8	The dynamo and alternating current	804
7.20.9	The transformer	804
7.20.10	Alternating current circuits	805
7.20.11	The induction coil	807
7.20.12	Interchange of energy between electric and magnetic fields	808
7.21	The nature and propagation of electromagnetic waves	
7.21.1	Electromagnetic waves in the school syllabus	810
7.21.2	The association of a magnetic field with a changing electric field	811
7.21.3	The propagation of electromagnetic radiation	818
7.21.4	The generation of electromagnetic radiation	826
7.21.5	Electromagnetic waves in the laboratory	835

Ch III

Chapter 8: Light as wave motion, in the senior secondary school

8.1	The wave nature of light	838
8.2	Refraction and the particle and wave models of light	840
8.3	Wave motion and refraction by lenses	844
8.4	Diffraction of light, I	844

		page
8.5	Two-slit interference	846
8.6	Interference in thin films	850
8.7	Colour, wavelength and frequency	852
8.8	Diffraction of light, II	853
8.9	The use of the laser in demonstrating interference and diffraction	855
8.10	Polarization	857
8.11	Intensity, illumination and the inverse-square law	861
8.12	Identification of light with electromagnetic waves	862
8.13	Dispersion	867
8.14	The electromagnetic spectrum	868
8.15	Introduction to the study of spectra	868

Chapter 9: Electronics

9.1	Some general considerations	871
9.2	Thermionic emission, rectification and semiconductor diodes	873
9.3	Triodes and amplification	875
9.4	Transistors and amplification	876
9.5	Oscillators, triode and transistor	878
9.6	The cathode-ray oscilloscope	879
9.7	Modulation, transmission and reception of radio waves	882

Chapter 10: Photons, electrons, atoms and nuclei

10.1	The particle-like behaviour of light	
	10.1.1 Introduction: teaching objectives	883
	10.1.2 Introductory demonstrations	884
	10.1.3 "Problematical" aspects of the photoelectric effect	885
	10.1.4 Randomness of absorption of photons	886
	10.1.5 The photoelectric effect: dependence on frequency	892

		page
10.1.6	The nature of photons	898
10.2	The interpretation of quantum theory and its implications for elementary teaching	
10.2.1	Wave-particle duality and quantum theory	904
10.2.2	The Copenhagen interpretation	905
10.2.3	Other interpretations	910
10.2.4	The Ψ-function, "matter waves" and reality	922
10.2.5	The "Uncertainty Principle"	933
10.2.6	Implications of the foregoing discussion for the elementary teaching of quantum mechanics	940
10.2.7	A possible teaching approach to the nature of the electron	946
10.3	The atom and its nucleus	
10.3.1	Introduction	953
10.3.2	The mass of an atom	955
10.3.3	The size of an atom	958
10.3.4	The unstable atom	959
10.3.5	The discovery of the atomic nucleus	964
10.3.6	The size of the nucleus and the size of the atom	972
10.3.7	A critical survey of some possible approaches to atomic structure	976
10.3.8	A suggested teaching approach to atomic structure	982
10.3.9	The composition of the nucleus	987
10.3.10	Nuclear reactions and "nuclear energy"	991
Chapter 11:	Relativity in the school syllabus	
11.1	Introduction	1003
11.2	The difficulties presented by elementary relativity	1006

		page
11.3	The use of inertial frames	1008
11.4	The postulates of special relativity	
	11.4.1 The "Principle of Relativity"	1009
	11.4.2 The Principle of the Constancy of the Velocity of Light	1011
	11.4.3 Relativity and history	1020
11.5	Problem-solving in elementary relativity	1021
11.6	The teaching of relativistic kinematics	
	11.6.1 The relative velocity of inertial frames	1023
	11.6.2 Should one start with the Lorentz transformation?	1024
	11.6.3 Derivation of some of the standard results directly from the fundamental postulates	1026
	11.6.4 Bondi's K-calculus	1032
	11.6.5 Experimental evidence for relativistic kinematics	1037
11.7	Relativistic dynamics	
	11.7.1 Mass and momentum: thought experiments	1041
	11.7.2 Mass and momentum: empirical approach	1045
	11.7.3 Energy	1047
	11.7.4 Relativistic momentum and Bertozzi's experiment	1052
	11.7.5 Relativistic momentum: experimental confirmation	1053
	11.7.6 Photon mechanics	1056
	11.7.7 Which approach to relativistic dynamics should one use?	1061
11.8	Special relativity via mechanics?	1062
11.9	Conclusions	1064
Chapter 12:	Summary and conclusions	1069
References		1073

ERRATA

Page 13, line 23: the reference should read: "Joint Committee of the South African Chemical Institute and Institute of Physics 1965".

Page 60, line 14 below heading: the reference should read: "Department of Education, Cape of Good Hope 1961".

CHAPTER 1: GENERAL INTRODUCTION

The author's interest in the subject-matter of this thesis dates back to his own schooldays, when he was an interested but frustrated science pupil. In more recent years his professional duties as a university lecturer, physics method instructor for University Education Diploma students, physics examiner for the Joint Matriculation Board, co-drafter of a new school physics syllabus and part author of teachers' guides for new science syllabuses (Helm 1963, Prinsloo et al 1968) and school textbooks (Helm et al 1968, 1969a, 1969b) have compelled him to take a much more active interest, which has helped him to understand his early frustration. It has also strengthened his conviction that any attempt to teach fundamental physical concepts and principles at an elementary level should be based on a very careful study of those concepts and principles.

The main purpose of the present study is to give a detailed critical survey of possible teaching approaches to fundamental physical concepts and principles which could reasonably be taught at school, and to consider the implications of this survey for the teaching of physical science in South Africa. In general, this study will not deal with those aspects of the problem which are more properly the concern of specialists in educational theory and method. Thus little or nothing will be said in the following pages about the details of classroom technique in the presentation of the topics under discussion. Instead, the author will seek to establish those points which, in his opinion, must be covered in any proper presentation of these topics.

A survey of this kind on the scale of the present work does not appear to have been attempted before. The nearest equivalent to it in English seems to be Warren's book, "The

/Teaching.....

Teaching of Physics" (1965), which is mainly concerned with exposing some of the more serious errors and faulty traditions in the teaching of physics in Great Britain. Dr. Warren makes a number of valuable points, but the terseness with which he deals with many of them may leave the unsophisticated reader with a sense of bewilderment or frustration. The correct approaches to particular topics or sub-topics are often only briefly indicated, and no real attempt is made to come to grips with the underlying philosophical issues. Some of these issues are discussed, sometimes at some length, in the excellent Teachers' Guides to the PSSC and Nuffield Physics courses (PSSC 1960, Nuffield Guides I - V 1966), but it is not to be expected of such works that they should devote too much space to arguing points of principle: their main concern is with giving aid to the teacher in presenting these courses.

That a study of the present kind might prove useful is suggested by some evidence recently obtained in the course of an enquiry conducted by the British Committee of Chemical Education into the teaching training of chemistry graduates (cf Elton 1968). Practising teachers completing a questionnaire issued by the Committee ranked "guidance related to the teaching of one's subject" as the most important aspect of a teachers' training course. It is not really surprising that practising teachers should think this, since it is in trying to teach his subject at an elementary level that the conscientious science teacher becomes most strongly aware of the many difficulties of presentation which confront him and of the extent of his ability to cope with them. There can be little doubt that in South Africa at present, the subject training provided for intending science teachers is inadequate. At certain South African universities, methods of coping with this problem are

/currently.....

currently under review. In the absence of foreknowledge of what these deliberations may produce, we would strongly suggest that an important part of successful subject training must consist of a careful study of the content of the subject, with especial reference to its underlying logic and philosophy. In fact, we think that the intending teacher should be exposed, at least in a small way, to the kind of analysis of fundamental concepts and principles that will be attempted in the main body of this thesis. He will then have had the opportunity of meeting and thinking about some of the crucial problems in the teaching of his subject before he meets them in the classroom.

In what follows, analysis of fundamental physical concepts and principles and teaching approaches to them (Chapters 3 to 11) will be preceded (Chapter 2) by a discussion of a number of important general issues related to physics teaching. The attitude of the present writer towards his task is inevitably coloured by his views on these general issues, so it is important that these should be made known at the outset. First of the issues to be discussed is the relationship between physics teaching and educational theory. This is taken first because it can be said to be fundamental to the whole enterprise. If, that is to say, the teaching approaches one favours towards topics in physics are in serious conflict with educational theory and the research which supports it, then the foundations of one's thinking are clearly insecure. The remaining general issues need not be listed in detail here. The order of their presentation has been determined not by their relative importance, which would be difficult to assess, but largely by the convenience of treating some of these topics in close succession. Thus the first four are fairly closely related to one another, all being in some way concerned with the underlying logic of the subject. The next two, while not as

/closely.....

closely related as the first four, nevertheless can be said to belong together to some extent, since both relate to the need to present the subject in a stimulating way. An essay on the role of practical work comes next, and finally in this chapter we include a survey of the physics syllabuses in use in South Africa prior to the introduction of new Common Basic Syllabuses, and a discussion of the content of the latter. It is fitting that this should immediately precede the main portion of this thesis, since the latter is presented against the background of the newly-introduced syllabuses.

CHAPTER 2: GENERAL ISSUES RELATED TO THE ELEMENTARY TEACHING OF PHYSICS.

2.1 Science teaching and educational theory

2.1.1 Concept formation and the pupil

It seems fair to say that some, at least, of the recent developments in curriculum reform in school science have taken place independently of educational theories or of the discoveries of educational psychology. Certainly this has been the case in South Africa, where the planning of syllabuses in senior school science has been left to specialists in education rather than Education. As one who was concerned in the revision of the physics syllabus for the South African Joint Matriculation Board, the present author knows that decidions concerning the content of this syllabus were guided to a large extent by the feeling that the existing syllabus was inadequate because of its lack of contact with the world of the present-day physicist. It was also felt that careful planning of the new syllabus might lead to improved approaches to the teaching of traditional material. There was no conscious attempt to use the results of research in educational psychology on what a pupil could safely be expected to absorb at a particular stage of his development. Instead, it was assumed that the accumulated experience of the teaching of certain topics at school and university had established that these topics could be "understood" at an elementary level (in so far as any topic could be said to be "understood" at any level), so that the introduction of topics of supposedly comparable difficulty should not cause serious embarrassment to pupils or teachers.

Justification for proceeding in this way is that similar methods have achieved some success in the past - pupils have managed to learn some physics at school, and it seems reasonable to argue that the less confusingly the subject is taught,

/the.....

the more use it will be to pupils later on, particularly if they are to continue their studies. But so long as relevant psychological research is ignored or not taken explicitly into account, there exists the danger that one may not be taking full advantage of the ability of the child to learn concepts at a particular age, or that one may be overestimating that ability. As far as the present study is concerned, there would be little point in elucidating the nature of certain fundamental concepts and outlining the way in which they might be taught at school, if in fact it were found that at school-going age, the pupil could not be expected to absorb these concepts to the required extent.

The classic work on the child's ability to form and handle mathematical and scientific concepts is that of Piaget and his co-workers (cf Lovell 1962 and the references given therein). According to Piaget, the intellectual development of the child proceeds through a sequence of well-defined stages, viz. sensi-motor intelligence (birth - 2 years), intuitive thought (4 - 7 years), concrete operations (7 - 11 years) and formal operations (11 - 15 years). (This summary is given by Wilkinson (1965); according to Lovell (1968a) "In Piaget's developmental system, formal operational thought sets in around 14 - 15 years of age in ordinary pupils although earlier in the very able pupils.") It is in the last-mentioned stage that the child is able to think about concepts in something like the way a scientist does and to understand "higher-order" concepts such as those of momentum, work, specific heat, electric intensity and potential difference. Therefore it is only when the pupil has reached the stage of "formal operational thought" that it becomes feasible to teach him science along the lines ordinarily laid down in science syllabuses. At present, it is accepted in South Africa that this sort of teaching will begin in Standard Six.

/Ordinarily.....

Ordinarily a pupil in Standard Six will either turn thirteen in the first half of the year or fourteen in the second. Younger pupils will also be found in Standard Six, but they will be among the brighter ones. Accordingly, if Piaget's ideas are correct, South African pupils begin their more formal study of science at approximately the right age. To this extent we may feel confident that the planned syllabuses are not attempting the impossible, at least as far as the introduction of most fundamental and derived concepts is concerned. In respect of fundamental principles or laws such as those of the conservation of momentum and energy, there is also encouraging evidence from the psychologists. Experiments on the ability of children to grasp conservation of various quantities such as volume and weight have shown that while the idea of a conserved quantity takes some time to develop, it is quite well established by the time the child has reached the age of eleven (cf Lovell 1962).

One might be tempted to conclude from the above account that the findings of the educational psychologists merely confirm what experience and common sense have already shown to be true, and that there is thus no real need to take them into account. However, it may well turn out that when it is proposed to introduce new topics into a syllabus or to try a new method of presentation of traditional material, the psychologists may have useful advice to give. Certainly it would have been useful to know in advance whether the available evidence suggested that the teaching of, say, the mechanism of propagation of electromagnetic waves would be a practical proposition at school. At the moment, relevant evidence of the kind that might pass muster as being "scientific" does not seem to be at hand. Lovell (1968b) has stressed that in the case of this topic, the pupil must be familiar with a formidable array of concepts and principles in order to grasp what is being

/taught.....

taught, but he has not gone so far as to condemn the inclusion of the topic in a school syllabus. In general, the present position appears to be that the practicability or otherwise of introducing new and possibly difficult topics into school syllabuses can only be established by the degree of "success" which greets the teacher's efforts in the classroom. This is clearly unsatisfactory (cf. Skinner's remarks below), and it is to be hoped that controlled experiments in this connection will soon be undertaken. In the meantime there is, we suggest, no need for excessive timidity on the part of the planners of syllabuses. By all means let them try to include new material in a syllabus if this does not appear to represent too great an advance in "difficulty" on what has previously been taught. Some teachers at least would welcome the challenge that this provides, and their pupils might welcome it too if they felt that it could plug a hole in their knowledge that desperately needed plugging. Until the psychological researchers can offer more definite information in this respect, there would seem to be no harm in exploring new territory in the classroom, provided that this territory is kept within bounds generally conceded to be reasonable.

2.1.2 The objectives of physics teaching

Another matter of fundamental importance to which educational theorists are wont to draw attention and educationalists to discuss, is the question of the objectives of a course of instruction and the extent to which these can be or are being realized. No doubt it is true that many of those concerned with teaching science at school or university have not always been careful to formulate their aims clearly and to consider whether or not their own teaching is properly directed towards those aims; perhaps, too, not all the recent attempts at

/curriculum.....

curriculum reform have been as soundly motivated in this respect as they might have been. In any event, educational theorists and psychologists have from time to time reminded science teachers of their obligations in this respect, though they have seldom done so in terms quite as forthright as those recently used by B.F. Skinner (reported by Schwartz 1968). Skinner has condemned science teaching in general for its reliance on "aphorisms and folk wisdom". He suggests that "classroom observation.... is probably one of the poorest settings and situations in which to conduct an effective analysis of teacher practices since the significant consequences of teacher behaviour, namely student behaviour many years after he has had contact with his teacher, cannot be discerned in the classroom and hence cannot be correlated with teacher practices." And he maintains that "The key step in improving the status of science education is discarding the high sounding but meaningless concepts that we have been relying on to describe what is to be taught and how it is to be learned - such notions as intellectual structure and concept formation - and to emphasize instead concrete, definable behaviours and the processes which can be used to shape and control these behaviours." Thus it is important, in Skinner's view, to specify the desired "terminal behaviours" expected of pupils in order to shape and control them.

Perhaps there is a need for reform of present approaches to science teaching along the lines Skinner indicates. But such a reform will, we think, be mainly concerned with techniques of instruction - what Skinner calls "schedules of reinforcement" - rather than with the content of syllabuses, although the latter may conceivably be affected by the former. Clearly, no matter how one may choose to reform one's "schedules of reinforcement", one still has to teach a <u>subject</u>; and it will

/remain.....

remain important, we suggest, that the subject-matter should be taught in a way which respects the fundamentals of the subject, as well as being in accordance with what Wilkinson (1965) calls "sound educational theory."

What the desired "terminal behaviours" are, Skinner does not discuss. Thus one does not know whether or not he would approve as acceptable "behavioural objectives" those listed by Nedelsky (1965), viz. knowledge, understanding and ability to learn. But it is likely that many science teachers would now acknowledge these as worthwhile objectives. Conferences of science teachers tend nowadays to affirm objectives for their teaching which amount in essence to the above three (e.g. American Association of Physics Teachers 1957, Michaelhouse 1968), with particular emphasis on "understanding;" it is not too much to say that "teach for understanding" has now become a rallying cry at such conferences. Likewise, individual teachers who discuss the problem in detail arrive at much the same formulation of aims (e.g. Rogers 1962).

One can argue that the above behavioural objectives are equally well suited to any subject, and that the physics teacher ought to specify in detail what he intends under the headings of knowledge, understanding and ability to learn. Nedelsky has in fact done this: under "knowledge" he lists possession of information about laws and principles, theories, facts, definitions and even history; knowledge of relations between laws of nature and specific phenomena, theories and phenomena, experiments and conclusions, and broad concepts and classifications and specifics, and laboratory knowledge; under "understanding" he lists understanding the individual topics and principles taught in the course, understanding the interrelationships of physics principles and their relationships to principles of other disciplines, understanding the nature and structure of physics, intuitive understanding and laboratory understanding; and under "ability to learn" he lists the ability

/to.....

to understand prose, the ability to understand non-verbal symbolism, disciplined thinking, imaginative thinking and ingenuity, knowledge of sources of information, and ability to learn from experiment or observation. Again, teachers would probably agree that this detailed list of objectives covers most or all of what they would like to see achieved in any introductory physics course. Some might be inclined to observe somewhat cynically that the primary "behavioural objective" for most teachers (and parents and pupils), viz. ability to get through the appropriate examination, has been omitted from the above list. In fact, this objective has been implicitly included, because any physics examination, however carelessly it may be set, does in effect test some or other of the objectives listed above.

If these are valid objectives for an introductory physics course, then they add up to a powerful argument in support of the present undertaking. It is difficult to see how a teacher can contribute positively toward the attainment of these objectives if his own knowledge and understanding of the subject and ability to learn more of it are seriously deficient. If he is content to present the garbled version of the facts on which he may have been brought up, then the best that can be hoped for is that the pupil will come to realize that his "live" source of information is unreliable and turn elsewhere for guidance. If the teacher has a poor appreciation of the nature and structure of physics, and if his "understanding (of) the individual topics and principles in the course" is deficient because of failure to probe these matters deeply enough, or clouded by misconceptions derived from faulty traditions, then one cannot hope that he will be able successfully to "teach for understanding". To achieve this and the other objectives, he needs ideally to study the kind of analysis of approaches to fundamental

/physical.....

physical concepts and principles which is attempted in the main body of this thesis.

2.1.3 Summary

So far as can be assessed at present, it would appear that the newer physics syllabuses do not seriously exceed the capabilities of a reasonably intelligent pupil in the senior school. More research is undoubtedly needed into the ability of pupils to cope with such topics as electromagnetic wave propagation, the ideas of the quantum theory or even special relativity, but until the results of such research are available, we do not think that attempts to teach these topics in the classroom should be barred. Although the psychologists may exhort teachers to reform their approaches to and outlook on physics teaching, it will still be necessary to teach some physics; and if the teaching of physics is to be taken seriously, teachers will have to be concerned about the underlying philosophy of their approaches to the content of the subject. The generally-agreed objectives of the teaching of physics (or of any science subject) are the attainment by the pupil of knowledge and understanding of the subject and ability to learn more of it. For the teacher to guide the pupil towards the attainment of these objectives, he must be properly equipped. For this reason, a study of the present kind, which is primarily aimed at analysing and clarifying the underlying logic and philosophy of the presentation of fundamental physical concepts and principles, can be claimed to be relevant to the needs of the present-day teacher of physics.

2.2 The problem of simplification: myths and half-truths in physics teaching

A fundamental problem in physics teaching is how to make difficult ideas intelligible to beginners. To the best of the

/author's .....

author's knowledge, there exists as yet no accepted general "theory of simplification" which can be used as a guide to the elementary teaching of a subject. Instead, supposedly satisfactory elementary approaches to particular topics appear to have evolved by a process of trial and error, and these approaches have come to be accepted as part of the tradition - one might almost say the dogma - of physics teaching. Voices of criticism have occasionally been raised against this tradition, but for a long time they seemed to have little effect on its principal guardians, viz. the examining authorities, the teachers and the authors of elementary textbooks. In recent years, however, the critical voices have become stronger, and certain authorities responsible for planning new physics courses have made known their desire for increased rigour in physics teaching. For example, persons connected with the Physical Science Study Committee course in physics have been heard to say that it is the intention of the course to present the subject in such a way that nothing learnt by the pupil would have to be unlearnt later; and the committee responsible for drafting the new Common Basic Syllabuses for physics, physical science and chemistry for South African pupils state in their preamble that "What is taught at high-school level must unavoidably be incomplete; it should, however, never be wrong" (Committee of the South African Chemical Institute and Institute of Physics 1965). Warren (1961,1965) has mounted well-aimed attacks on specific errors in textbooks and in the tradition of physics teaching, and the recently-established journal "Physics Education" has followed this lead by running a feature devoted to the criticism of examination questions called "Why is this a bad question?" and by publishing articles and correspondence concerned with the correction of errors in physics teaching.

As might have been expected, this tendency towards increased rigour has been welcomed by professional physicists,

/but.....

but it has elicited an equally predictable reaction from experienced teacners, who have insisted that some sacrifice of rigour is essential if difficult concepts are to be successfully taught at school level. As examples of this attitude we may take the reactions of Ruchlis (1964), who believes that "...we should select the maximum of precision consistent with the goal of achieving comprehension of the concepts we seek to teach", and Jardine (1964), who writes as follows in the preface to his series of books, 'Physics is Fun" : "To say 'you must simplify without falsifying' is unrealistic. Every attempt to simplify falsifies to some extent. Let us accept this and go on to eliminate the falsification progressively." Tomer (1966) has expressed views closely similar to these. He advocates a "combination of oversimplification with removable limitations" and goes on to stress that "the ability, willingness, even eagerness to unlearn is of extreme importance in the world today." One can readily sympathise with these views, since the difficulties of making concepts comprehensible at any level are not to be minimized. Unfortunately, there are clear dangers in the uncritical adoption of these views. It would be all too easy for the lazy or incompetent teacher or examiner to take a much too conservative view of "the maximum of precision consistent with the goal of achieving comprehension." Likewise, a policy of "oversimplification with removable limitations" might pay dividends if applied by the right hands, but in the wrong hands the "limitations" might never be removed, even if their existence was recognized in the first place.

At any rate, let us grant for the moment that Ruchlis, Jardine and Tomer have something of a case, and see whether their attitudes are in any degree reconcilable with the pursuit of increased rigour. In fact they are, since it is obviously

/impossible.....

impossible to present the whole truth (if there is such a thing) about any given concept within the confines of a school physics course. Is it possible to present nothing but the truth? Tomer's use of the word oversimplification and Jardine's reference to "falsification" would seem to imply that it is not possible, and perhaps here too one can agree. But it should surely be the aim of any conscientious teacher to tell as much of the truth as possible at all times, to be constantly aware of the "whole" truth and to regard it as his goal. If he does this, he will indeed be satisfactorily reconciling the requirements of the classroom with the pursuit of rigour.

The question of oversimplification and its possible consequences is so important that it must be given further attention here. Tomer takes as an example of excusable oversimplification the use of the concept (if such it can be called) of "electrical pressure" for voltage. He says: "Before a student can understand energy he can be taught that voltage is electrical pressure and this will enable him to feel right about the flow of current [sic] in metallic circuits." There is something to be said for this: the concept of "electrical pressure" is not "wrong", it simply does not mean very much, and if it is used very carefully it can be made to serve as a satisfactory basis for the understanding of currents in circuits. But what if the teacher has been so conditioned that he does not think of improving on the teaching of potential difference as electrical pressure, or if the requirements of the syllabus seem to prevent him from going beyond this? The author well remembers the frustration he experienced when being taught this part of the school syllabus. The fundamental concepts associated with electric currents were loosely "explained" by way of the water analogy, and no attempt whatever was made at giving an account of these concepts in electrical terms. Nor did the approved

/textbook.....

textbook provide any help. Its effectiveness as a teaching aid in this connection may be judged from its "definition" of E M F :- (Gordon and Neser, no date) "The E M F of any generator is the pressure it has available for causing a current".

It is easy to cite other examples of oversimplifications that have long been sanctified by tradition but which can easily lead to confusion or total lack of comprehension. Consider the introduction of the concept of mass as "quantity of matter". As a means of starting the pupil off on the study of a subtle and elusive concept this may perhaps be excused; we are, after all, conditioned to the practice of paying for commodities according to their "weight", and the term "quantity of matter" thus provides some link with experience. But to <u>define</u> mass in this way and to leave it at that, as so many of our text-books have been content to do, is inexcusable. The pupil who is left with the impression that this is all that there is to be said about mass has gained a totally false view of the situation.

One may reasonably complain, too, about the properties with which magnetic lines of force are occasionally endowed in order to explain certain electromagnetic phenomena. Pupils are encouraged to think of lines of force as "tending to shorten" and "tending to repel one another sideways" (cf. Gordon and Neser, op.cit.), and also as entities capable of jumping into and out of a coil in which the current is switched off or on.
Now it is true that Faraday thought of lines of force in this way (cf Tricker 1966), and they were undoubtedly of great help to <u>him</u>. But our understanding of electromagnetism has deepened since Faraday's time, and it is at least questionable whether one should invite pupils uncritically to accept such statements as the following:-

/(a).....

(a) "The lines of force are forced to crowd together within this space. Since lines of force tend to repel one another, this crowding tends to cause the wires to repel each other." (An explanation of the mutual repulsion of parallel wires carrying oppositely-directed currents.)

(b) "In tending to shorten and repel one another sideways these lines of force drive the wire forward as if it were an arrow impelled by a bowstring under stress." (An explanation of the force on a current-carrying conductor in a magnetic field.)

(c) "The alternating current in the primary coil produces a varying magnetic field whose lines of force cut across the windings of the secondary coil to induce an E M F in it." (An explanation of the induction of an E M F in the secondary coil of an alternating current transformer.)

[The above quotations are taken from Gordon and Neser, op. cit.]

No attempt is made in the context in which these quotations appear to stress that lines of force exist only in the pupil's imagination or on the paper on which he draws them. Thus instead of being taught to regard lines of force patterns simply as helpful pictorial representations of these phenomena, the pupil is misled into supposing that they explain what is happening. (Of course, (c) above is not even correct as a pictorial description; an E M F would be induced in the secondary coil even if no part of it were in a magnetic field, so that to talk in this connection of "lines of force cutting across the windings of the secondary" is to mislead the pupil inexcusably.) Compare this usage of lines of force with what one encounters in, say, Rogers's "Physics for the Inquiring Mind" (Rogers 1960) or in the Nuffield Physics Project Course (Nuffield Guide I). (Note: The Nuffield Teachers' Guides will in future be referred to in the following way:

Nuffield Guide I ≡ Nuffield Physics Teachers' Guide I, 1966)

In these courses extensive use is made of lines of force to

/aid.....

aid visualization and comprehension; for example, the term "catapult force" is used to characterize the force on a current-carrying conductor in a magnetic field (Rogers op. cit. and Nuffield Guide (III)), and this is surely of great help to anyone in remembering the salient features of this force. But students of these courses are not allowed to gain the idea that lines of force are "real"; thus on page 572 of "Physics of the Inquiring Mind" the reader is firmly told that "Neither set of lines [electric or magnetic] is really there, of course."

The attribution of fictitious physical properties to lines of force may charitably be placed in the category of "possibly useful myths." How is one to categorize the traditional definition of a magnetic line of force as "the path along which a north-seeking pole tends to move in a magnetic field"? (Gordon and Neser, op. cit.) Here is a statement which has the weight of a long-established tradition benind it; even Rogers (1960) gives what is essentially this definition, although he qualifies it in the next sentence by saying "...if we have doubts about obtaining a test pole all by itself, we define them as lines along which a tiny test compass needle would point". There is no doubt that this definition is mythical; no-one has ever found a free magnetic pole (cf. Section 7.15.2) and even if such a thing did exist, it would not follow a line of force unless it were subject to certain special conditions. Yet this definition is still enthusiastically taught at various levels; with the introduction of magnetism into the syllabus for primary schools in this country, children now meet it quite early in their school careers. Is there really anything to be gained by the use of such a mythical definition? Surely there is no point whatever in dragging in the fictitious concept of a free magnetic pole when

/the.....

the pupil is already familiar with the compass needle and the way it behaves. One is forced to conclude that the definition of a magnetic line of force in terms of free north poles must be put in the category of "pointless myths."

The examples given above of myths current in physics teaching form only a small part of those known to be in common use. The available evidence suggests that in many cases these myths are used quite uncritically and without any awareness that they are myths. In the face of this situation it is surely a tactical error to place as much stress on the value of oversimplification in elementary physics teaching, as Jardine and Tomer would have us do. We should rather insist that while oversimplification is certainly necessary, it must be used with a proper sense of responsibility if its effects are not ultimately to be harmful. In the words of Noakes (1969), "... making things easy is a difficult job, which cannot be done unless you approach it from above." This will be the standpoint adopted in the detailed critical survey of the teaching of fundamental concepts to be found later in this thesis.

It remains to consider Tomer's argument concerning "the ability, willingness, and even eagerness to unlearn." He poses the question: "In science classes should we not stamp out the idea that success depends on learning, for once and for all, the 'correct' models?" One can unhesitatingly answer this question in the affirmative, and still feel bound to point out that pupils will derive little benefit and possibly a good deal of harm from being taught ideas that are needlessly incorrect. The pupil who enters university with an idea of potential difference based on an insecurely defined "electrical pressure" and is then given the true story by his physics lecturer, is not likely to feel grateful for having had to learn something at school which he must now unlearn. For him, the only retrospective value of the experience may well be that the

/suspicions.....

suspicions he had entertained about his teacher and his text-book have now turned out to be well founded! There is a great deal of difference between this sort of thing and the careful study of various models as alternative explanations for a particular physical phenomenon. Whereas the latter forms an essential part of the study of physics, the former is an inexcusable waste of the pupil's time. One says "inexcusable" because not only could the time have been more profitably spent, but the misconceptions acquired in this way might prove very difficult to remove later on.

The conclusion seems inescapable that current efforts to reduce the amount of "unlearning" by the rooting out of errors and faulty traditions and the intelligent planning of courses deserve all possible support. In the following pages, attention will be drawn wherever necessary to the existence of such errors and faulty traditions and alternative suggestions for improved presentation will be offered.

2. 3 The place of the atom, the molecule, the electron and other sub-atomic "particles" in elementary physics teaching

In the previous section, we noted the emergence of a tendency towards increased rigour in elementary physics teaching. Another feature of contemporary physics teaching is the tendency to include more and more "modern" material in elementary courses in order to add life and interest to them. In this section, we examine a particular form of this "modernization", viz. the early use in an elementary physics course of concepts specifically associated with "modern" physics.

One of the papers distributed among delegates to the International Conference on Physics Education in Paris (1960) was entitled "The Teaching of Physics in Schools" (Brown and Clarke 1960). This was the report of a special committee of European physicists under the chairmanship of Mr. Norman Clarke, whose work had been sponsored by the Organisation for European

/Economic.....

Economic Co-operation. Much of the report is taken up with the outline of proposed new school physics courses and the principles upon which these should be based. One of the committee's recommendations in this connection has attracted so much attention that it must be quoted in full here:

"We have recommended that the introduction to physics should be through a consideration of atoms, molecules, and electrons. We do this for two main reasons. First, most children these days are aware of the terms and have at least a vague idea that matter consists of these entities. A teacher is therefore able to begin the course by building upon the children's interest and by dealing with a branch of the subject which also has the virtue of being able to provide a good deal of stimulation. Second, and more important, the teacher should start from the very beginning to present physics in as modern terms as possible. Children should obtain an early grasp of the idea that the physical world can be interpreted in terms of moving particles and of wave motion."

This kind of proposal is not new. Compare the following statement, made by no less a person than Ernest Rutherford (1923) nearly forty years earlier:

"It seems to me important that new conceptions and results that are generally accepted by scientific men should be incorporated at once in elementary teaching, provided that they simplify current explanations or assist the formation of a clearer picture of natural processes.....We are quite certain that the existence of the atom is no longer a hypothesis but a verified fact, and we are confident that tne masses of the individual atoms and the number of atoms in a known weight of material are known with errors which are certainly less than 1 percent. In teaching both physics and chemistry, there seems to be no

/objection.....

objection to presenting atoms at an early stage as real constituents of matter about whose concrete existence there is no longer room to doubt."

The proposal of the OEEC committee thus has eminent backing. Implementation of it is one way of ensuring that pupils will encounter something of recent (i.e. post-1895) developments in physics. Is this the best way? The author's experience leads him to believe that a number of science teachers, some of whom will never have heard of the OEEC report, or Rutherford's statement, think that it is the best way, for reasons that are very much the same as those quoted above. Are the pupils of these teachers receiving the right kind of introduction to physics?

In attempting to answer that question, let us consider carefully what kind of introduction this is. It would seem that the pupils must take the <u>existence</u> of atoms, molecules and electrons for granted from the start. As the course develops, more and more of the nature of these entities will be revealed, but their existence is to be assumed throughout. In other words, the pupil is apparently to be denied the opportunity (and, it might be added, the fun) of following the arguments and observing some of the supporting experimental evidence in favour of the existence of these entities. This seems a very serious limitation. When a pupil embarks upon the study of science, he really has no more reason for believing in atoms, molecules and electrons than he has for believing in fairies, except that he will have gathered by this time that his elders and betters don't believe in fairies whereas they do believe in atoms. It is on this basis - i.e. on "half appreciated modern concepts floating without visible means of support" (Tricker 1962)- that his serious study of physics is expected to rest. This kind of situation has been well dramatised by Rogers (1968a) who, in a lecture entitled "Why should

/People.....

People Believe What Scientists Tell Them?" , began by spending several minutes in lucidly summarizing the known properties of molecules and atoms before announcing "That's not science - that's gossip! ".

When looked at in this light, the case for the OEEC recommendation seems so shaky that we feel almost compelled to try to re-present it in terms rather more persuasive than those used by the OEEC. Such an attempt might run along these lines: "In teaching physics to our pupils, we have to get through a lot of material most of which won't seem very exciting to our pupils. If they have to wait until we have marshalled all the evidence in favour of the existence of atoms, molecules and electrons before they can see these concepts being used to explain various phenomena, their interest in the subject may already have been killed. As the late Professor Francis L. Friedman put it: "Every professional science book rests on the philosophy of starting with the known and going to the unknown. And this is the surest way to bore the children to death. Start with the unknown and work to the known" (Brown, Clarke and Tiomno 1964). So let us tell our pupils something about atoms, molecules and electrons right at the beginning; then we can use these concepts to help brighten up the routine physics. And let us not worry too much about presenting things in a strictly logical order: life is short and the time available in school minimal, and we shall inevitably have to do a lot of corner-cutting and dogmatic presentation in order to get through the syllabus, so we may as well be content to let the particle structure of matter form part of the dogma."

Perhaps the argument sounds a little better when presented this way; in any event, it does disclose a basis for a compromise in which the advantages of the "tell them about atoms at the beginning" approach can be enjoyed without any real sacrifice of the principles lying behind a strictly logical approach.

/In.....

In such a compromise approach, pupils would still be given some account of the particle structure of matter at the start of their course, but this time it would not be as a slice of dogma waiting to be swallowed. Instead the particle picture of matter would be treated as a plausible <u>model</u> of matter, the details of which are subject to verification and elaboration by experiment. The model can be used to interpret and "explain" phenomena and to serve as the basis for a detailed theory whose predictions can be checked; shortcomings in the explanations based upon it can be dealt with by modifications in the model. In this way the pupils do have physics presented to them in modern terms without being cheated of the intellectual excitement that can accompany the unravelling of the truth about the nature of things. This is essentially the approach followed by the Nuffield Physics Project: as stated in their Guide (Nuffield Guide I), "... we shall treat atoms as familiar things from the beginning and encourage children to learn more about them; but we shall maintain a warning flag by asking again and again in the early years: 'How do you know there are atoms? How do you know how small they are?... Did that experiment tell you for certain, or did you have to make a risky guess (assumption) in it?'."

The celebrated PSSC course is also planned along these lines (White 1960, PSSC 1960). Quite early in the course, before the pupil encounters any formal study of mechanics, he is confronted with the atomic theory of matter. However, this section of the course is no mere recital of facts; some evidence in favour of the atomic view is presented and the pupil is also given some of the chemical background. Later, after having made some acquaintance with the fundamentals of electricity, the pupil hears plausible arguments, based on experiment, for the existence of electrons and studies the behaviour of

/electrons.....

electrons and positive ions in electric fields. He is now well prepared for the study of current electricity as the motion of charge carriers such as electrons and ions. In the last section of the course he ventures on a much deeper exploration of the structure of matter, again based on experiment, in which further details of the structure of the atom are revealed, as is the so-called "duality" of radiation and matter.

In both the Nuffield Physics Project and the more ambitious, higher-level PSSC physics course, the integration of the use of modern concepts with a presentation of the subject that is essentially logical has been successfully carried out. One might quarrel with details of either project (e.g. the use of the electronic charge as the fundamental unit of charge in the PSSC course - see Tricker 1962), but minor difficulties of this sort do not obscure one's admiration for the whole. That these courses represent major departures from traditional approaches to physics teaching is not in itself of consequence. If there are any dogmas to be accepted in physics teaching, one of them must surely be that there is no One True Method of presenting the subject. Unfortunately, there have been some who have sought to follow the lead of PSSC physics (which was first in the field) without paying proper regard to the underlying philosophy of that course. The discussion of heat and temperature by Stollberg and Hill in their recently-published textbook (Stollberg and Hill 1965) may be used as an illustration, since it is typical of the approach to this topic adopted by a number of new textbooks. After a rapid review of molecules and intermolecular forces, and a treatment of some of the properties of solids, liquids and gases in terms of the "kinetic-molecular" theory, they move on to a discussion of temperature and heat in which the conclusion is arrived at that the temperature of a substance depends on the kinetic energy of its molecules.

/In .....

In the next paragraph, the idea of the absolute zero of temperature is introduced. The difference between temperature and internal energy is then explained, followed shortly by an account of the different modes of heat transfer. Here the authors really rush ahead: in quick succession the reader is told that heat radiation is produced by the vibrations of atoms and molecules, that their electrons are displaced, sending out waves which belong to a family called electromagnetic waves - all this and a good deal more. On the whole, this breathless exposition is accurate; but one cannot help feeling that its impact is somewhat reduced by the fact that it precedes the treatment given in the book of the measurement of temperature (which follows traditional lines and in so doing commits some of the traditional blunders) and of waves. We venture to doubt whether this kind of thing was really what the OEEC committee had in mind when advocating the presentation of physics in modern terms.

In South Africa, the report of the OEEC committee has been noted and the inclusion in the Common Basic Syllabuses for Physics and Physical Science of a section entitled "The Particle Structure of Matter" is an avowed attempt by the drafters of these syllabuses to give effect to the recommendations of this report. Indeed, in the detailed commentary accompanying the first draft of the Common Basic Syllabus for Physical Science, their only reference to this section consists of the same quotation from the OEEC report that we have given above. However, teachers were initially given no indication of how the OEEC recommendation was to be implemented. The present writer, entrusted with the task of writing the portion of the Teacher's Guide relevant to this section (Prinsloo et al 1968), took it upon himself to suggest that the particle picture of matter should be treated as a plausible model and recommended the use of certain experiments to provide supporting evidence for the particle picture.

/Apart.....

Apart from this section and later sections dealing with the kinetic theory of gases and the nature of an electric current, there are no explicit indications in the syllabus of the way in which the concepts of the atom, molecule and electron might be used in teaching these parts of the course not directly concerned with them. In general, one might voice the criticism that insufficient attention has been given in the drafting of South African science syllabuses to the satisfactory incorporation of modern ideas into the general teaching scheme. Understandably enough, the main trends in syllabus revision have been the inclusion of material previously untaught and the exclusion of material now thought to be less "relevant" than it once was. The stage has now been reached in which a better-integrated presentation can be worked out. The working-out of such a presentation is not the primary aim of this thesis, but it will be both necessary and expedient to give some consideration to this problem at various points in the succeeding pages.

2.4 The use of "black boxes" in elementary teaching

The term "black box" probably comes from electronics and electrical engineering, in which fields it is used to denote a device whose internal constitution and/or mode of operation are unknown to the user. Its incorporation into the vocabulary of the schoolteacher is probably attributable to the PSSC, who include a section on black boxes in the chapter on "Measurement" in their physics textbook (PSSC 1960). As they put it, "we speak of a physical system of any kind as a 'black box' when we have made use of it without analyzing how it functions, without lifting the lid." They point out that in our everyday activities as professional scientists or science teachers, we constantly and unavoidably make use of "black boxes", ranging from our own sense organs to exceedingly complex items of electronic equipment. The PSSC authors also remind us that for any one

/one.....

one of us, some black boxes appear blacker than others, and that as our work and understanding progress we can "open" more and more of them. Perhaps most important of all, these authors outline the process whereby we acquire the judgment to know when and how to use black boxes; this we do by using the boxes to measure what we already know - we calibrate them.

The lecturer may take it for granted that he will be using black boxes, real and conceptual, throughout his career; in this, he has no choice whatsoever. Where he does have a limited choice is in respect of what black boxes he may reasonably use at a given stage in his presentation of a topic. This is an important problem to which the conscientious teacher and anyone seeking to guide him must give careful thought. Today, apart from the black boxes which have for years formed part of the teacher's stock-in-trade, there are many wonderful devices available which seem to offer exciting possibilities for classroom demonstration or even laboratory work. Cathode-ray oscilloscopes, D.C. amplifiers, electronic timing devices, audio-frequency oscillators and many other alluring items of equipment - can the teacher safely use any or all of these at the first opportunity or, if not, at what stage ought he to use them, and how should he decide this?

Of course it is the last of these questions which is really fundamental. On what grounds does one decide that it is permissible to use, say, a scaler-plus-quartz-oscillator as a clock in an experiment to determine "g" by free fall? The difficulty that lies at the back of the teacher's mind is this. The operation of black boxes used in physics teaching very often depends on principles not yet explained to or understood by the pupils. So when one uses such a device to illustrate or explain a given principle, is one not perhaps cheating, especially when the principle being illustrated is one on which the

/operation.....

operation of the black box actually depends? An advocate of a strictly logical approach to teaching would claim with some justice that one undoubtedly was cheating. But any teacher of other than the "armchair" variety cannot be a strict logician in practice; in the words of the Nuffield Guide, "In practical teaching, where our aim is some sense of understanding, rather than a structure of rigorous logical building which we shall never achieve, we all of us have to leave out some of the argument and indulge in occasional loosening of logic." (Nuffield Guide I). Does this mean that the teacher is free to use all available black boxes as and when he pleases, without regard for logical succession of ideas? One might doubt this, since indiscriminate use of black boxes is hardly consistent with an "occasional loosening of logic". We deal with the general problem of logic in physics teaching in the next section, but at this stage we can make the assertion that unless one pays a good deal of attention to logic in one's presentation, one's pupils are bound to receive a very confusing idea of science.

Having looked at the underlying principles, we may now return to our fundamental question. We can give the following short answer to it: one can justifiably use a given black box to illustrate a point or principle if one can provide some sort of convincing demonstration that the device does what it is intended to do. Then the black box is not just a beautiful toy which bewilders as much as it amuses; its potentialities as a measuring instrument stand revealed for all to appreciate. Consider how we could apply this principle to our example above of the scaler-plus-quartz-oscillator used as an electronic clock. The working of this device would be completely obscure to a pupil meeting it for the first time in a mechanics lesson, although the teacher could lessen the obscurity a little by drawing attention to the fact that the (invisible) quartz crystal is executing vibrations; if pupils had previously been given some idea of the association of clocks with repetitive motions, this point

/might.....

might be worth making. In any case, the function of the scaler-plus-oscillator can be made plain by showing that the seconds "ticked" off by it are the same as those of a conventional clock. From this it is a reasonable extrapolation that the electronic clock is accurately ticking off tenths, hundredths, thousandths and ten-thousandths of a second, and the teacher can use it to measure the short time intervals characteristic of the "g" - by free-fall experiment with some confidence that his pupils will believe what he is doing. (The simple make-and-break circuits used to start and stop the clock should not be difficult to understand.) On another occasion he could use this electronic clock to measure the even shorter times taken by sound pulses to travel distances of the order of a metre or two. Conventional mechanical clocks or stopwatches are, of course, black boxes to most people, in the sense that relatively few of us ever trouble to find out exactly how they work. But they will be rather less black than electronic clocks to many pupils, who will know something about the role of the balance wheel or the pendulum and be familiar with the principle of gearing down.

The use of conventional clocks or stopwatches to make measurements in mechanics might itself be thought to raise a little issue of logic. As the Nuffield Guide (I) points out, "We never worry about using a stopwatch (whose balance-wheel controls the time by executing simple harmonic motion) to time a pendulum or even a loaded spring. We use the watch and make perfectly good discoveries concerning simple harmonic motion." The discoveries are indeed "perfectly good", as can be seen from the background story to them, which runs something like this:- "We base our measurement of time on repetitive motion and so design our clocks to be driven by repetitive motion of some kind. Because of this it does not surprise us that many of the repetitive motions we observe around us are found to have

/a.....

a constant period. More careful study of such motions shows us that when they have small amplitudes, they have a number of noteworthy and easily describable characteristics: the amplitude remains nearly constant (if we are lucky!); the acceleration is proportional to the displacement from the equilibrium position; the displacement, velocity and acceleration vary sinusoidally with time; and so on". There is nothing circular in the logic of all this; the clocks used in our laboratories may have depended on simple harmonic motion for their operation, but they have to be calibrated against other clocks (e.g. atomic ones) which are not thus dependent. This fact also makes our observations concerning the constant period of repetitive motions less trivial than they otherwise might have seemed.

The cathode-ray oscilloscope is rapidly becoming standard equipment in school laboratories. A knob-twiddler's paradise outside and a maze of complicated electronic circuitry inside, it is the very archetype of a black box. Its possible uses as a teaching instrument are so many that we would do well to consider how it might be introduced to pupils. In elementary teaching one might first want to use it to show the "waveforms" of musical sounds. In this application, it is used to draw graphs of pressure versus time. If we do not wish merely to assert that the combination of oscilloscope and microphone can do this for us, we must think of a way of convincing pupils that it does. The essential steps in the argument would be (i) to show that if a point is simultaneously subject to horizontal movement at constant speed and vertical movement that varies in some arbitrary way with time, the resulting curve traced by the point is a graph of the variation of the vertical displacement of the point with time:
(ii) by setting the time base of the oscilloscope to a slow speed of sweep to show that the spot on the screen does appear

/to.....

to move across the screen at uniform speed:
(iii) to point out that this horizontal speed of travel is something that is determined by the controls on the instrument and not by what we put in from outside:
(iv) to show that the up-and-down movement of the spot is determined by what goes into the microphone - the harder one shouts or claps one's hands, the more the spot jumps. These are the essential points; it is not necessary to say anything about the adjustments that have to be made to obtain a stationary pattern, though pupils sufficiently interested to ask about this could have the point explained to them easily enough. Teachers who still feel that too much mystery surrounds the operation of the oscilloscope in this type of application might like first of all to show their pupils photographs of oscillograms obtained with Dayton C. Miller's "Phonodeik" (Miller 1926), which does reproduce the "waveform" of a sound without electronic assistance.

The oscilloscope is a happy example of the kind of black box whose lid can gradually be lifted as the pupil's knowledge and understanding of physics increases. Once he has learned something of the fundamental concepts of electricity and been shown the phenomenon of thermionic emission, he will be able to understand what an electron gun is and how the cathode-ray tube works. Then he will understand that in many applications the oscilloscope draws graphs of voltage against time, and that in the oscilloscope demonstration of sound waveforms, the microphone provides the link between the pressure variations of the sound waves and the voltage variations portrayed on the screen. He can then draw full benefit from oscilloscope demonstration of the "waveforms" of alternating voltages from the mains or a simple A.C. dynamo. This is most valuable, because only in this way can the pupil see direct confirmation of the theoretical curve that has been sketched on the blackboard or drawn in the textbook. Note that we have said "full benefit"; the

/pupil.....

pupil is not handicapped in his appreciation of this demonstration by his ignorance at this stage of the principles of electronic amplification. Proof that the spot really is deflected by an amount proportional to the voltage applied at the vertical input terminals is easily obtained by a simple exercise in calibration, making use of known voltages applied to these terminals. This is an important exercise for the pupil to carry out, for it brings home to him the fact that the oscilloscope can be used as a voltmeter (see also Nuffield Guide IV in this connection).

If the syllabus includes a survey of the rudiments of electronics, which it should, the lid of the oscilloscopic black box can be lifted even further by virtue of the pupil's newly-acquired familiarity with the principles of amplification. Perhaps, too, something of the idea behind the operation of the time-base circuit can be explained, together with an initiation into the mysteries of "synchronization", "flyback", and "triggering". The pupil's understanding of how an oscilloscope works would still be far from complete, but it has at least advanced to the "block-diagram" stage; moreover, he may by now have quite a good idea of what happens inside some of the blocks. It will by now be realised that there would be nothing offensively illogical about the use of this eminently electronic instrument in teaching the fundamentals of electronics; we can even use it to demonstrate amplification, though we should preferably keep its own amplification setting constant during this process.

Not all black boxes lend themselves so readily to a general revelation of their contents as does the cathode-ray oscilloscope. The previously-mentioned scaler is a case in point; another is the D.C. amplifier which, of course, has to operate in conjunction with a galvanometer, ammeter or voltmeter.

/The.....

The unquestionable advantages of this type of instrument are that it makes possible the detection and measurement of extremely small currents (e.g. 10^{-12} A or even less) such as are typical of weak ionization, photoelectric emission in certain types of photocell, and charge and discharge by electrostatic induction. The more elaborate D.C. amplifiers include provision for measurement of a wide range of currents and voltages and even of electric charge. The Nuffield Guide IV remarks that "These devices can hardly rank as 'familiar black boxes' like stopwatches"; however, they should not be ruled out of order on this account. As has been suggested above, the satisfactory use of an instrument as a teaching device will depend on how the pupil is prepared for it. Since in elementary teaching the D.C. amplifier is most likely to be used to measure very small currents, its introduction to the class should take the form of a demonstration that it can work like an ammeter or galvanometer. For this purpose, the combination of D.C. amplifier and ammeter can be connected in series with a galvanometer or microammeter and used to measure currents that give easily readable deflections on the latter. The class can then reasonably accept the use of the instrument to measure much smaller currents, and provided <u>absolute</u> values of these currents do not have to be used (this would raise awkward questions of calibration) there would be no serious transgressions of the canons of logic. Since absolute values of currents are not needed in introductory experiments on ionization by radioactivity, the photo-electric effect and electrostatics, this condition is easily satisfied in elementary teaching.

It is clear that use of the D.C. amplifier in the way just described presupposes previous acquaintance on the pupil's part with another black box - the ammeter. This rules it out for

/use.....

use in introductory demonstrations in electrostatics, at least in a context in which the serious study of electricity begins with electrostatics. Perhaps the view expressed by Harrap (1966) is rather too optimistic (or pessimistic, depending on the point of view): "... in five years' time the availability and convenience of D.C. amplifiers, together with their output being displayed on an obviously electrical meter, may make the gold-leaf electroscope obsolete." It is quite likely that the D.C. amplifier will establish itself as an increasingly useful teaching tool, but its very nature (and cost) would seem to preclude its use too early in a physics course, even though it may one day become almost as familiar as a stopwatch.

So far our discussion of black boxes has been confined to measuring instruments. These are by no means the only devices which fall into this classification, however. Many of the things that are commonly used in teaching laboratories are also black boxes, in the sense that the teacher or his pupils do not know all that there is to know about them. Pupils may use magnifying glasses in order to read the scales on certain instruments more easily; perhaps they have not yet learnt anything of the theory of lens behaviour, but this does not stop them from using these devices with complete confidence. The use of mercury in barometers and manometers is accepted, possibly long before the pupil has found out anything about surface tension or the properties of liquids and metals in general. To this extent, mercury is a black box for him. And what of that indispensable device in the teaching laboratory, the primary (or secondary) cell? How far would one get if one insisted on lifting the lid of this black box before using it in demonstrations and experiments? This example is so interesting that it deserves detailed consideration. When one uses a cell or a battery in elementary physics classes and wishes to explain

/something.....

something of its function, one might adopt the following approach: "Here is a two-terminal device whose internal constitution we shan't bother about for the moment. When we connect its terminals to the ends of a conductor, a flow of charge is maintained in the conductor. Let us now try to find out in physical terms what the cell or battery has to do in order to maintain such a flow of charge." The result of the investigation suggested in the last sentence would be a description of the cell's action in which one talks about the maintenance of concentrations of opposite charges on the terminals and the need to supply energy to any additional charge pushed on to these terminals. How the cell accomplishes these things is not of immediate concern. Later, this aspect of the black box may be exposed to view - probably in a chemistry lesson. In the meantime, cells and batteries can be used as D.C. sources - a fortunate circumstance, for how else is one to operate D.C. circuits unless it be by the use of electronic power supplies (black boxes again) - and their EMF's and internal resistances can be measured.

Those who feel that this is somehow "wrong" can perhaps derive a little comfort from history. The first primary "cells" were made by Volta (cf. Magie 1935) who took as his starting-point Galvani's work on the contraction of frog's legs under the action of (supposedly) electrical stimuli. Volta realized that his pile of "cells" ..." [provided] an unlimited charge or [imposed] a perpetual action or impulsion on the electric fluid," but neither he nor his contemporaries had any idea of the mechanism whereby this was achieved. This did not deter people from making and selling Voltaic piles and using them for further research into electricity. Indeed it was through the use of Volta's newly-invented black box that Nicholson and Carlisle were able to achieve the electrolysis of water

/and.....

and Davy that of molten salts of potassium and sodium, thereby discovering important clues to the action of the Voltaic cell. Even now, the details of what happens inside primary and secondary cells are occasionally the subject of controversy, but we go on using these cells just the same. If this was the actual sequence of events in history, then surely we need not feel any misgivings of conscience at using cells or batteries as black boxes in physics classes.

Enough has now been said in respect of these selected "black boxes" to show that the enthusiastic teacher who is also thoughtful can devise a satisfactory rationale for the use of almost any black box with which he might want to enhance his teaching. And this is just as well, because he is constantly going to have to use such devices, whether they "rank as familiar black boxes" like stopwatches or are more sophisticated, like some of the examples discussed above. As physics advances, more and more black boxes will become available to the elementary teacher (consider the laser, which did not exist prior to 1960 and is now widely used in the teaching of physical optics) who must be prepared for this situation and be able to derive maximum benefit from it. We shall accordingly include references to black boxes in the following pages wherever we feel it appropriate.

2.5 The use of strict logic in the teaching of concepts and principles

To some extent, the above topic has already been under discussion in Sections 2.2 - 2.4. This is as it should be, for this is a fundamental issue, underlying all that is done in the name of physics teaching. The trainee teacher, preparing a model lesson for his examination in the practice of teaching, is likely to select a topic in which he can demonstrate his powers of helping his pupils to appreciate how one thing follows

/from.....

from another. Unless he has been conditioned to be a mere fact-swallower, incapable of thinking critically, he knows that his own acceptance of scientific facts and principles is partly based on the logic of the arguments presented in their support, and he would like to foster a similar attitude in his pupils. We know that his instincts are right in this; that there is a "logic of scientific discovery" which should be respected in the teaching of science. Awareness of this makes one wish to avoid circularity in theoretical arguments or in the reasoning with which one supports one's experiments - one wants to escape the embarrassment of a "Zanzibar" argument or experiment (Rogers 1960). (The association of the name "Zanzibar" with logical circularity derives from a fable concerning a retired sea captain and a jeweller, both of whom lived on the island of Zanzibar. The jeweller set his clocks according to the time-signal provided by the sea captain, who fired off a gun at exactly noon every day; and the sea captain ensured the accuracy of his time signal by constantly checking his chronometer against the clocks kept by the jeweller!) Rogers and Tricker (1960) have given examples of this type of absurdity, which may still be found in elementary textbooks; to these and others we shall draw attention later in these pages. Tricker has also commented on another kind of circularity of reasoning which, he says, is becoming widespread: "... the use, in elementary stages, of concepts derived from later stages, the consideration of which depends on an understanding of earlier ones." The examples he gives take the form of the bald assertion of known facts about the electron, proton, neutron, the atomic nucleus and the hydrogen atom at the beginning of elementary courses in physics and chemistry. In one of the examples cited, the subject of electricity is actually built up deductively from some of the basic facts of atomic theory. This can be logical in its way, but one has to ask

/whether.....

whether this is the kind of logic that is wanted. The best justification one can invent for this kind of approach is that in it, the basic ideas are treated as postulates to be justified a posteriori by the experimental confirmation of the theory based on them. Surely, some might say, this is exactly the way subjects like relativity and quantum mechanics are taught, and nobody complains about this? It is instructive to compare the latter situation with the former. None of the kinematical or dynamical predictions of special relativity lie within the direct experience of the student (seeing the formula $E = mc^2$ in comic strips hardly comes under the heading of "direct experience"). A sensible way of teaching the subject is therefore to state Einstein's postulates after an adequate introduction and then to derive the standard predictions from them, preferably by way of properly-chosen "thought experiments." Essentially the same could be said of quantum mechanics: many (in a sense, all) of its predictions do lie within the direct experience of students, but the links between these facts of their experience and the basic principles of the theory are so complicated that it would be impracticable to teach the subject in a sequence of inferences from everyday and laboratory observation. With "classical" electricity and electromagnetism the situation is altogether different. The links between the relevant phenomena and experiments and the principles inferred from them are easily demonstrable and so make possible the general pattern of teaching the subject which has long been in use. To supplant this with a teaching order in which, for example, the ampere is defined as 6.25×10^{18} electrons per second and Ohm's Law is deduced from Drude's electron theory before any evidence for the existence of electrons has been presented, seems a perverse refusal to take advantage of the aids to comprehension so generously made available by Nature.

/In.....

In answer to our earlier question, then, purely deductive logic is not the kind that is wanted in elementary physics teaching, if the alternative of a combination of inductive and deductive reasoning can be used. Only in the latter case can the teacher build up the subject with frequent references to the pupil's own experience and so have some chance of implanting understanding. Tricker (op.cit.) makes the point that "... the deductive method gives no insight into the mind of the originator of the concepts and the problems he faced". This may seem to be an appeal for preference to be given to the historical approach, but in our view it can be interpreted more widely than this if we alter the wording slightly to read "... gives no insight into the way in which concepts originate." This insight is a vital part of the understanding of science, and a teaching approach in which the pupil can be led towards it is obviously to be preferred to one in which concepts are placed before him as "gifts from Heaven" (Rogers 1960). This kind of teaching approach does not have to be given a specifically historical slant, although this might be helpful (cf. Section 2.6).

The title of this section refers to "strict logic". Thus far we have concerned with the importance of logic in physics teaching and not with its degree of strictness. How strict, in fact, should the application of logical principles be in elementary physics teaching? In this connection, the Nuffield Teacher's Guide (Nuffield Guide I) commits itself to the following observation: "Children are not logical in their learning. They will not thank us for strict logic in building our teaching order and in providing experimental checks before each new step. That is what we should perhaps do in the sixth form and must do in university physics; but here we want to build some understanding rather than puzzle our pupils with formal requirements.

"We are concerned with what the pupil actually learns and knows and understands rather than fulfilling our own needs of

/proper.....

proper logic in building sound science. We must of course build sound science as far as we can, but we should accept the limitations of pupils' viewpoint and skills and make the best of them." If the opening statement of the quotation is correct, some compromise with strict logic is needed if pupils are not to be "puzzled with formal requirements". The Nuffield way of handling the problem has already been illustrated in the section dealing with the use of the concepts of the atom, molecule and electron in elementary teaching; we saw that in the Nuffield course atoms are treated as "familiar things" from the beginning, but that this familiarity is constantly subjected to challenge in the form of reference back to observation and experiment. (Of course, this is very different from the <u>postulational</u> use of the concepts of the microscopic world.)

There are other ways in which the requirements of strict logic can be relaxed somewhat in order to present an image of physics which is not too forbidding at first. An obvious one is deliberately to omit certain steps in the argument, though if possible this should not be done without some warning being given as to what is being done. For example, suppose the concept of force is being introduced in the lower standards of the senior school. The treatment could begin in the usual way with reference to pushes and pulls and to examples of forces encountered in the pupils' everyday lives. But to make the subject quantitative, it is desireble to bring in the idea that force can be measured. The pupils will not yet be at the stage when they could derive proper benefit from a treatment of the relation between force and acceleration, so that some other means has to be found of introducing the measurement of force. An obvious possibility is to use a spring balance calibrated by hanging known masses on it (cf. Helm et al 1968). For the spring balance to be calibrated in this way requires the use of the fact that gravitational force is proportional to the mass of the object on which it acts. In the absence of a proper

/dynamical.....

dynamical treatment of force, in the context of which it can be shown that gravitational force is proportional to (inertial) mass, this fact cannot be given adequate justification. Should one then abandon the idea of having pupils measure force at this stage of their careers? Only if one insists throughout on a rigorously logical approach. If one does not so insist, then it is still possible to continue by saying something like "We just don't know whether it is true or not that the gravitational force on an object is proportional to its mass, but it seems reasonable that it might be. Let's assume that it is, and see how we get on." In this way, the wool has not been pulled over pupils' eyes, but not too much fuss has been made about the logical difficulty. In the meantime, the pupil has learned to think of force that something that can be measured, and he is able to use this notion in forming his early ideas about work and energy. (For an alternative and more rigorous approach to this topic, see Section 3.8.3.)

It is generally agreed that there are some logical difficulties which one should not worry too much about even at a fairly advanced level of teaching; indeed, too much concern for this kind of thing would drive the over-conscientious teacher mad. Nevertheless, the number of steps omitted from a given argument should be kept to a minimum, at any rate for more advanced pupils, who frequently show a praiseworthy reluctance to take scientific facts on trust - and obvious circularities in logic should certainly be avoided. It is up to the teacher, in fact, or to those who plan his teaching approach, to arrive at a kind of "norm of reasonableness" in logical rigour, so that on the one hand the logic is strict enough to introduce pupils to the discipline to which scientific arguments must submit, while on the other hand it is not so strict as to stand in the way of a presentation of a topic which conveys the essence of the truth but which is not completely immune from criticism on logical grounds. At this point we

/link.....

link up with our earlier discussion of myths, half-truths and oversimplification, for the decisions that the teacher has to make in this connection concern the extent to which he is going consciously to simplify what he is teaching. We have implied earlier that the teacher should be aware of as much of the whole truth about the subject as possible; here we should like to stress that this knowledge should be the background to any decisions on the 'norm of reasonableness" in logical rigour that the teacher might have to take. In the following pages, the attempt will be made both to describe this background and to suggest teaching approaches which conform to the criteria suggested above.

2.6 The historical or "humanistic" approach to physics teaching

To the best of the writer's knowledge, no authorities today seriously advocate the wholehearted adoption of the historical approach to the teaching of science. Quite apart from anything else, there simply is not enough time available at school or university to make such an approach practicable. On the other hand, a number of authorities have put forward eloquent pleas for science to be taught in a more "humanistic" way than before, and they make it clear that this implies the inclusion of a certain amount of history in the teaching plan. The most recent of these pleas has come from Rabi (1967), who sums up his message thus: "So what I propose as a suggestion for you is that science be taught at whatever level, from the lowest to the highest, in the humanistic way. By which I mean, it should be taught with a certain historical understanding, with a philosophical understanding, with a social understanding and a human understanding in the sense of the biography, the nature of the people who made this construction, the triumphs, the trials, the tribulations."

Various concrete suggestions have been made concerning the

/application.....

application of this kind of teaching philosophy in classroom and laboratory. Professor W. Schaffer (Brown and Clarke 1960) thinks that "... the laboratory is a very suitable place in which the historical approach can be introduced, in which we can keep men in the humanity of physics, and in which experiments can be carried out. We should carry out practical experiments due to men like Galileo and Newton, to Young and Fresnel. These can be introduced so that physics can be seen as the kind of adventure of the spirit, the human quest in which men with certain kinds of knowledge at their disposal and with a certain intellectual climate around them attempted to solve their problems." A new school physics course known as "Harvard Project Physics" (Holton 1967) represents a full-scale attempt to teach physics in a humanistic way; it achieves this, among other ways, by the use of carefully-selected areas of the history of science as "pedagogic aids" and by attention to the way in which science has influenced and can influence the world around us. Students of the course are provided with "readers" which include passages not only from scientific literature but also from a variety of other sources ranging from Shakespeare to science fiction. It is noteworthy that even in courses not otherwise committed to a humanistic approach, some provision is made for the inclusion of certain slices of history. Thus in the Nuffield Physics course, Year V, a considerable amount of time is devoted to a detailed study of the development of man's knowledge and understanding of the solar system. This is done in order to give pupils the full background to Newton's crowning achievement in "explaining" the solar system by means of a theory that was to prove even more fruitful in the years following his death. The PSSC course includes a section in which the same material is rather more rapidly covered. In the new Common Basic Syllabuses for Physics and Physical Science now

/coming.....

coming into use in South African schools, allowance is also made for the inclusion of historical matter, but the wording of the relevant instruction is cautionary: "The history of physical science should only be drawn on where the historical development is an aid to understanding." Since our ultimate concern in these pages is with the teaching of fundamental physical concepts and principles, we must now consider whether the limited use of the historical approach can be "an aid to understanding" in this connection. Apart from its value in making the presentation of the subject more "humanistic", does the historical approach help to sharpen the pupil's perception of physical concepts?

As an example, let us take the study of motion. In teaching this branch of physics, we need to introduce the concepts of displacement, velocity, acceleration, mass and force, among others. Does it help us if we appeal to history in trying to convey their meaning? Consider acceleration: in a non-dogmatic approach to this concept, a teacher would not simply confront his pupils with the approved definition of acceleration, and carry on from there; he would first talk to them about the idea of acceleration as being vaguely associated with change of speed, and then invite suggestions for a way of defining some measure of this concept. It is more than likely that a class led on in this way would suggest both $\frac{dv}{dt}$ and $\frac{dv}{ds}$ as possible measures (though of course they would not express them in these terms). In other words, they would bring to life one of the situations dramatized by Galileo in his "Discourses concerning the Two New Sciences". At this point in the discussion, the teacher might proceed to point out to his pupils that Galileo had discussed this very problem, and present them with his arguments and solution. Or he might go on to reach the same conclusion witnout any mention of Galileo. In either case he

/would.....

would hope that he had helped his pupils to gain some insight into the present-day meaning of acceleration, but whether or not he wished to bring Galileo into the discussion would surely have been a matter of taste. It might be argued that whether or not Galileo is mentioned in this case, the approach used is really the historical one. To this we might reply that there is nothing essentially historical about this approach - that it is the natural approach one might use in trying to teach acceleration by drawing on the pupil's own experience. History makes its entry when we link what we are teaching with specific historical figures and events.

What of the other concepts listed at the beginning of the previous paragraph? Take mass and force, the evolution of both of which has been carefully chronicled by Jammer (1957,1960). One would probably do more harm than good by spending appreciable time in the classroom on early ideas of what is now called "mass"; it would be much better to lead the pupils as directly as possible to the modern view. As for force, recent developments in the teaching of mechanics have shown how to introduce it in a satisfactory way without reference to history - except, of course, for the mention of Newton's name in the proper place! In general, we might say of the use of specifically historical matter in the teaching of physical concepts that in certain cases this could be helpful, in that one could thereby add a human touch to one's teaching and so help to fix an important idea in pupils' minds; but we have to add that the same end might well be achievable in other ways.

It may happen that certain principles or theories in physics might with advantage be taught along historical lines. There is much to be said for arriving at the Principle of Conservation of Energy along a route which, at any rate in its final stages, is essentially the historical one, i.e. via the work of Rumford and Joule. (Cf. the Nuffield physics course,

/in.....

in which pupils review an impressive accumulation of historical evidence in favour of a fixed quantitative relationship between heat and work, and are then invited to draw their own conclusions.) For the teaching of the characteristics of a good scientific theory, the story of the evolution of the Newtonian picture of the solar system is particularly well suited; we have seen that the Nuffield physics course accords it a prominent place on this account. But here also we could argue that it is not essential to make one's teaching explicitly historical: equivalent teaching approaches could be devised which, in principle, could be just as effective; or an approach derived from history could be used but within a non-historical frame of reference. (Cf. the PSSC approach to the nature of light, in which the particle model of light is discussed with only occasional references to history which are in any case not essential; only at the end are experimental results obtained by Fizeau, Foucault and Michelson cited in order to show up certain defects of the model.) One must emphasize that unless a great deal of time is at the disposal of the teacher, his account of history will be one in which the facts are subjected to a great deal of streamlining and editing. A close look at virtually any portion of the history of science will reveal the vast amount of effort spent on issues which seemed vitally important when they were first raised but are now regarded as unimportant or irrelevant. Science's obstinate refusal to progress along an unswerving logical path is certainly something that every science student should know about; he should be aware that science often takes wrong turnings, the wrongness of which may only become evident after some time has passed. This may well be the most valuable lesson which the history of science can teach him. But it is out of the question for him to be taken along every

/step.....

step of the original devious route followed by science. He could best gain a feeling for the way in which science progresses by doing a little "research" of his own under careful direction. (Cf. again the Nuffield physics course, which makes explicit provision for this kind of work.)

Even where history does seem to have given us a well-ordered way of presenting part of a subject, it does not follow that our adherence to this need be slavish. Newton's formalisation of the foundations of mechanics might appeal to us, for example; what was good enough for Newton is surely good enough for us. But is it? For one thing, his terminology would have to be rendered in modern language; for another, we might be a little unhappy about his definition of "quantity of matter" (see our later discussion of this point); and for yet another, we would probably not want to announce the Laws of Motion so near the beginning of the discussion. So for sound pedagogical reasons, we would be compelled to depart somewhat from Newton's way of presenting the subject. We need not feel conscience-stricken about this; in writing his "Principia" Newton was not trying to provide a textbook for use in schools or universities.

It has become common practice in physics teaching deliberately to follow certain of history's celebrated "wrong tracks". In some cases one may question whether this is altogether wise. For example, does one really need to spend very much time on the details of the Bohr theory of the atom when teaching atomic structure? In teaching special relativity, is it essential to make a tremendous fuss at the outset about whether or not motion relative to the ether can be detected? (This is not necessarily such good history as one might suppose. Einstein does not appear to have been specifically <u>motivated</u> by the null results of ether-drift experiments in formulating the special theory of relativity. Cf. Shankland 1963, 1964). In discussing the

/constitution.....

constitution of the nucleus, must one first talk about the proton-electron model and show wny it must be ruled out, before introducing the proton-neutron model? The answers to these questions and otners like them are not necessarily all in the negative; they will be given due consideration later in this thesis. What is important is that these questions should be asked, for in trying to answer them we might arrive at approaches to teaching these topics that are superior to the ones we have used before.

Explicit attention to history in a school physics course could have at least one unfortunate result. (The author is indebted to Prof. J.A. Gledhill for calling attention to this possibility.) History, or at any rate history in the form in which it might be dispensed in the classroom, lends itself to being stored up in the memory for future regurgitation in an examination of the appropriate kind. At a time when strenuous efforts are being made to improve examinations in physics as tests of <u>understanding</u>, anything which tends to shift the emphasis from understanding towards memory work should be firmly resisted. One can all too readily visualize a situation in which the teaching of selected slices of potted history would be made the excuse for the provision of dictated notes on the subject and for the performance of even less practical work than usual. Of course this does not happen in such a teaching programme as Harvard Project Physics, which is well supported financially, selects and trains its teachers for its particular ends and keeps careful control of its tests and examinations. But in the South African context introduction into the syllabus of specifically historical material could easily lead to the kind of abuse deplored above. Perhaps this is one of the reasons why planners of South African syllabuses have preferred to exclude this kind of material.

/Up.....

Up to this point, the tenor of our argument has been that it is not by any means self-evident that the use of historical matter in physics teaching is a good thing. This must not be taken to mean that the teacher can safely neglect the history of physics in his own training. On the contrary, it is desirable that his knowledge and understanding of it should be as extensive as possible. In order to teach basic concepts well, the teacher should preferably be familiar with the way in which those concepts have evolved. A realization of how painfully slow the process of evolution often was will help him to appreciate the nature of the task that confronts him in the classroom in making concepts clear. A knowledge of the facts of history will also make him aware of the resemblance between the misconceptions of his pupils and those of early thinkers in the scientific field - he will find many budding Babylonians or Aristotelians in his classes - and help him to direct his pupils' thinking along the proper lines. In short, history can give the perceptive teacher valuable hints.

2.7 Physics as an "open" subject

It is one of the fundamental tenets of PSSC philosophy that "... the tentative nature of all knowledge must be consistently stressed" (White 1960). In the pursuit of this goal, "Theory and experiment alike are presented as processes of successive approximation, in which final answers are neither expected nor sought. It is repeatedly made clear that any area of physics, however well-trodden, ends at last at a frontier beyond which lies the unknown." In the same spirit, Mr. Norman Clarke has asked "... whether or not an entirely new approach is needed, an approach that presents physics from the earliest stages as a living subject, a subject that did not, even as far as schools are concerned, cease to develop about

/the.....

the year 1900" (Brown and Clarke 1960). Closely allied to this attitude towards elementary physics teaching is the view that even a school course in physics should "...[reflect], as much as possible, the world of physics as it appears to the professional physicist, for only in this manner is it likely to have any clear relevance to the student himself." (White 1960).

It is unlikely that anyone would seriously challenge the first of the above-stated principles. The evidence that physics remains an "open" subject is so overwhelming that one cannot now imagine some present-day or even future Lord Kelvin announcing that all the great discoveries in physics had probably been made and that future progress would be confined to more accurate measurement of known quantities. Such is the rate of advance of the subject that any physics course which does not at least hint at this is guilty of presenting a damagingly false view. One does not encourage a prospective physics student by, in effect, suggesting to him that the subject is not one to which he can make a positive contribution.

An obvious way to stress the existence of a "frontier beyond which lies the unknown" is to specify in detail the location of that frontier, i.e. to tell pupils about work that is going on in physics at the present time, about "the world of physics as it appears to the professional physicist." And it is here that difficulties seem to arise, because one immediately starts to wonder whether it is possible to convey to school pupils a meaningful view of contemporary physics. One feels bound to concur with Holton that "... the pyramidal structure of physics, so beautiful and almost unique to our field, makes it practically impossible to talk in honest detail about most of the actual problems on which physicists are now working" (Holton 1967). But while one admits the truth of this, one

/also.....

also notes the use of the important qualifying phrase "in honest detail" in Holton's statement. Detail, honest or otherwise, is not vital to an introductory mention or discussion of the problems on which present-day physicists are engaged. Holton himself seems well aware of this, since the list of contents of the "Reader" for Unit 2 of the Harvard Physics Project course includes articles on Gravity experiments (Dicke, Roll and Weber), the expansion of the universe (Bondi), negative mass (Hoffmann) and quasars (Feinberg), to mention only a few of those concerned with present-day physics. The problem, then, seems to be capable of solution; but to see how it can be solved we need to glance at the field of present-day physics, in the hope that we might acquire some valuable clues.

It is customary to classify research in any of the sciences under the headings of "pure" and "applied" research. This classification might seem to be helpful to the puristically-inclined teacher, who might say "I'd like to tell my pupils about modern research in physics, but of course it's only pure research I need worry about, because only this is the real thing." Such a teacher would no doubt have particularly in mind the work of the "elementary"-particle and nuclear physicists in talking of "the real thing", though he might perhaps widen his scope a little to include, say, the activities of the solid-state physicists. At the same time he would set his face firmly against any discussion of the contributions of physics to modern technology, under which heading he might include almost anything from electric shavers to hydrogen bombs. There is no doubt that the view of modern physics of our hypothetical teacher is much too narrow-minded; in analyzing why it is so, we can perhaps achieve a clearer understanding of the nature of "pure" and "applied" physics and at the same time arrive at a view of modern physics which is both more broad-minded and of greater help to pupils.

/In.....

In the first place, a clear definition might help. A short definition of "pure research" might be "research pursued for its own sake, i.e. without any particular applications in view." It is at once evident that this definition is much wider in its implications than that which our hypothetical teacher may be supposed to have had in mind. Whereas for him, pure research in physics would be concerned only with advances in knowledge of the basic principles of the subject, the definition just given embraces any research directed towards the advance of knowledge within the boundaries of physics. Even this definition might be held by some to be too narrow. How, for example, is one to classify the work of Georg van Békésy (winner of the Nobel Prize in physiology and medicine in 1961) Crick, Watson and Wilkins (joint winners of the Nobel prize in physiology and medicine in 1962), and Perutz and Kendrew (joint winners of the Nobel prize in chemistry in 1962)? Of these men, all but Watson and Kendrew are or were physicists, but their research has consisted in the application of physical principles to the unravelling of problems of pressing interest in other fields. Békésy, for example, is an experimental physicist of great ingenuity who, by the application of the principles of classical physics, has succeeded in clarifying many aspects of the functioning of the cochlea, while the other physicists named above have achieved eminence by the successful use of X-ray diffraction techniques in revealing the structure of the DNA, haemoglobin and myoglobin molecules. In a sense this is certainly pure research, but by its nature it does not lead to any advance in physical knowledge. Indubitably, however, it is part of "the world of the present-day physicist" and as such it is relevant to the concern of any teacher wanting to tell his pupils about that world. The fact is that professional physicists today are engaged in a wide range of research

/activity.....

activity, virtually all of which is potentially capable of providing stimulation both for those engaged in it and those who come to hear of it. Division of this range into "pure" and "applied" categories is not as meaningful to the pupil as it might be to his teacher and, as we have seen, it is not always so easy to fix the dividing line. In the circumstances it is sensible to adopt the broadest possible view of what constitutes "modern" physics. The purist attitude exemplified by our hypothetical teacher above was a natural enough reaction to the once-popular attitude among teachers and textbooks that the justification of physics (or indeed of any science) was to be found in its "applications". This attitude transferred itself readily enough to pupils; indeed the typical modern pupil may still be heard to ask "what's it <u>for</u>?", when looking at some demonstration on a physics exhibition. Certainly this is a distorted view of the function of physics but it would be unwise for any teacher to go to the opposite extreme.

Among the most prominent fields of research in which present-day physicists are engaged are nuclear and elementary-particle physics, solid-state physics, plasma physics, spectroscopy, and space physics and geophysics. At first glance, one might be inclined to classify all these as "pure", but this would be risky: plasma physics probably received its biggest stimulus from the need to develop controlled thermonuclear fusion, which is essentially technological; although space and geophysics pose many fascinating problems of a "pure" nature, research in these fields would not be as well supported as it is were it not for the possible technological consequences; and much of solid state research is specifically directed towards technological ends. But this is beside our present point, if our arguments above be accepted; all we are concerned with for the moment is to define some areas of interest in present-day physics and suggest

/how.....

how they might be linked with physics teaching in school.

It is not difficult for the teacher to indicate the location of the present-day frontiers of knowledge in nuclear physics. No discussion of the proton-neutron model of the nucleus would be complete without at least some consideration of how such a structure could be held together; and once one has introduced pupils to the idea of a specifically nuclear force, one has brought them into a realm of physics in which all the answers are not known. The same is true of the details of the nuclear structure itself - here the teacher need do no more than mention the fact that it has not yet proved possible to invent a model of the nucleus that will account for all aspects of its behaviour. A syllabus which includes the proton, the neutron and the electron provides an adequate starting-point for some remarks about "elementary" particles in general, with especial reference to the problems confronting modern research workers in the field. The pupil will have learnt enough about the techniques by which sub-atomic particles can be identified to have some appreciation of the way the evidence for these particles is built up, and while his and the teachers' speculation about the place of all these particles in the scheme of things is not likely to get very far, it does at least help to show something of the nature of the problem.

Virtually any modern physics syllabus for school use includes a certain amount of very elementary physics of matter in bulk, including solid state physics. Again there is enough background material here for the teacher to be able to point out that the fundamental difficulty in this field lies in the vast numbers of particles with which one has to deal, and that it is largely on this account that the discovery of a number of phenomena in solid state physics ante-dated their theoretical explanation by many years.

/Plasma.....

Plasma physics presents one with essentially the same difficulty with this difference: that whereas with solid-state physics quantum mechanics has to be used in order to give the full story, classical mechanics and electromagnetism suffice for this purpose in the case of plasma physics. It is neither desirable nor necessary for the teacher to introduce pupils to any of the jargon of this subject (Debye length, plasma frequency and so on); on the other hand, since most of the matter in the universe is in the plasma state, pupils ought not to remain ignorant of what a plasma is, and they should also be given some indication of the importance of plasma research both to technology and to other branches of physics. Here is a link with space physics, in which a knowledge of plasma physics is needed for the understanding of the behaviour of planetary ionospheres (if any) and magnetospheres, to mention only two topics of importance. A significant feature of space physics which sets it apart from those branches of modern physics so far considered, is that its phenomena cannot be made the subject of controlled experiment in the same way as other phenomena can. To take one example only: one might like to find out whether the variation of electron density with height in the earth's ionoshpere depended in a particular, predicted way on electron and ion temperatures at the given heights. Techniques exist for measuring these quantities with fair accuracy, so that in principle the problem might seem simple enough to solve. In practice, it is impossible to take a portion of the ionosphere and isolate it so that its electron density distribution is influenced by temperature alone; several other complicating factors are always present. The only way out of this kind of difficulty is to take as many observations as possible at various times and see whether a pattern emerges which can convincingly be explained in one way and no other. For pupils taught to try to "vary one thing at a time" in their experimenting, this

/would.....

would be a new situation which they certainly ought to know about; for one thing, it will help them to appreciate the difficulties faced by the weather forecaster!

In short, the interested teacher will be able to find "leads-in" to almost any branch of research in modern physics, which will make it possible for him to say something interesting about these topics without confusing his pupils with "honest detail". He may as well make the effort, because there is every likelihood that the scientific society in his school might anticipate him in this, and it would be a shame for him to leave all the exciting contemporary material to an extra-curricular body. Naturally he will be unable to fulfil his proper function here if he has not formed some acquaintance with "the world of the present-day physicist". One could argue that an essential part of the training of a competent physics teacher is the provision of such an acquaintance, for even if he were not to take the initiative in the classroom in telling his pupils about these matters, he would not be able to avoid answering questions on them put to him by some of the brighter members of his class.. It is of course desirable that the answers he would give to these questions or the information he gives of his own accord should be correct. The kind of thing to be avoided is illustrated by the following extract from a recently-printed supplement to a much-used physical science textbook:

"Neutrons hold the positive protons in the atomic nucleus together. The real nature of the binding force is still to be explained, but μ and π mesons, recently discovered, appear to play a part. Mesons have a mass between that of a proton and an electron and may be either positively or negatively charged. The exchange of a meson between a proton and a neutron is suggested as the binding force." (Gordon, Neser and Pienaar, no date [the book appeared in 1964, however].).

/The.....

The present author has criticized this passage in detail elsewhere (Prinsloo et al, 1968), but it may be remarked here that its score of half-truths and plain mistakes per line is enviably high. It is impossible to justify this shoddy treatment of the topic on the grounds either of necessary oversimplification or that the source material is too erudite for even the conscientious teacher or textbook author to comprehend. A correct account at the same level would be no more difficult to understand than the original, which in any case consists largely in a recital of (alleged) facts; and as for the second suggested excuse, many popular and correct accounts are now available for the benefit of anyone interested. One cannot expect a teacher to have read all of the available source material, but he should at least be able to refer his pupils to some of it if the need arises.

Our suggestions concerning the establishment of links between syllabus material and the "present-day world of the physicist" were made on the assumption that the syllabus could provide the necessary opportunities. It will be clear from the above discussion that a syllabus does not have to be wildly "modern" in order to do this. Obviously, however, the more a syllabus is oriented "in the general direction of modern physics" (White 1960) and the further it goes in this direction, the easier it will be to link up with present-day work. This raises the important question of how far a school syllabus can reasonably go "in the general direction of modern physics". At this stage, we do not propose to answer this question in any but a general way: later in the thesis it will be dealt with more specifically with reference to particular topics. The main point we wish to make now is that <u>for the teacher's sake</u>, syllabuses should be designed to go as far in the direction of modern physics as is consistent with the pupil's potential

/understanding.....

understanding of the material. We say "for the teacher's sake" for this reason: that one of the most powerful incentives anyone has for keeping abreast of the development of a subject is having to explain it to someone else. We have come across many saddening examples of teachers who in past decades received their training under the most respectable auspices, and who have now drifted so far away from their subject in its present-day guise that they would appear to have little or no chance of regaining contact. It is at least a reasonable assumption that these teachers would not have been in this position if the requirements of the syllabus had compelled them to know a little of what was going on in the world of physics. On the other hand, we have come across heartening examples of teachers both young and old who have been prepared to accept the challenge of the partially "modern" content of recently-introduced syllabuses, and who have found that they could teach it better than they dared hope. The success achieved by this latter category of teachers carries with it this lesson: that one should not be too hasty in deciding that this or that topic cannot be taught. Subject to the limitations mentioned in Section 2.1, it is surely better to decide that some topic <u>should</u> be taught, and then try to develop ways of teaching it effectively, than to shrug the problem off at the outset. Take the principles of quantum physics: by general consent, these are difficult to grasp, even at quite an advanced level, and yet they are directly concerned with a great deal of present-day physics. On this latter account, several authorities have taken the bold step of including some quantum physics in school syllabuses, and much effort has been and is being put into the development of satisfactory ways of teaching this material at an elementary level. It is probably too early to judge the success of this experiment, but one may hope that trial and error will in the end achieve

/the.....

the desired result of the stimulation of pupil and teacher alike.

We have seen that principle of teaching physics as an "open" subject presents challenges to the teacher in respect of both his general acquaintance with present-day physics and his understanding of the concepts and principles upon which it is based. Thus once again the teacher's need for a well-developed understanding of fundamental concepts and principles is brought into prominence.

2.8 The role of practical work in the elementary physics curriculum

2.8.1 Introduction

Since physics could not exist without observation, experiment and measurement, no-one would seriously contest the importance of practical work in the training of the physics pupil. We find this attitude receiving official support in the published syllabuses of various examining bodies in South Africa. Thus the syllabuses of the Joint Matriculation Board in Physics, Physical Science and Chemistry were, in the years preceding 1965, each prefaced by the statement that "Although due regard should be paid to fundamental...theories, it is intended that the syllabus should, as far as possible, be treated from the experimental point of view, and the question paper will lay stress on experimental methods." (Joint Matriculation Board 1962). Similarly, the "Remarks" prefacing the syllabus in Physical Science for the Cape Senior Certificate (Cape Education Department 1961) include the comment that "It is very desirable that pupils should be given ample opportunity of handling apparatus and material, and of making their own observations and deductions". Unfortunately, the stress laid on experimental methods in examination papers has provided no guarantee that the candidates writing the papers would ever have encountered these methods. The author knows of several pupils at reputable girls' schools who had never even seen such familiar items of apparatus as

/ammeters,.....

ammeters, voltmeters, galvanometers and rheostats in their school laboratories. He has also coached African candidates for the afore-mentioned examinations who had never been inside a physics or chemistry laboratory. Worst of all, he has it on reliable authority that the headmaster of a reputable boys' school virtually eliminated practical work from his science classes, on the grounds that the time thus saved would enable his pupils to be better prepared for the written examinations. Whether the last-mentioned horror story is true or not, it is clear that the system hitherto in force has been open to serious abuse. Any improvement on it should be based on a careful assessment of the aims and value of practical work and detailed consideration of how these aims could be achieved; in this way it might be possible to secure a better treatment of the subject from the experimental point of view.

2.8.2 The aims of practical work

Various authorities have drawn up their own lists of the aims of practical work in school physics and, not surprisingly, there is a fair measure of agreement among them. Boulind (1961) lists the aims of school practical work as follows: development of skills, learning facts and principles, inculcation of scientific method, and maintenance of interest. Santur (cf. Brown and Clarke 1960) speaking actually of university physics courses, states the following as the aims of practical work: the development of manual skills, the development of faculties of observation, the evaluation of results, the gaining of familiarity with physical equipment, and the recording of observations and the presentation of results. Dechene and Eurin, (cf. Brown and Clarke 1960) representing the French point of view, think that practical work is valuable because it enriches personal experience, constitutes a means of developing the pupils' initiative, helps the understanding and interpretation of measurement and, when well-conducted, can develop team spirit.

/In.....

In recent years, those responsible for the development of the Nuffield Physics Project in Great Britain have built up a "Physics for All" course for pupils from "11+" to O-level which places considerable emphasis on practical work by the pupils themselves. They point out that "A very strong influence in young peoples' understanding of science and scientific work is their own experimental work....They need this personal experience of science." (Nuffield Guide I). Nedelsky (1965), as we have already seen (Section 2.1), lists laboratory knowledge and understanding and ability to learn from experiment and observation among the behavioural objectives of an introductory science course. Under "laboratory knowledge", he lists knowledge of apparatus and materials, laboratory procedures and relations between data and generalizations from data; under "laboratory understanding", he lists the understanding of processes of measurement, of experiment and of the relation between theoretical or "book" physics and phenomena, and the ability to design and carry out experiments and to interpret data; and under "ability to learn from experiment or observation", the ability to pursue experimental inquiry, possession of laboratory skills and imaginative thinking in the laboratory.

As might be expected, there is a good deal of overlap amongst these various statements of objectives. In particular, Nedelsky's extended list of precisely specified objectives includes most of what appears in the other lists. Even where an objective in another list does not coincide directly with one of Nedelsky's, closer examination may reveal that it embraces several of the objectives on his list. One can claim this to be true, for example, of such objectives as "learning facts and principles", "inculcation of scientific method", "providing personal experience of science" and "development of faculties of observation". For our purposes, it will be convenient to

/group.....

group the above objectives under eight headings, viz. development of skills, development of faculties of observation, evaluation of results, the gaining of familiarity with physical equipment, learning facts and principles, inculcation of scientific method, the recording of observations and the presentation of results, and maintenance of interest and the provision of personal experience of science. These aims will be examined critically and an attempt will be made to assess their relative importance for a high school course in physics.

(i) Development of skills Boulind has argued that this aim of school practical work is important for those who are to become research workers. He points out the obvious value of this kind of training for persons in this category and claims that the years spent at school provide much more opportunity for such training than do the three undergraduate years of a university course. He himself, however, estimates that no more than 15% of school pupils are likely to enter scientific careers in which practical skills are necessary and questions whether, for the other 85%, the school physics laboratory should be the place where the pupil learns the sort of skill which will enable him to mend fuses or replace tap washers. Boulind is, of course, discussing the problem in the context of the British educational system; in South Africa one would assert with much less confidence that the physical science pupil had more opportunity for development of manual skills at school than at university, and one would estimate the proportion of pupils likely to take up scientific careers as less than Dr. Boulind's 15%. Thus it would be even less easy in South Africa to defend this aim of practical work on statistical grounds than it is in Great Britain. A realistic approach suggests that the development of manual skills should be

/treated.....

treated as an incidental aim of school practical work; that as a result of carrying out a fair number of experiments during their school careers pupils would, with good guidance, acquire some ability to carry out simple measurements, repairs and constructional work. But it must be stressed that it is more important for the physics pupil to know what makes a fuse blow than to know how to repair it.

(ii) Development of faculties of observation Essentially the same arguments can be advanced in favour of this aim as for aim (i) above, and it can be argued with almost equal force that this is an incidental aim: "almost equal" rather than "exactly equal", because the ability to observe accurately comes closer to the real stuff of physics than does the acquisition of manual skills. But one can still claim that this aim can be achieved for the pupil as a result of his performing a number of experiments not primarily designed with this end in view.

(iii) Evaluation of results As physics is vitally concerned with the results of measurement, there can be no question of the importance of this aim. Unfortunately there are clear signs of its almost total neglect in South African science teaching. It is still possible to come across second-year physics students who will naively write down the result of an experiment to an implied accuracy of four significant figures without regard for the actual experimental situation - and this despite their having received instruction on this point during their first year at university. In many such cases there is good reason to believe that this undiscriminating attitude towards experimental results was acquired at school; and since the matriculation examination, despite its alleged stress on experimental methods, has never tested this kind of thing, there has been no incentive for the teacher to set his pupils right in this respect.

/Of.....

Of course we are not advocating the presentation of a rigorous theory of the treatment of errors as part of school physics. What is necessary is that the teacher should remind pupils that there are limits to the accuracy of any given measurement, and that these limits should somehow be reflected in the way the results of the measurement are recorded. Without some understanding of this point, the pupil cannot be said to have made proper contact with physics at all.

To clear up any other possible misunderstanding on this issue, we should also stress that it is not necessary for a teacher to spend a great deal of energy in classroom and laboratory in urging his pupils to be "accurate." A great deal of insight can be gained into certain areas of physics by means of measurements and calculations that may be quite rough; what is important is that the pupil should be to some extent aware of the degree of roughness. An appreciation of the manner in which experimental techniques can be refined to yield "accurate" results in particular circumstances is part of the advanced training of the physicist.

(iv) The gaining of familiarity with physical equipment

Everyone would surely agree that pupils should come to know something of the equipment with which physics is partly "made", and that they should preferably have the opportunity to see and handle such equipment. There are obvious limits to the extent to which opportunities of this sort can be provided in the elementary laboratory, and the point should thus not be over-stressed. But one would hope that pupils would have formed direct acquaintance with at least some of the simpler tools of the physicist's trade, such as the metre stick, the stop watch, the thermometer, the ammeter, the voltmeter and the cathode-ray oscilloscope - to mention only a few measuring instruments in common use. Any course in school physics which

/makes.....

makes reasonable provision for practical work can be made to achieve this aim.

(v) Learning facts and principles If by "learning" is meant the imprinting of facts and principles upon the memory by whatever means, there can be no doubt that practical work by the pupil is not essential to this activity. Those who have experience of the use of programmed instruction claim unrivalled success for it as a means of inculcating facts for regurgitation in the appropriate kind of examination and, as has already been mentioned, many South African pupils have already shown themselves capable of passing their matriculation examinations in physical science without having wasted their valuable time on practical work. But the conscientious teacher with the immediate as well as the ultimate welfare of his pupils at heart will not be content with this situation: he will want his pupils to learn their facts and principles by means less sterile and soulless, possibly by discovering them for themselves in the laboratory under his careful guidance.

In thinking about how to solve this problem, the teacher may fall prey to a certain temptation: that of doing it all by demonstration. In the paper by Boulind already cited, there is an impressive list of the advantages of demonstration experiments by the teacher over practical work by the pupils; demonstrations save time for pupils and teacher, they save money - "one set of apparatus costs less than 15" -, the teacher can emphasize crucial results or readings and underline the principles, he can get more accurate results than the pupils can, and he can use apparatus that would not be safe in pupils' hands. All this is undeniable, and the teacher could hardly be blamed for concluding that it would really be better if he just did it all himself. But in so doing he would unwittingly have decided to deprive his pupils of the fun they could derive from carrying out certain experiments individually. More will

/be.....

be said about this point later under the heading of "Maintenance of interest". In the meantime it is suggested that practical work by the pupils can be of considerable value in learning facts and principles, where by "learning" is meant both "learning to know and understand", and that careful thought should be given in the design of any syllabus to the selection of suitable experiments for this purpose.

(vi) Inculcation of scientific method In listing this among the possible aims of practical work, Boulind was careful not to speak of "The Scientific Method"; as has been forcefully pointed out by various physicists (e.g. Rogers 1960, Zemansky 1962) there is no single Scientific Method, but a variety of possible methods followed by scientists in pursuing scientific problems. By setting out to solve a problem along experimental lines, a pupil can gain experience of and insight into such methods. As Boulind points out, the type of experiment which a pupil performs with the aid of a detailed list of instructions is completely useless in this regard; a completely different approach in experiment-setting has to be adopted by the teacher. (Of course, "cook-book" methods are not entirely to be despised; all of us are content to follow the step-by-step instructions of the handbook provided when confronted with some unfamiliar appliance or piece of apparatus. But we do not need exhaustive training at school or university to fit us for this kind of activity.). Given a suitable problem and the time in which to solve it, the pupil can gain something from his own experimental work which it is virtually impossible for him to acquire in any other way.

(vii) The recording of observations and the presentation of results That individual practical work can provide training in the recording of observations and the presentation

/of.....

of results is obvious enough. The question to be considered here is the extent to which such training is necessary in the elementary stages. Much of our science teaching has been dominated by the assumption that it is necessary and has sought to impose upon pupils a rigid form of presentation of results. That this kind of training has not achieved the desired ends is only too evident from the kind of work produced by university students called upon to "write up" their experiments. In the main, these students betray little or no sign of being able to give a decent account of what they have done in the laboratory. Is there any point in attempting to train school pupils in the presentation of results if those whom it ought to benefit gain little from it? And what of those pupils who will never find their way into a research laboratory? Is this kind of training of any use to them?

In reply to the first of these questions one might suggest that if the training were improved, it might obtain better results; and in reply to the second one might say that if a pupil does experiments at all, he should get used to the idea of recording his results in a way that is intelligible to himself and preferably to others as well. The two replies are really interdependent. Improved training in the "writing-up" of experiments would lay greater stress on individuality, which would help to make the pupil more interested in what he was doing; this increased interest would in turn make it easier for the pupil, be he future scientist or not, to get used to recording the results of his laboratory work. This does not mean to say that every experiment carried out by the pupil should be made the subject of a detailed laboratory report, to be corrected in due course by the already overburdened teacher. Some experiments might be so simple that little would be gained by making notes of them (cf. the Nuffield Guide (I), "... they are things to do and see, and we should

/no.....

no more ask for notes after the experiment than we do after a visit to the circus."); others could be adequately dealt with by informal notes (again the Nuffield Guide is worth quoting: "... a pupil's own notes, written informally like the notebooks of most research men, will keep him nearer to real science than formal writing up of each experiment"), and others again could be given detailed reports, although this would only be appropriate in the later school years.

(viii) Maintenance of interest and the provision of personal experience of science.

These two aims have been coupled together because they are very closely related. In the above-cited article, Boulind asserts that "Perhaps the primary justification for practical work is, in fact, the arousing and maintenance of interest." This interest can be aroused and maintained by letting the pupil feel that he is experiencing science at first hand - that he is "a scientist for the day", in the memorable phrase of the Nuffield Guide (I). He will be especially able to enjoy this feeling if he is not given too much advance information concerning what is to come out of the experiment he is engaged in. There will have to be a limit set to what the pupil can discover for himself in this way; moreover, a good deal of intelligent stage-setting will have to be done by the teacher if this kind of teaching is to succeed. Those responsible for devising new teaching approaches which emphasize experimentation by the pupil (e.g. the Nuffield Physics Project) appear to be at least as well aware as anyone else of the limitations inherent in the method; there seems to be no need for the concern expressed by Warren (1965), who regards the emphasis placed upon pupils "finding things out for themselves" as "very disturbing". (In this connection see also Helm 1966).

A particular advantage of this "personal experience of science" which links up with "learning facts and principles"

/has.....

has been neatly summarized by Nedelsky (1965): "Laboratory work... is practically irreplaceable...for helping the student understand the relation between theoretical 'book' science and real phenomena". Here we have perhaps the most convincing single reason for extensive practical work by the pupil.

2.8.3 The need for practical work

A number of possible aims of practical work in physics have been considered above. It has been shown that all of these aims can be served by practical work to a greater or lesser degree, and that some in particular are uniquely and excellently served by it. All this adds up to a strong argument for a great deal of practical work to be done by school pupils as part of a carefully designed syllabus, in which experiments have been included which will serve some or other of the aims discussed above. In the main section of this thesis, attention will be given to the recommendation of experiments suited to form part of a course in which experimental and "blackboard" work (one hesitates to call the latter "theory") are closely linked.

2.8.4 String and sealing-wax apparatus

A word should be added here about the use of "string and sealing-wax" apparatus in school physics.

From Maxwell's time to the present, the praises of this approach to elementary practical work have been sung by eminent authorities. In favour of it one must say at once that there is more to it than the mere making of a virtue out of necessity. As Maxwell (quoted in Tricker 1967) has put it: "The simpler the materials of an illustrative experiment and the more familiar they are to the student, the more thoroughly is he likely to acquire the idea which it is meant to illustrate. The educational value of such experiments is often inversely proportional to the complexity of the apparatus. The student who uses

/homemade.....

homemade apparatus, which is always going wrong, often learns more than one who has the use of carefully adjusted instruments, to which he is apt to trust and which he does not take to pieces." However, both pupils and teachers sometimes grumble about simple apparatus on the grounds that it is not accurate enough or that it is unnecessarily crude. Since the use of such apparatus is a prominent feature of newly-developed courses in school physics, these criticisms are worthy of some consideration. First one must insist that the simplicity of a given experimental set-up becomes self-defeating if it produces a sense of irritation or frustration on the part of the experimenter. The author's first acquaintance with Newton's rings was made with the aid of a sodium flame as light source and a travelling microscope which was unable to track across the field of view in a straight line. In order to obtain a reasonably well-illuminated interference pattern, he had to move the flame as close as possible to the microscope. The result was that instead of experiencing the thrill of studying a beautiful and hitherto unfamiliar phenomenon, his major concern was with trying to look down the microscope without singeing his hair. Obviously the occasion was a memorable one, but not quite in the way intended. In this case the unfortunate experimenter would have been much happier if the apparatus provided had been just a little more sophisticated.

If a certain standard of accuracy is expected in the performance of a given experiment, then the apparatus should be capable of yielding up such accuracy without the pupil having to spend too much time and trouble in achieving it. Some of the complaints about inaccuracy may arise from a faulty approach by teacher and/or pupil. If the pupil has been conditioned to believe that the primary purpose of his practical work is to obtain a "good" value for an already known numerical result, then the mere sight of apparatus of obviously simple design is likely to put him off. But let us suppose that his upbringing

/has.....

has been different; that he has grown accustomed to using simple apparatus to make simple discoveries, qualitative or quantitative, to suggesting specific improvements in the design of an experiment or the apparatus used for it, and to forming a clear idea of what accuracy is and of its relative importance in different situations. Then he is unlikely to develop a resistance to "string and sealing-wax" as such, and his appreciation of more elegant and sophisticated forms of apparatus will be all the greater for the experience he has had of simpler forms.

In short, apparatus that is simple in design and construction has a vital role to play in building up a pupil's feeling of what physics is all about, but it can only fulfil this function if given its proper setting by the teacher. Provided that this type of apparatus is not used for purposes for which it is plainly unsuited, no difficulties need arise. The pupil may feel that the experiments he performs at home or in the school laboratory are very far removed from the experiments currently being conducted with apparatus in the mega- or giga-dollar range. The teacher can reassure him by discussing with him some of the reasons why certain modern apparatus costs so much. Such a discussion might prove very fruitful. A pupil curious about the 114,000,000-dollar Stanford linear accelerator or about the even more expensive accelerator projects at present being planned, might learn a good deal from a teacher who took as his starting point the energies attained by the electrons accelerated in the cathode-ray tube of an oscilloscope. In this way a link can be established between the pupil's experience and the activities of present-day professional physicists.

/2.9 .....

2.9 The content of school physics syllabuses

2.9.1 Some general considerations Of the making of school syllabuses in physics there is no end. This is by no means to be deplored. There is certainly no ideal school physics syllabus, suitable for all pupils for all time, and in the absence of such a thing, there will always be room for discussion of what should be taught at school under the heading of physics.

A variety of attitudes towards their task is to be found among planners of syllabuses. There are some who feel that an introductory physics course should be closely related to the everyday life of the pupils, if necessary to the exclusion of "modern physics"; there are others who, as we have seen, feel that a twentieth-century physics course should be orientated "in the general direction of modern physics". Others again believe that school courses in the sciences should be influenced by the technological needs of the country and should therefore be the basis for fairly specific technological training. Still others, mindful of the needs of the university student, will want to apply a principle enunciated by Mendoza (1965), viz. "...that subjects which are touched on at school, however superficially, are later accepted by university students as 'easy' or at least as approachable."

Whichever category they fall into, conscientious planners of syllabuses will want to consider whether or not their proposed choice of content makes possible the realization of the "behavioural objectives" (cf. Section 2.1.2) they have in mind. As Nedelsky (1965) has emphasized, this means choosing topics not only for their intrinsic importance but also for their "illustrative value". Thus one may want to include a brief treatment of a "naive" particle model of light in an introductory physics syllabus, (cf. Section 8.2), not because this

/adds.....

adds anything vital to the content, but because it illustrates the need to modify a theory if certain of its predictions clash with experiment. Consideration of the illustrative value of a topic can be useful in reducing the content of a syllabus, as Nedelsky has also pointed out. When a number of topics seem to have equal claims for inclusion on the grounds of intrinsic importance, some may be found to have more illustrative value than others, and this can justify the exclusion of the less illustrative ones. Application of these principles of topic selection still leaves a good deal of scope for variety in the resulting syllabuses and, it should be added, for possible error - the finished syllabus may after all not permit wholly successful realization of the desired objectives. For these reasons, no syllabus, however august its sponsorship, should be treated as a sacred document, immune from criticism or amendment.

2.9.2 Recent developments in syllabus-making

It has been implied in the preceding section that introductory physics syllabuses drawn up by different teams are likely to be different, and that this is no cause for concern. Despite this, many of those concerned with syllabus-making from the inside cannot help feeling that there ought to be some measure of general agreement concerning what should be taught in school physics. And indeed something like this does now seem to be emerging. Looking at syllabuses that have originated outside South Africa in recent years, we find a number of common features. Consider the recommendations for the "Essentials for a Minimum Course in Physics" (Schoene 1965) based on the report of the special committee of the OEEC (cf. Section 2.3 above), and the syllabuses drawn up by the Science Masters' Association and Association of Women Science Teachers of Great Britain (1961), the Nuffield Physics Project (Nuffield

/Guide.....

Guide I), the Physical Science Study Committee (PSSC 1960), the Harvard Physics Project (Holton 1967), a committee set up by the Nuclear Research Foundation of the University of Sydney (Parry 1964), and a special committee appointed by the State of Victoria, Australia (Caro 1963). (In connection with the last, it should be noted that the Leaving Certificate and Matriculation Physics Course for Victoria is substantially identical with the PSSC Physics course.) All these syllabuses are distinguished by an emphasis on fundamentals - the basic principles of mechanics, wave motion and electricity receive a good deal of attention - and by a generous though not disproportionate treatment of modern physics. Also, it has been accepted by those drafting these syllabuses that if an increased amount of modern material is included, some traditional material will have to make way for it. Consequently topics whose inclusion has long been taken for granted have had to be omitted from these new syllabuses.

It is this last point which has aroused and still continues to arouse most controversy. The arguments against have been succintly expressed by Zemansky (1962): "When an entire subject, such as geometrical optics, or thermodynamics, or acoustics is skipped, the student is presented with a distorted view of what physics is. Such an omission might conceivably affect the student's future career....Why should it be regarded as legitimate to skip the subjects of statics and hydrostatics in order to concentrate on atomic physics at the introductory level? Why is the world of the unseen more important than the familiar world of ladders leaning against walls, hammocks stretched between two trees, or boats floating on water? I can see no justification for leaving out whole branches of physics that infringe directly on the students' everyday life." In fact, Zemansky was thinking specifically of introductory college physics when he made these remarks, and his views regarding

/physics.....

physics as taught at school might conceivably be quite different. We may nevertheless take the above-quoted passage as a convenient statement of the argument against deletion of familiar and well-loved topics from school physics in favour of the inclusion of more modern physics, since it is typical of the views of teachers devoted to the established tradition. Zemansky makes two main points:-

(a) omission of certain topics gives the pupil a distorted view of physics; and

(b) in our teaching, we should not neglect the familiar world in favour of the unseen world.

The short answer to (a) is that any view of physics based on an acquaintance with it acquired only at school is bound to be distorted. To present physics to a pupil in such a way that his impression of it is coloured by what some present-day physicists are doing is unlikely to be very harmful. It is certainly unlikely to affect his future career, although Zemansky might well be right in implying that this could happen to a college or university student. The assumption that seems to lie behind (b) is that a pupil should first achieve a thorough understanding of the physics of the familiar world before attempting to probe the unseen world. In the abstract, this may seem a sound principle on which to base one's teaching of physics, but it is difficult to uphold it in the face of certain difficulties encountered in the "familiar world" of education. How many pupils are actually excited by the prospect of learning about those aspects of the familiar world which are "covered" by such topics as statics, hydrostatics, thermal expansion and calorimetry (especially specific heats)? How many teachers are there who are capable of presenting this material in an exciting way so as to "give pupils great insight into science"? (Rogers's phrase.) One suspects that there are very few in either case.

/On.....

On the other hand, some pupils are excited at the prospect of learning something about the unseen world, and many of their teachers are both excited at the prospect of teaching this material and capable of transmitting that excitement. In short, the present-day educational climate does to some extent favour the inclusion of modern physics in the school syllabus, and it is realistic to take this into account. In Section 2.7 we have also argued in favour of the inclusion of some modern material on the grounds that this is a way of compelling the teacher to keep up with his subject. We conclude that we must come down on the side of those who have designed the new syllabuses listed above, at any rate as far as their attitude towards modern physics is concerned.

2.9.3 What should we exclude?

What topics should make way for the new material? Undoubtedly statics, hydrostatics, thermal expansion and calorimetry must be favourably considered for exclusion. On the subject of hydrostatics, Professor Rogers has this to say (Nuffield Guide I): "Hydrostatics is not undesirable; but it is unnecessary. Studies of the Principles of Archimedes are not important as preparation for other topics in physics; and as they are taught they do not, I think, give pupils great insight into science. In many a school, pupils weigh things in air and water till it is almost a mechanical habit." We need not add anything to this summing-up. As for statics, its chief value in an introductory course is as a means of illustrating the fact that force vectors add up to zero in equilibrium situations. Of course it also embraces the Law of the Lever, which is very well suited to investigation in the elementary laboratory and can be used as an early illustration of a physical law. But the tedious elaboration of these basic principles in the solution of problems concerning gates, ladders and doors can safely be omitted.

/The.....

The qualitative facts of thermal expansion can easily be dealt with in the primary school, and it is good to see that they have now been relegated to this sphere in South African schools. Quantitative work on thermal expansion with its definitions of the various kinds of expansion coefficient and the use of these in "standard" problems almost unrivalled in their lack of point, would seem to serve no useful purpose in a school physics course. Thermal expansion of gases can be dealt with in the context of thermometry and the gas laws. The basic ideas of calorimetry - "heat measurement" - are a necessary prerequisite for the quantitative treatment of energy conservation and can therefore scarcely be avoided. Likewise it is valuable for a pupil to know of the existence of latent heats of change of phase, because this helps to fill in an important detail of his picture of matter. But the value of extended laboratory work in calorimetry, involving the measurement of the specific heats/water equivalents/thermal capacities of various materials and objects by the method of mixtures, is doubtful in the extreme. There would be some point in finding the values of specific heats of substances if these were then converted into "atomic" or molar thermal capacities and the results were compared with theoretical prediction. But this extension of the subject cannot be accommodated in a school course, although at least one South African textbook has tried half-heartedly to point the way by including a brief reference to the Law of Dulong and Petit (Gordon and Neser, no date). As no indication is given by the authors of the significance of this law and the deviations from it, one cannot help wondering why they bothered to include this material at all. Perhpas they imagined that conscientious pupils would store up this information and patiently await its clarification at some later date.

/Among.....

Among other topics nowadays regarded as "expendable" are geometrical optics and sound. "Expendable" should not in these instances be equated with "valueless", however. As one group of syllabus authors puts it, "Sound is one of our two means of communication and as such deserves a place in the syllabus." (Joint Committee of the South African Chemical Institute and Institute of Physics 1965). A good deal of fun and genuine enlightenment can be derived from an elementary study of sound, and for this reason it is a pity to have to exclude it from a school syllabus. On the other hand it is not essential to the development of other topics now regarded as important, so that it can be left out without doing <u>much</u> harm. (One wonders a little about this, of course; a newspaper reporter once informed his readers that a research group at Rhodes University were constructing equipment to receive sound waves from the sun. Perhaps a smattering of knowledge about sound might have prevented him from perpetrating this gaffe.) It would be difficult to justify the total exclusion of geometrical optics from introductory courses. It appears to be generally agreed nowadays that such courses should include a study of light, leading on to a preliminary investigation of its nature. An investigation of this kind has to take into account the known behaviour of light, and the fundamental principles of geometrical optics embody important features of that behaviour. Significantly, both the PSSC and the Nuffield physics courses allow some space for this topic. Admittedly one should not overdo this indulgence towards it. To quote the abovementioned group of syllabus authors again, "There is little basic physics in [geometrical optics] and, in any case, the quantitative aspect leads to formulae which are not exact enough to be used in instruments." In other words, little purpose is served by <u>theoretical</u> quantitative treatment of reflection from curved mirrors, image formation by lenses or

/even.....

even deviation of light by a prism. (Dispersion by a prism and its use in spectrometers, which <u>are</u> important, do not require reinforcement by derivation of the formula

$$\mu = \frac{\sin \frac{(A+Dmin)}{2}}{\sin \frac{A}{2}} \quad .)$$

2.9.4 <u>Should technological material be included in physics syllabuses?</u>

The inclusion of technological material is against the present-day trend in the planning of physics syllabuses, at any rate in the Western World. Perhaps this is on account of the excessive attention given to technology in earlier syllabuses. The tendency to exclude technology has been deplored by some critics, notably by Elton (1967), who in criticizing the absence of technology and the applications of science from the Nuffield Physics Course, has stressed that "it is the mutual interaction of [science and technology] which fertilizes both... neither can flourish unless the community as a whole accepts their importance and for that reason 'Science for All' [i.e. a general introductory course in science] must give an appreciation of both science and technology to all those who will not take the subject any further". Accordingly he suggests that "we should not just teach descriptive technology...we should teach a good deal of it". There is clearly some merit in this suggestion, but it raises a number of questions. Should "descriptive technology" be taught as <u>technology</u>, i.e. should one segregate a portion of the science syllabus from the rest and call it "technology"? Exactly what should be taught under this heading? What principles should be followed in trying to teach it - should it be taught as "applied science", with emphasis on the fact that it is the application of scientific principles which is under discussion, or should the

/emphasis.....

emphasis be placed on engineering aspects? The answers to these and related questions will have to be found if Elton's suggestion is to be implemented, but most of them lie outside the scope of the present work, which is concerned with the teaching of physics. What is more directly relevant is the question of whether time which might otherwise have been devoted to teaching for increased understanding of physics should instead be given over to the teaching of technology. Elton is firmly of the opinion that it should; purists in the field of physics education will maintain that it should not. Many will feel that it is sensible to apply the kind of compromise solution that some have applied for many years, namely to add interest to the teaching of certain topics in physics by the discussion of exciting technological applications of physical principles. (Cf. Section 2.7). In the right hands, this can lead to genuine enhancement of physics teaching while at the same time it can help pupils to learn some basic technology.

There is one specifically technological topic that could with advantage be included in a school physics course, viz. electronics. Many boys have an interest in and aptitude for electronics; many of the black boxes currently used by the physicist are wholly or partly "electronic" in content, so that there is an obvious utilitarian link between physics and electronics; and it is possible to base one's teaching of electronics on underlying physical principles. This last point is important, for it can reasonably be argued that if technology is to be included in a course labelled "physics" or "physical science", the point of departure for it should always be the underlying scientific principles. In the past, the incorporation of technological applications into a science syllabus has sometimes deteriorated into a "descriptive technology" of the worst kind, with pupils being expected in examinations to give detailed descriptions, preferably accompanied by annotated sketches, of various contrivances without having

/to.....

to show any scientific understanding. This suggests that if certain specific concessions to technology (e.g. the inclusion of elementary electronics) are to be made in a physics syllabus, the method of approach to the chosen topics must be carefully delineated. We shall see below that the Common Basic Syllabus includes an optional section on electronics. In Chapter 9 of this thesis, consideration will be given to the teaching of this material.

2.9.5 Past and present syllabuses in South Africa

Our survey of recent trends in syllabus-making outside South Africa has shown that by and large the new syllabuses are characterised by an emphasis on fundamentals and on the inclusion of "modern physics", the latter at the expense of the omission of certain traditional material, if necessary. We now turn to an examination of the South African scene, both before and after the introduction of new syllabuses in this country. First, we look at the syllabuses in physics and physical science of the Joint Matriculation Board that were in use until 1965, when new syllabuses approved by the board first became the subject of examinations. At that time, only a few pupils (between a hundred and a hundred-and-fifty) from the private schools entered for the J.M.B. examination in pure physics. [The number of pupils studying pure physics at government schools (i.e. those responsible to Provincial education departments) was far fewer, as only one or two schools offered it as a subject .] The influence exerted by this syllabus was thus small, (although it should not be forgotten that some of the best university entrants in the country came from the group of pupils affected by it), but it is nonetheless worth considering as an indication of the prevailing attitudes of the time. It begins with a two-page preamble, most of which is taken up with a discussion of electrical units and

/their.....

their definition. Quotation of a few sentences will serve to illustrate the flavour of this section: "Electrical quantities are measured by the effects they produce and in particular by the electromagnetic, electrothermal and electrochemical effects. The use of any of these introduces difficulties when the attempt is made to relate them in a simple and obvious manner with absolute units. Thus the use of the electromagnetic effect involves the rather difficult concept of the unit magnetic pole, while the use of the electrochemical effect relies upon the value of the Faraday and its relation to the fundamental units of the c.g.s. system." In the paragraph following this quotation, the ampere is defined (in the 1960's!) as being "one-tenth of the absolute electromagnetic unit of current, namely, that current which produces a magnetic field of one dyne per unit pole at the centre of a circle one centimetre in radius when it flows through an arc of the circle one centimetre long," and definitions are also given of other electrical units. Some remarks on the subject of "international" units follow. The purpose of the inclusion of this section is nowhere made clear. No recommendation is even made concerning which units should be taught, absolute or international. The contents of the syllabus then follow. In summary, these are the measurement of area, length and volume; Hooke's Law; surface tension; statics of a particle and a rigid body; the fundamentals of dynamics; work, energy and power; elementary hydrostatics; temperature and thermometry (including maximum and minimum thermometers); linear and cubical thermal expansion, including coefficients of expansion; the gas laws and absolute temperature; changes of phase, including saturated and unsaturated vapours; calorimetry; the nature of heat and determination of "J"; transport of heat; rectilinear propagation of light; photometry; laws of reflection and refraction; thin lens

/formulae.....

formulae; optical instruments, including the eye and the correction of its defects; dispersion of light and spectrum formation; general characteristics of sound; velocity of sound; standing waves on strings and in a resonance tube; elementary magnetism; magnetic moment; field plotting; the earth's magnetic field; magnetic effects of an electric current, including field of currents in circular coil, solenoid and long straight wire; the tangent galvanometer; potential difference, Ohm's law; simple problems on current, voltage, resistance and energy dissipation; Wheatstone bridge; heating effect of an electric current; qualitative treatment of force on a current-carrying conductor in a magnetic field; moving coil instruments; laws of electrolysis; Leclanché cell and lead accumulator; electromagnetic induction; A.C. and D.C. dynamos; the transformer.

No one can deny that this syllabus included a good deal of solid, worthwhile physics. Unfortunately it also has other rather less praiseworthy features. Perhpas the most obvious of these is the fact that it might have been drafted in 1890 - none of the developments of the last eighty years are mentioned at all. (Apart from the sections on electricity and magnetism, the syllabus bears a close resemblance to that offered by Benjamin Worster at his Academy in London in 1722 (Bishop 1961).) The kinetic theory of matter is not even mentioned, and one cannot assume that it was supposed to be implied. Another significant omission is that of time and its measurement, and of course the syllabus is also noteworthy for its inclusion of various "expendable" topics discussed earlier in this section. The whole thing seems to have been constructed on the principle of "what was good enough for my grandfather is good enough for me". Despite its virtues, it is not a syllabus calculated to inspire teacher or pupil in the mid-twentieth century.

/After.....

After this, one turns to the physics part of the corresponding physical science syllabus with no high hopes of finding any redeeming features there. Nor are there any. This material consists simply of a watered-down version of tne above syllabus. Its mechanics section is made up of the following topics: "Elementary notions of force. Balancing of parallel forces; the moment of a force". Similarly the section on electricity has been pruned of most of its potentially worthwhile features.

The corresponding syllabuses formerly prescribed for the Cape Senior Certificate (Department of Education, Cape of Good Hope 1961) present a somewhat more inviting appearance. In the case of pure physics, the entire contents of the Joint Matriculation Board syllabus are included, with a number of valuable additions, e.g. the electromagnetic spectrum, cathode rays, thermionic emission, the photo-electric effect and X-rays. But this syllabus catered for an even smaller number of candidates than did the J.M.B. physics syllabus - only one or two government schools in the Cape Province offer pure physics as a subject - so its influence was negligible. The physics portion of the Cape physical science syllabus did to its advantage include some of tne additional topics contained in the pure physics syllabus, but on the otner hand contained no mechanics. Naturally this was a crippling omission and conscientious teachers have been severely handicapped by it in their attempts to present a coherent picture of the subject. They and their fellow teachers in the other provinces have been further handicapped by certain other difficulties, viz. tne textbooks recommended and tne type of examination question tnat was set. The tnen typical South African textbook in physical science, such as the above-cited work by Gordon and Neser, was seriously misleading on most matters of importance.

/Despite.....

Despite this, it was treated by most teachers as having scriptural authority. Examination questions exemplified all of the faults exposed by Prof. Rogers in his discussion of examination question-setting (Rogers 1962, Brown, Clarke and Tiomno 1964, Nuffield Guide I). Candidates (and their teachers!) could confidently expect to be asked to describe an experiment to verify this law or determine that specific heat, to draw labelled sketches of the maximum-and-minimum thermometer, the vacuum flask or the Leclanché cell, or to work out problems which required them to plug in data into some remembered formula. Of testing for understanding of the subject there was little or none. Thus even teachers attempting to make the best of a bad job by teaching for understanding within the confines of the then existing syllabus were thwarted by the style of the final examination.

It was in this prevailing atmosphere that a conference of science teachers from South African private schools resolved in December 1960 to ask for the introduction of modernized syllabuses. Considerably to their surprise they found that they had been anticipated by the Joint Matriculation Board, who at that time already had a new draft syllabus for pure chemistry in their hands and had agreed in principle to the compilation of new syllabuses in physics, physical science and biology as well. In pursuance of this objective they asked Prof. J.A. Gledhill and the present writer to draft a new syllabus for pure physics and the physics portion of a new syllabus for physical science. These authors completed their work in 1961, the new syllabuses were duly published and were examined for the first time at the end of 1965 (Joint Matriculation Board 1963). The syllabus authors were strongly influenced by recent developments in the U.S.A. and Great Britain and sought to incorporate some of the best features of these in the new

/syllabuses.

syllabuses. They tried also to improve the presentation of some of the traditional material by specifying a more quantitative treatment in certain places (this was especially true of electricity and electromagnetism) and listing the syllabus headings in such a way as to suggest a more logical teaching approach than might previously have been used. At the same time they did not eliminate a great deal of the material formerly included, with the result that while the new syllabuses were commended on account of their inclusion of modern material, they were also criticized (in the case of the physical science syllabus, sharply criticized) on the grounds of being overcrowded. Nevertheless, science teachers tackled the task of teaching the "new look" syllabuses with a will and a number have testified to enjoying the experience. In the meantime, the Joint Matriculation Board had taken certain other decisions of great importance to South African science teaching:-

(i) It accepted that when a new syllabus was to be implemented, mere publication of the syllabus was not enough; this had to be supported by the publication of a teachers' guide setting forth detailed information about the content of the syllabus and, where necessary, recommending possible teaching approaches. The present writer was asked in 1962 to undertake the writing of the teachers' guide for the new physics syllabus and to contribute to the teachers' guide for physical science.

(ii) It accepted the principle of regular revision of syllabuses, so that the danger of fossilization could be removed.

(iii) It agreed to the adoption of "core" syllabuses which were to be common to all the examining authorities in the country. These authorities were to be allowed to add a certain amount of extra material to the "common basic syllabus" in accordance with their inclinations.

/The.....

The present writer has been told by a member of the Joint Matriculation Board that it now disavows any intention of recommending possible teaching approaches (Gledhill 1968). However, when it originally commissioned its "teachers' guides", the authors of the guides were given clearly to understand that they should outline possible teaching approaches. For example, in a letter to the author from the editor of the guide to the new Common Basic Syllabus in physics, it was stated that a "Representative Committee of the Joint Matriculation Board" had "requested that the guides should clearly indicate...possible teaching approaches to a number of topics especially those unfamiliar to teachers". There were very good reasons why the Board should have made such a request. By and large, South African science teachers needed (and still need) guidance in the teaching not only of topics unfamiliar to them but also of those with which they had long been familiar. It would have been imprudent, to say the least, to have inflicted a new syllabus upon them without at the same time offering them some assistance in planning the teaching of its "new" topics and taking the opportunity of correcting serious errors in their approaches to "familiar" topics. To its credit, the Joint Matriculation Board fully appreciated this at the time, whatever doubts it may have had on the subject in later years.

In addition, the South African Institute of Physics commissioned Mr. (now Professor) J.W. Brommert of the University of the Witwatersrand to write a physics textbook for South African schools. Although this book is unlikely to sell widely amongst school pupils, most of whom will be taking physical science or biology if they take science at all, it could be a most useful reference work for South African science teachers.

Acting upon the third of the above-listed decisions of the Joint Matriculation Board, the South African Chemical Institute and the South African Institute of Physics set up committees to

/draft.....

draft new common basic syllabuses in physics, chemistry and physical science. These syllabuses were accepted in due course by the Board, and other examining authorities have since adopted their own slightly-enlarged versions of them. Thus the present position in South Africa is that the older syllabuses which, as has been seen, were in greater or lesser degree unsatisfactory, have now all been replaced. In view of the crucial importance of the new common basic syllabuses in physics and physical science, they must be closely examined here. Their aims and the guiding principles followed in their drafting are set forth in a preamble which is substantially the same for both syllabuses, and which deserves extended quotation:

"(i) The objective of the syllabus should be to provide the School-leaving child with a picture of Physical Science [or Physics] which reflects the status of our subject in our contemporary civilization, and in so doing, to contribute to the provision of useful and responsible future citizens of this country - of civilized people as distinct from technical barbarians.

"(ii) It should be recognised that school-leaving children fall into two categories - the <u>majority</u> who will receive no further instruction in Physical Science [Physics], and the <u>few</u> who will continue with their studies in this subject at university. The former require a rounded-off picture of physical science - [Physics] whereas the latter must be provided with a basis for further study.

"(iii) It should be recognised that contemporary civilisation sees two main facets of physical science [physics]; namely physical science as a <u>science</u> ..... and physical science as a technology..... Both facets should be adequately dealt with."

"(iv) The teaching should be visualised as a structure consisting of an orderly framework of laws and concepts,

/erected.....

erected by the tools of mathematics, on which useful facts are co-ordinated.....

"(v) The history [of the subject] should be drawn on only where the historical development is an aid to understanding.

"(vi) The syllabus should contain material from the frontiers of modern physics and chemistry. Illustrations should be drawn from everyday life..... Where possible, material of this nature should flow from scientific principles and concepts; but where this is not possible it should not be considered as an insurpassable obstacle, or as an excuse for omitting such material. Instead qualitative, and particularly visual (pictorial) methods should be resorted to.

"(vii) What is taught at high-school level must unavoidably be incomplete; it should, however, never be wrong. Furthermore irrelevant, archaic and useless material must be rigidly excluded.

"(viii) Physical science should be taught as a coherent, unified subject..... A division into physics and chemistry is bound to cause confusion and misconceptions." [For obvious reasons, this point does not appear in the preamble to the pure physics syllabus.]

"(ix) The factual content should always maintain a correct perspective in relation to physical science as a whole. Overemphasis of irrelevant details should be avoided at all costs.

"(x) At all times principles should be reduced to the most general possible form; it should be emphasized that principles cover a wide range of phenomena."

It will be noticed that points (iv), (v), (vi), (vii) and (ix) are in direct agreement with certain of the views advocated in the preceding pages. Of the statement as a whole it can be said that the principles it enumerates are an adequate foundation on which to base a science syllabus. The contents

/of.....

of the syllabuses themselves represent a carefully thought-out attempt to follow these principles, and if properly taught could provide an excellent introduction to physics or physical science. (Although the only explicitly-stated objective (c.f. pt. (i)) is scarcely a clear-cut behavioural objective, the syllabuses are in fact so designed as to make possible the attainment of the desired objectives of knowledge, understanding and ability to learn.)

In some ways, the very first section, common to both syllabuses, is the most controversial. Entitled "Science and Technology", it includes among its list of subjects "The objectives of science and technology; the elements of scientific method; the scientific meaning of facts, concepts, laws, models, theories and hypotheses." It is actually included within double square brackets, which indicates that it should "normally be dealt with before Stds IX and X", i.e. before the last two school years. Although the authors of the syllabuses do not explicitly say that topics should be dealt with in the order in which they are listed, the general design of the syllabuses and the accompanying comments do seem to suggest that this was the intention. Assuming this to be so, one must doubt the wisdom of trying to teach "the elements of scientific method" <u>as such</u> to pupils whose level of scientific sophistication is still distinctly low. Most professional scientists would probably agree that they learnt "the elements of scientific method" along with learning about science itself - that as their scientific experience grew they realised more and more the significance of what they were doing in the laboratory and of scientific work generally. The commentary accompanying the Physical Science Syllabus (Joint Committee of the South African Chemical Institute and Institute of Physics, 1965) says: "Scientific method must be carefully, although briefly, explained. The concept of a scientific model must be particularly stressed."

/This.....

This seems extraordinarily unrealistic. How can one possibly explain scientific method "carefully, although briefly", to a school pupil? And would it not be better if "the concept of a scientific model" could be progressively understood by a pupil through his experience of a great many such models, than through an isolated discussion of models in general?

Fortunately, although the tone of this section of the syllabuses suggests that the fear of those responsible were not firmly enough anchored to the ground, the contents of the remaining sections restore one's confidence. Section 2 of both syllabuses is concerned with "Classical Mechanics", under which heading a proper foundation is provided for the rest of the course. The physical science syllabus here includes vectors and their addition, resolution of vectors into components, the fundamental principles of dynamics, including conservation of momentum, Newton's Law of Gravitation, work, energy and power, conservation of energy and the relation between heat and work. To this list the physics syllabus adds the study of rectilinear motion, centripetal force and circular orbital motion, and the First Law of Thermodynamics. (Incidentally it is assumed by those drafting the syllabus that pupils will have met <u>some</u> calorimetry before reaching the senior standards.) Section 3 is "The Particle Structure of Matter" in both syllabuses, and deals with Brownian motion, the existence of electrons, protons, neutrons, atoms, molecules and ions, and the kinetic picture of matter. Its inclusion was evidently inspired by the report of the OEEC Committee cited in Section 2.3 of this thesis. Some might question its placing at this point, but with careful teaching this could be justified. Section 4 of the physics syllabus is taken up with "qualitative description of Van der Waals forces" as a prelude to section 5, which deals in some detail with the particle picture of gases, solids and liquids. Under "gases", we find elementary kinetic theory of gases,

/kinetic.....

kinetic theory interpretation of temperature, Kelvin temperature scale, thermal equilibrium, ideal gas laws, heat capacity and Dalton's Law of Partial Pressures. Under "solids" the topics listed range from crystallinity to lattice energy, crystallization and melting. "Liquids" includes only the liquid state and surface tension. In the physical science syllabus, all this material appears at a much later stage in what might be called (though not officially) the "chemistry" section of this syllabus; its introduction is preceded by treatment of the atomic nucleus, the atom and molecules. Indeed, with the exception of the sub-section on gases, this would appear to be the more logical place to have it. One might have this reservation about the inclusion of this material: that in syllabuses by no means underweight, there might be little time available for the teacher to give anything but a brief, "telling-the-facts" treatment of these topics.

"Electricity and Magnetism" constitutes Section 6 of the physics syllabus; it is Section 4 of the physical science syllabus, coming immediately after the particle structure of matter. This section follows the lead of the corresponding section of the immediately preceding Joint Matriculation Board syllabuses in outlining as full a treatment of the subject as seems compatible with pupils' level of understanding. It begins with a consideration of the inverse square law, conservation of charge, the electric field, potential and potential difference. In the words of the drafting committee, "These are the concepts which will be required later and without which no real understanding of the behaviour of electricity is possible." With this statement one can only agree, while at the same time regretting that it has taken so long for the principle it asserts to achieve official acceptance. The determination of the electronic charge by Millikan's method is

/included.....

included in this sub-section on "electrostatics." A preliminary treatment of "current-electricity" follows, with consideration being given to various modes of charge transport. Qualitative treatment of the magnetic effects of currents precedes quantitative electromagnetism, in the course of which the unit of current is properly defined and "magnetic induction" ($\vec{B}$) is treated as the fundamental magnetic field quantity. Resistance and EMF are introduced, together with some elementary circuit theory. Electromagnetic induction is approached quantitatively via Faraday's Law. It is heartening to see that most of what is included in Section 6 of the physics syllabus finds its way into Section 4 of the physical science syllabus as well.

The fundamentals of vibration and wave motion come next in both syllabuses. The treatment outlined takes the pupil as far as interference, diffraction and polarization, all of them treated qualitatively. The stage is now set for the introduction of electromagnetic waves, which are actually considered <u>before</u> light as such is brought in. When it is, the emphasis is placed on a study of the nature of light; at this stage evidence is advanced for light being a form of wave motion, culminating in the revelation that it consists of electromagnetic waves. The next section, "Geometrical Optics", basically dealing with elementary geometrical optics, includes a brief treatment of dispersion and of the measurement of the velocity of light. Sound is also included in both syllabuses; we have already quoted the drafting committee's defence of its inclusion. Those aspects singled out for quantitative treatment are standing waves on strings and in pipes, and the resonance of air columns and the determination of wavelength. Elementary vacuum tube electronics, from the diode and triode to the modulation, transmission and reception of radio waves, forms section 11 of the pure physics syllabus; it is listed also as

/section.....

section 9 of the physical science syllabus but appears there in single square brackets, which means that it is a "suggested extension of the basic syllabus." The drafting committee point out that although there is little basic physics in this section, "there is an almost irresistible demand for some understanding of electronics and this should be seriously considered." This is a realistic approach, and one can only regret that in the mid-nineteen-sixties it was not thought necessary to include some reference to solid-state electronics as well.

In "Elementary Treatment of the Nature and Properties of Electrons and Radiations" (section 12 in pure physics, Section 10 in physical science), the evidence for the particle-like and wave-like properties of electrons and radiations is surveyed; this includes electron diffraction and the photo-electric effect, the latter leading to a consideration of energy quantization. The inclusion of this section is to be welcomed, although it is clear that its presentation will present a considerable challenge to teachers - perhaps the biggest of any in the new syllabuses. In the physics syllabus, a section entitled "The Atom" follows next, although the one after that ("The Atomic Nucleus") leads off with "The evidence for the nuclear atom as provided by Rutherford's experiments on the scattering of alpha-particles". This seems odd, as the bulk of the section on "The Atom" deals with atomic models, atomic energy levels and the electronic structure of atoms. As the corresponding section in the physical science syllabus comes <u>after</u> "The Atomic Nucleus", as one might expect, one is left wondering why the order of these sections was inverted in the physics syllabus. The section on the nucleus starts, as has been seen, with Rutherford's alpha-particle scattering experiments and moves on to the sizes of atom and nucleus, the constituents of nuclei, nuclear reactions (including fission and fusion), $E = mc^2$,

/particle.....

particle accelerators and artificial radioactivity.

Those drafting these Common Basic Syllabuses made no specific recommendations concerning accompanying practical work. Instead, they included the following directive among the preliminary notes to the syllabus: "Physical science is an experimental science. All examining bodies must submit with their syllabus the minimum illustrative practical work which they prescribe for Matriculation Exemption purposes." The examining bodies have duly done this, furnishing detailed lists of experiments, some of which are to be done by the teachers as demonstration experiments and the remainder by the pupils themselves. These lists show that those responsible for drafting them have taken proper note of the work done by the PSSC and the Nuffield Physics Project in designing suitable experiments that can be done with simple apparatus. Many of the recommended experiments are taken directly from the PSSC and the Nuffield courses. There can be no reasonable objection to this: if experience suggests that certain experiments have obvious pedagogical virtues, by all means let these experiments be adopted as widely as possible. A welcome feature of these lists of experiments is the absence from them of experiments intended to "verify" certain laws or principles. To be sure, some experiments which were formerly designated in this way are still included, but the pretence that they constitute "verifications" of any sort has now been abandoned. In general it can be said of these new "practical syllabuses" that they are adequate for the realization of the aims of school practical work in physics discussed in Section 2.8. It is always easy to suggest additions to lists of this kind, and indeed one would hope that teachers themselves would want to supplement the "official" experiments where they could. But what is there already is certainly capable of giving the pupil an interesting and worthwhile time in the school laboratory. Unfortunately, by no

/means.....

means all South African pupils at present studying the new syllabuses have been given a fair opportunity to carry out the recommended experimental work. In the Cape Province, for example, many schools lack the necessary equipment, and the Provincial Department of Education is apparently not yet able to help them. Some teachers have sought assistance from university science departments; others are not in a position to do so, while still others will no doubt continue to do what they have always done, namely to "teach" it all from the book. The author has been assured by officials of the Cape Education Department that this lamentable state of affairs is only temporary.

The only serious doubt that arises in connection with these syllabuses is whether too much material may not have been packed into a two-year period of study. The experience of the next few years will help to decide this issue.

In helping to implement the syllabuses derived from the new common basic syllabuses, the Joint Matriculation Board has commissioned a new Teachers' Guide for Physical Science. For unavoidable practical reasons, this has had to be produced in two parts; a "physics" part, for which the present writer was partly responsible and which has already appeared in print (Prinsloo et al 1968), and a "chemistry" part, which has not yet appeared. (This division, it will be noted, is in apparent contradiction to the principle quoted above that "A division into physics and chemistry is bound to cause confusion and misconceptions". However, the contradiction is only apparent, because the second, "chemistry" half of the syllabus does in fact follow smoothly and naturally upon the first.) At the same time, the provincial departments of education have set about organizing refresher courses for their teachers as a matter of the highest priority. As textbooks for these courses, small-scale teachers' guides were produced for private

/circulation.....

circulation; the present writer was asked to contribute to the "Guide" used by the Department of Education of the Cape of Good Hope for its refresher courses for standards 9 and 10 during 1967 (Department of Education, Cape of Good Hope 1967).

A most welcome feature of the "new-science-syllabuses" movement is the attention currently being given to the provision of better examination papers - i.e. examinations which constitute better tests of understanding than their predecessors did. Indeed it is essential that attention should be given to this matter, as the best-intentioned efforts at syllabus reform can be nullified by poor examination-setting. Many teachers are aware of this danger and are taking steps to combat it. Thus at a conference of science teachers from the private schools held at Michaelhouse, Natal in January 1968 (Michaelhouse 1968; see also Bosco 1968), a considerable amount of time was spent on this problem, with delegates criticizing questions from past examination papers and individually and collectively devising better ones of their own. As they were assisted in their deliberations by Professor Eric Rogers, the result was a great clarification of ideas; delegates felt that they had fought their way through towards more enlightened examination-question-setting, which would help them in their classrooms and in their appraisal of the work of the official examiners.

As a footnote to this account of the current situation in South African science teaching, it can be mentioned that the various educational authorities have drawn up and begun to implement science syllabuses for standards Six, Seven and Eight (junior secondary school) and standards Three, Four and Five (primary school). The syllabus for standards Six, Seven and Eight is intended as an introduction to the syllabuses for standards 9 and 10; the primary school syllabus is intended

/to.....

to provide a qualitative introduction to the physical sciences and biology. The new primary and junior syllabuses that the author has seen do show that their planners were keen to ensure that pupils would be adequately prepared for senior school science. The planners are obviously aware of the work of the Nuffield Physics Project, which has exercised a beneficial influence on them. There is a welcome emphasis in these syllabuses on experimental work to be done by the pupils themselves with the simplest possible apparatus, and an attempt is made to familiarize pupils with the concepts of the microscopic world at an early stage. Unfortunately, perhaps because the work was done in too much haste and because it was not thought fit to seek the advice of university scientists in drafting these syllabuses, there are signs of carelessness in matters of detail and of failure to consider fully enough the implications of some of the outlines given for certain topics. Some of these difficulties will be discussed in the following chapter of this thesis.

To sum up, we can say that the new science syllabuses for South African secondary schools reflect an outlook which is "modern" in the best sense of the term. In time the content of these syllabuses may be somewhat pared down and possibly modified in other respects as well, but they contain a hard core of basic science which is unlikely to be altered very much in the near future. In the following, main part of this thesis, the present writer will concern himself with what he has chosen to identify as this hard core, (or at any rate the "physics" portion thereof) which will also be found in science syllabuses newly introduced in other countries. His chief aim will be to set out in some detail (a) the points that should be made in any adequate presentation of the topics concerned at an elementary level (b) the background to this presentation with which an "ideal teacher" ought to be acquainted.

/Thus.....

Thus the main section of this thesis will consist, in effect, of a glorified "Teachers' Guide to Physics for Standards Six to Ten", with a much more detailed discussion of certain issues than would ordinarily be found in such a work. Some of the material will consist of an enlarged version of the author's contribution to the various above-mentioned "Teachers' Guides"; some of it will also be based on contributions he was asked to provide for new textbooks. The author likes to think that this material could serve as the basis for a textbook for the "method" training of physics teachers and that even those concerned with university teaching of the subject might find it of some use.

CHAPTER 3: PHYSICS IN THE JUNIOR SECONDARY SCHOOL.

3.1 Introduction

Most of this chapter (Sections 3.4 - 3.19) deals with topics which are currently included in South African junior secondary school syllabuses. Before these topics are taken up, some attention is given to a general problem which actively concerns the secondary school teacher of physics, namely the nature of physical definitions and laws and their correct treatment in the classroom. Accordingly, Sections 3.2 and 3.3 are devoted to the discussion of definitions and laws respectively.

3.2 The nature and importance of definitions

3.2.1 The role of definitions in the teaching of physics

In "traditional" physics teaching, definitions of concepts and units have always played a prominent role. One can confirm this statement very easily by looking at the textbooks on which much of this kind of teaching has been based. For example, the first two pages of the first chapter of the celebrated "Textbook of Physics" by Duncan and Starling (1918) are largely taken up with the statement of "preliminary definitions", and in the rest of the book definitions are drawn to the reader's attention by being printed in black type. (A very reliable test of whether a given textbook is "traditional" or follows a "new approach" is whether or not it gives definitions this kind of prominence.) In many cases these definitions are set down without preamble, as though they were self-explanatory or even self-evident; thus the sum total of the discussion of the concept of mass is the sentence "Mass means quantity of matter."

The practising physicist finds definitions useful be-

/cause.....

cause they enable him to be sure that other physicists know what he is talking about. The teacher's enthusiasm for definitions is understandable for two reasons in particular: (i) It is obviously convenient to have at one's disposal a neat summary of the meaning of a physical concept or unit; (ii) definitions can be memorized by pupils for subsequent regurgitation in an examination. With the first of these reasons one cannot really quarrel. If it is possible to convey something of the meaning of a physical concept in a few well-chosen words, such a "definition" is useful as a reminder of what the concept is, and as such it also links up with the use made of it by the professional physicist. This is an entirely acceptable function of definitions in elementary teaching and one which they will no doubt always fulfil. While approving it we should nevertheless lodge a protest against the practice noted above of beginning rather than ending a discussion of a concept with a definition. More importantly for our present purposes, we must also express the hope that care would be taken to ensure the "correctness" of the definitions used; if the pupil or student is expected to remember a definition, there should at least be something right about what he is trying to commit to memory.

In recent years, certain textbook authors have experienced prickings of conscience on this point and have gone to some trouble to avoid the errors of their predecessors in the field. Somehow the view seems to have gained ground that an "operational definition" might be the kind of definition to aim at and, although one might have doubts about the status of operational definitions as <u>definitions</u> (cf Sections 3.2.3-5), the results of this trend have undoubtedly been beneficial. Few authors and, one hopes, fewer teachers are nowadays inclined to say "mass means quantity of matter" and leave it at that. The definition of EMF quoted in Section

2.2 would be laughed out of court today by any competent physical science teacher, whereas formerly he might have let it pass. Yet there still seems to be room for confusion. In a recently-published school textbook (Stollberg and Hill 1965) we read: "A force can change the shape of an object; it can change the velocity of an object. This is called an operational definition of a force because it describes an activity to make clear what is meant by the word." It is doubtful whether this example of an operational definition would be found entirely acceptable by the originator of the term. Or consider the following sentence from a recently-produced Teachers' Guide (Orange Free State, 1967): "For the purpose of this course [Standards Six to Eight] all definitions are to be limited to operational definitions." The phrase "limited to" seems to imply that at some later stage, operational definitions could be replaced by some other, more acceptable form of definition. This Guide contains other observations on definitions which seem to lend strength to this view, e.g. "The volt is the capacity of a source of electricity to use 1 coulomb of electricity to perform 1 joule of work. At this stage [the volt] is rather a difficult concept, and even then the above definition is only an operational one, not exactly phrased as the scientist would prefer it." (The use of the word "only" in the second of these sentences is significant.) But nowhere in this Guide can one find any example of what a satisfactory definition is, as opposed to an operational one. These quotations suggest that there is room for clarification not only of the term "operational definition" but of the nature of definitions in general. We shall consider this problem below, and attempt to set up criteria for definitions suitable for use in elementary teaching.

/3.2.2.....

3.2.2 The nature of definitions A definition, one might say, is a statement of what something is in terms of something else. Thus we have Bunge's (1967) definition of a definition as "the expression of an identity the stipulation of which enables us to eliminate one concept in favour of others." Presumably any definition we should wish to use in elementary physics should meet this basic specification. To see what it implies, let us consider the following fictitious definition of (linear) momentum: "The momentum of a particle is a measure of its quantity of motion." On the face of it, this definition conforms to the basic requirements: it expresses something (momentum) in terms of something else ("quantity of motion"); if the latter be held to be a concept, then it might seem that we are eliminating one concept in favour of another. The obvious difficulty with this definition is that it will convey little or nothing to a pupil or student who has not been given an exhaustive account of what might be meant by "quantity of motion." In contrast, consider the accepted definition of momentum (Newton's own, rendered in modern terms): "The momentum of a particle is the product of its mass and its velocity." This definition states a precise quantitative relationship between the new concept, momentum, and two previously established concepts, viz. mass and velocity. In the context of Newtonian mechanics, no-one has been able to improve on it as a definition of momentum, and it seems unlikely that anyone ever will. One might still argue that the pupil would not gain a "feel" for momentum from this definition alone; that would only come from some study of certain dynamical situations. On the other hand, even the pupil armed only with this definition would know exactly what to do if he were asked to measure the change in momentum undergone by a body in a collision. By this time he would know how mass

/and.....

and velocity were measured, and all he would have to do in the above situation would be to measure the mass and velocity of the body before and after the collision and then evaluate the "mv" products in accordance with the definition. Thus the "mass times velocity" definition of momentum has a tremendous advantage over the pseudo-definition given earlier: it enables one to use the concept even though one might not have understood its full significance. It tells one, in principle, how to measure momentum.

3.2.3 Operational and constitutive definitions

A definition that seems to lay down a prescription for measurement of a concept is what is ordinarily understood by the term "operational definition". Many twentieth-century physicists have taken the view that the proper definition of a physical concept must be in terms of measurement operations. Cf. Alonso and Finn (1967): "[An] important requirement is that the definitions of physical quantities must be operational, in the sense that they must indicate explicitly or implicitly how to measure the quantity defined." The attitude expressed here is typical of what may be found in many modern textbooks. It would appear to derive directly from that formulated in the following terms by Bridgman (1927) in his "The Logic of Modern Physics": "To find the length of an object, we have to perform certain physical operations. The concept of length is therefore fixed when the operations by which length is measured are fixed: that is, the concept of length involves as much as and nothing more than the set of operations by which length is determined. In general, we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations. If the concept is physical, as of length, the operations are actual physical operations, namely, those by which length is measured; or

/if.....

if the concept is mental, as of mathematical continuity, the operations are mental operations, namely those by which we determine whether a given aggregate of magnitudes is continuous"..".. the proper definition of a concept is not in terms of its properties but in terms of actual operations." As has been pointed out by Hesse (1952,1961), Eddington (1923) anticipated Bridgman by a few years in the publication of a clear summing-up of "operationalism", viz. "A physical quantity is defined by the series of operations and calculations of which it is the result;" and indeed the seeds of the operationalist viewpoint are to be found in the writings of earlier physicists and philosophers such as Mach (1902). But of all modern writers on the subject, it is undoubtedly Bridgman who has been the most influential; certainly it is not surprising that his name more than any other has come to be associated with operationalism. Whiteman (1967) remarks that "In view of Bridgman's repeated references to 'mental operations' ..... it seems rather strange that contemporary opinion should have widely imputed to him a positivistic insistence on purely physical operations." He describes as a "non-Bridgman Operationism" the doctrine that (in Popper's words) "theoretical concepts have to be defined in terms of measuring operations" (Popper 1957, 1963). But as the above quotation from his "The Logic of Modern Physics" suggests, there is good reason for believing that in his earlier writings Bridgman did uphold just such a doctrine. With the tenability or otherwise of this doctrine we deal below; our immediate concern is to establish whether or not definitions such as that of momentum given above can fairly be described as "operational". If we regard the definition of momentum as the product of mass and velocity as satisfactory, and we accept that this kind of definition is operational, then we would presumably

/incline.....

incline to the view that operational definitions of physical concepts are the kind to be aimed at.

There is little doubt that many today would accept the above definition of momentum as operational. Those authors who make a point of avowing their allegiance to the operational approach (e.g. Arons 1965 and Alonso and Finn 1967) use as specific examples of satisfactory definitions statements which have similar characteristics to those of the definition of momentum as mass times velocity. (E.g. Alonson and Finn: "...to say that <u>velocity is the distance moved divided by the time</u> is an operational definition of velocity.") Now the logical soundness of operational definitions has been questioned by a number of modern critics - we consider their criticisms below - and it is of interest at this point to note what these critics might regard as <u>acceptable</u> definitions of physical concepts. As a good example of an outspoken contemporary critic of operationalism we may take Mario Bunge, who in his book "Foundations of Physics" (1967) sets about a thorough re-examination and constructive criticism of the foundations of physics from the viewpoint of the philosopher of science. Let his definitions of linear momentum and velocity be taken as test cases, since supposedly operational definitions of these concepts have been quoted above. Turning to Bunge's exposition of the "Background and Primitives" of Particle Mechanics, we find that he defines the linear momentum (symbolically) as the product of the mass and the velocity of the particle, while his symbolical definition of instantaneous velocity is given in the form " $v \equiv \frac{df}{dt}\frac{dx}{dt}$ " . But these are exactly the definitions given earlier as apparent examples of operational definitions. Evidently the confusion that surrounds the interpretation of the term "operational definition" is not limited to the authors of school texts and teachers' guides.

/A.....

A helpful discussion of the problem for our present purposes is that of Margenau (1942,1950). He prefers to use the term "epistemic definition" (or "instrumental definition") in place of "operational definition", but his discussion makes it clear that his "epistemic definitions" are the same as "operational definitions" in Bridgman's sense. He focuses attention on the essence of these "Bridgman" definitions, i.e. their instrumental character - "an actual use of physical instruments is involved." (Cf. Bridgman's original statements, and also Hesse's (1952) first rule for operational definitions - "The definitions are to include only physical operations which are or could be actually performed.") Illustrating the point with reference to speed, Margenau points out that speed can be defined instrumentally or operationally as the reading on a speedometer, an instrument which measures speed directly. Now it is true that the definition of speed as "distance divided by time" can be made "instrumental" if the terms "distance" and "time" are taken to refer to actual measurement operations of distance and time; but it is also true that the idea of speed as "distance divided by time" seems complete in itself without one having to think of specific measurement operations. As Feynman puts it, "We believe that there is something to measure before we build the speedometer" (Feynman et al 1963). The "something" is distance travelled per unit time or, more precisely, $\frac{dx}{dt}$. Thus this form of definition of speed is not essentially operational or instrumental, in contrast to that involving a speedometer. The definition of speed as "distance divided by time" or $\frac{dx}{dt}$ falls into the class of what Margenau calls constitutive definitions, which link together several different "constructs" (Margenau's preferred word for concepts); the definition of momentum as "mass times velocity"

/is.....

is obviously of the same type. The essential characteristic of constitutive definitions is that they meet the above-quoted condition of Bunge (1967), viz. they "enable us to eliminate one concept in favour of others"; in another phrase of Bunge's, they establish "concept-concept correspondences."

Many of the concepts met with in physics are "derived" in the sense that conventionally they are derived from other concepts. Generally speaking, it is possible to formulate constitutive definitions for concepts in this category, although in certain cases closer analysis may reveal that apparently constitutive definitions are more correctly regarded as operational. Such cases will be considered when necessary in later sections of this thesis. For the moment, we note that many definitions commonly thought of operational are more accurately described as constitutive, and that constitutive definitions are acceptable to both physicist and philosopher.

3.2.4 Is operationalism tenable?

The reason why constitutive definitions have come to be regarded as operational by some is plain enough. It is the implied quantitative content of these definitions that has led to this confusion. Since these definitions usually (though not invariably) contain clues to the means whereby the concepts defined may be measured, they have the superficial appearance of being based on measurement operations. However, they are not so based; not, that is, unless one chooses to insist, following Bridgman's original line of thought, that concepts have no meaning whatsoever except in terms of measurement operations or of the numbers representing the results of measurement. This extreme operationalist doctrine still lingers on in some quarters, so it is fitting now to examine some of the criticisms that have been voiced against it.

(1) As has been shown by Ramsay (1931), Braithwaite (1953)

/and.....

and others, it is logically impossible to define all the concepts of any theory in terms of operations. One might be surprised to find this said of classical physics, but it is a familiar situation in modern microphysics, in which it has been found expedient to employ concepts which cannot be linked directly with measurement operations and which cannot therefore be given "operational definition". An obvious example of such a concept is the state function of quantum mechanics, which has no direct physical significance. Only its absolute square can be assigned a physical interpretation. It is interesting to note the attitude of the leading operationalist of our time to theoretical concepts of this kind (Bridgman 1961). He has summed it up by saying that "It is desirable that the paper-and-pencil operations of the theorist be capable of eventual contact, although perhaps indirectly, with instrumental operations." With regard to the Ψ-function itself, he comments that "...defined as a probability amplitude, [it] is at first a pure construction of the theoretical physicist, but again it makes its connection through mathematical operations, in this case operations of integration, with the mean density of electric charge, which does have instrumental significance." In other words, Bridgman does not appear to be insisting that "instrumental interpretation or validation must be given to every element in [a] theory" [the quotation is from Lindsay (1961), who was putting the point to Bridgman]; indeed his general attitude as expressed above would surely be found acceptable by all or most physicists.

(ii) According to the critics of operationalism, the view that we can define theoretical concepts in terms of measurement operations really puts the matter back to front; we need theories and the theoretical concepts embedded in them to confer meaning upon the observations or measurement

/operations.....

operations we carry out. In Popper's phrase (1963), "Measurements presuppose theories". Virtually any measurement operations take place against a background of some sort of theory, and the data we obtain from our measurements can be said to be "theory-laden" (Hanson 1958). Measurement of time with a clock presupposes a "theory" of time-keeping with clocks (cf Section 3.6); measurements of length with measuring rods is bound up with ideas about temperature and thermal expansion (cf Popper 1963); careful analysis of methods of measuring mass (cf Sections 3.7.2 and 4.7.10) shows that these are inextricably bound up with the theoretical structure of classical mechanics; and so on.

(iii) If the fundamental concepts of theories were confined in meaning by the operations by which they are measured, these theories would be incapable of growth, because they would not have the flexibility which could enable them to assimilate new observations without radical change of meaning (cf Hesse 1961). The need for flexibility in the meaning of fundamental concepts can be illustrated with reference to momentum and its relationship with mass. In the special theory of relativity, we may retain the definition of momentum as "mass times velocity", provided that we are prepared to broaden our interpretation of mass. We may, for example, have "defined" mass previously as force divided by acceleration, but mass defined in this way is not the quantity we must multiply by velocity to obtain "relativistic" momentum. (As we shall see in Sections 3.7.2 and 4.7.4, there are difficulties associated with the quoted "definition" of mass even in Newtonian mechanics, but these do not affect the point at issue.) Alternatively, we may prefer to treat mass in relativity as an invariant quantity (cf Section 3.7.2), in which case we shall have to re-define momentum in

/relativity.....

relativity as mass times proper velocity. The willingness of physicists to accept this kind of "open-endedness" in their attitude towards fundamental concepts makes possible the extension of old theories and the development of new ones.

(iv) The "incompleteness" of the meaning conveyed by an operational definition is also apparent from the fact that, as it stands, such a definition implies that the defined concept only comes into existence when the operations by which it is measured are carried out. To overcome this difficulty, operational definitions have to be understood to mean that the concept defined is "there all the time" and would be "obtained" if the specified operations were carried out (cf Bergmann, Hempel 1961). Otherwise we should have to conclude that the operational definition of speed with a speedometer means that if the speedometer of a car is broken, the car could have no speed - and this is manifestly not the way in which we think of speed (cf Feynman et al 1963).

In the face of these criticisms, it is difficult to uphold the tenability of extreme (non-Bridgman) operationalism. Lindsay's (1968) summing-up puts it well: "Thoroughgoing or rigid operationalism appears to be outside the mainstream of successful physical inquiry. The tempered operationalism of Bridgman's middle years, which incorporated mental or paper-and-pencil concepts along with those instrumentally defined, seems to be nothing more than the standard scheme of physical theorizing now current, with varying degrees of emphasis, among modern physicists."

We must now consider the status of operational definitions in the light of these criticisms. We have seen that it is not possible to define all concepts operationally, and that definitions of this kind encounter the objection that they place restrictions on the meaning of concepts

/which.....

which may turn out to be crippling. In view of all this, it may be asked whether "operational definitions" can properly be called definitions at all. Bunge (1966,1967) has strongly urged that operational definitions cannot qualify as definitions in the strict sense of the term since they are not "logical identities" or "concept-concept correspondences". It is plain that operational definitions establish correspondences between concepts and operations rather than between concepts and concepts. Therefore, unless we hold with Bridgman that a concept is "nothing more than a set of operations" - as we have seen, such a standpoint would be insecure, to say the least - we have to admit the correctness of Bunge's argument.

3.2.5 The role of operational definitions

In view of the above considerations, do we need "operational definitions" at all? Margenau (1942,1950, 1961) claims that we need both constitutive and operational definitions. According to him, "... it is necessary... that every accepted scientific measurable quantity have at least two definitions, one formal [constitutive] and one instrumental [operational]" (Margenau 1961). (Lindsay (1968) has expressed essentially the same view: "... every physical construct that represents a physical quantity subject to measurement must contain both epistemic [operational] and constitutive aspects in its definition"; the only difference is that whereas Margenau prefers to talk of separate constitutive and operational definitions of the same concept, Lindsay sees constitutive and operational elements as being fused in a single definition.) The reference to a measurable quantity in these quotations is significant; it implies that concepts not susceptible of direct measurement and therefore incapable of being given operational

/definition.....

definition are not to be ruled out of order on this account alone. We might summarize Margenau's argument in favour of the need for operational definitions as follows:-

Operational definitions are, by common consent, links between concepts and immediate experience (even those who object to the use of the word "definition" here would surely admit the correctness of this description) and have value for precisely this reason. Without such links we should have no physics. It is an eminently worthwhile aim to determine whether or not a given physical concept has "operational meaning" (i.e. yields up measurable values) and to establish the conditions under which it can have such meaning. In this way one can fix a borderline between useless speculation and profitable theorizing.

Looked at in this light, "operationalism" has undoubted value and operational definitions can play a useful role, although whether they can correctly be called "definitions" is, as we have seen, open to debate; "prescriptions for measurement" might be a more generally acceptable term.

While it is obvious that it must be possible to furnish "operational definitions" or "prescriptions for measurement" for any measurable physical quantity, it does not necessarily follow that it is possible to furnish <u>constitutive</u> definitions for all such quantities within a given conceptual scheme. Those concepts in physics which are customarily treated as "fundamental" instead of "derived", such as length, time, mass, temperature and electric charge, cannot be given constitutive definitions unless one chooses to regard them as derived from other concepts which are instead to be treated as fundamental. This does not contradict what we take to be the true import of Margenau's assertion that "all measurable quantities must be susceptible to formal definitions that link [tnem] non-operationally to other

/concepts.....

concepts of the theory". (Margenau 1961). Fundamentally, Margenau is merely insisting on the need for formal relations between concepts which form part of the structure of a valid scientific theory; and this condition can be met even if some of the concepts cannot be constitutively defined. Although it may not be possible to connect some of the "undefined" concepts of a theory by means of constitutive definitions, they may nonetheless be connected by axioms or laws.

The points that we have made concerning definitions of and relations between concepts will receive further elaboration when we come to consider the fundamentals of classical mechanics (Chapter 4), where we take up such matters as whether or not it is possible to define force and mass and whether the relation between them is characterized by definition, axiom or experimental law.

3.2.6 Definitions and the teacher

(i) Concepts. The implications of the foregoing discussion for the practising teacher must now be examined. Perhaps the foremost of these is that physical concepts are not necessarily capable of "precise" definition (in the sense that their full content of meaning can be compressed into a short statement), or even of definition at all in the strict sense of the word. It is essential for the teacher to appreciate this point, lest he seek to propagate the contrary view and so give his pupils the wrong idea of how physics works. He would do well to realize that it is far wiser not to attempt to define certain concepts, than to hand out pseudo-definitions of them to pupils. Next, the teacher should have some understanding of the difference between constitutive and operational definitions, and be aware that some doubt attaches to the claim of the latter to be regarded as true "definitions" at all. Acquaintance

/with.....

with these ideas on definitions should incline him towards casting a critical eye on the definitions he encounters in textbooks.

A useful clue to the adequacy or otherwise of a definition in terms of the foregoing discussion is whether or not the proposed definition has quantitative content. A constitutive definition, which relates a given concept quantitatively to others, obviously has this characteristic. It does not necessarily follow that non-quantitative definitions of concepts are automatically bad or useless (see below), but the lack of quantitative content in a statement purporting to be a definition ought to put the teacher on his guard.

We have noted that it is impossible to furnish constitutive definitions for the fundamental concepts of a theory. How, then, if at all, is one to define them? In a number of cases it is possible to formulate operational definitions of such concepts, but because of the doubtful status of these statements as definitions, it is probably wisest not to treat them as such, particularly in elementary teaching. In a sophisticated treatment of the subject, one can simply treat fundamental concepts as undefined or "indefinables". Thus in a neat, formal "axiomatisation" of the foundations of particle mechanics (e.g. that given by Bunge (1967)), such fundamental concepts as time, length, mass and force are listed as being among the "primitives" of the theory, with no attempt being made to give them formal definition. Such a procedure is too forbiddingly austere to be of much use to the elementary teacher, who might well feel that there is something to be said for trying to sum up at least part of the essence of a fundamental concept in a short statement - a "definition", in fact. Non-quantitative "definitions" of this sort can undoubtedly be devised for

/some.....

some fundamental concepts. For example, we may "define" temperature as "that property of a system which determines whether or not it will be in thermal equilibrium with other systems" (Sears and Zemansky 1964). Such "definitions", which "give a sort of flavour of what the concept is or tell what it is not" (Margenau 1950) are called denotative definitions by Margenau. They have proved useful in the past in forming part of the basis of a pupil's understanding of a concept, and there is no objection to their continued use in this way, provided that (a) they are chosen or formulated with extreme care, and (b) they are treated as "definitions" and not as "Definitions", i.e. that they are not given the "heavy-print" treatment and made "swot"-material for examinations. A denotative definition would be worse than useless if it failed to meet condition (a); we have already noted unfortunate examples of this in the case of mass (cf Section 3.2.1) and EMF (Section 2.2). There is a very good reason for insisting on condition (b). Since denotative definitions can only give "a sort of flavour of what the concept is", it would be unwise to give them undue prominence. If it is desired to set an examination question on, say, the meaning of mass, it would be better to ask the pupil to try to tell his "Uncle George" (cf. Boulind 1966) about mass than to ask him to reproduce a "definition" of it. If the best the pupil could do towards answering such a question would be to reproduce a memorized "definition" of mass, he would then qualify for only some of the available marks. In the following pages we shall give attention where necessary to the problem of acceptable denotative definitions.

Of the currently available textbooks to which a teacher might want to turn for guidance, the present author has come across at least one which uses an approach to definitions

/closely.....

closely similar to that advocated here. This is 'University Physics" (Third Edition) by Sears and Zemansky (1964). These authors lay down the rule that "A definition of a quantity in physics must provide a set of rules for calculating it in terms of other quantities that can be measured." Though it may seem that this statement smacks of operationalism, this does not appear to be the way in which its authors view it. Suitably interpreted, the Sears and Zemansky rule amounts to a description of a constitutive definition, although one must take care here as not all prescriptions for calculating one quantity in terms of other measurable quantities can safely be taken as definitions. Sears and Zemansky recognize the existence of "fundamental indefinables", such as length, time and mass in mechanics, a fact which further suggests that they have a cautious approach to the problem of definitions.

To sum up: One of the most important tasks of the teacher of elementary physics is to convey to his pupils some feeling for the nature of physical concepts. Both the definitions he gives of such concepts and the teaching with which he backs them up should be directed towards making the pupil realize that physical concepts are not just "there" in Nature, waiting to be recognized and used - instead they are man-made, constructed by a sophisticated process of abstraction from experience and selected for their usefulness, i.e. their ability to be incorporated within the framework of a physical theory. No pupil could gain a full appreciation of this point during his years at school, but he could perhaps be guided away from certain misconceptions he might have acquired or be in danger of acquiring. The use of constitutive definitions where appropriate will certainly help in this respect in showing how concepts can be con-

/structed.....

structed from other concepts. Where it is not appropriate to use such definitions, as in the case of the fundamental concepts of a theory, it is necessary to emphasize the central importance of the quantitative relationships between these concepts and others in the theory. Suitable denotative definitions of concepts can be used, but only as points of reference for the pupil's understanding of the concepts so "defined".

(ii) Units. In defining the units in which physical concepts can be measured, one is less open to confusion and criticism than in seeking to define the concepts themselves. One can now set strict limits to the meaning of what is being defined, something which, as we have seen, is not necessarily possible for physical concepts. One's definitions must now have quantitative content if they are to be of any use and be closely linked to measurement operations, and this at once narrows the range of acceptable types of definition: denotative definitions are ruled out altogether, and acceptable definitions must either be constitutive or operational (see below). We can in fact lay down a simple programme for the definition of units in any given field of physics, viz.

(a) Decide which concepts are to be used as bases for standards in the system of units to be established (they need not necessarily be identical with those chosen as fundamental concepts), and then set up definitions of these standards. For example, in the MKS system, the chosen standards are those of length, mass and time, and internationally agreed definitions establish the standard metre, kilogram and second.

(b) Formulate definitions of the units of all "derived" concepts in terms of the units of the concepts from which they are derived. For example, since momentum is defined as the product of mass and velocity, the unit of its

/magnitude.....

magnitude can be compounded suitably from the units of mass and velocity (and of course the unit of the magnitude of velocity must in turn be compounded from those of distance and time); hence one arrives at a definition of the unit of momentum of the following form: a body of mass 1 kilogram travelling with a velocity of magnitude 1 metre per second has a linear momentum of magnitude 1 kilogram-metre per second (1 kg.m.s^{-1}) (The significance of naming units in this way is discussed in Section 3.9.4.). Definitions of this sort can be regarded as constitutive since they are equivalent to defining certain units in terms of others.

It may be convenient to <u>define</u> the units of fundamental concepts in terms of those of other concepts. Thus in one presentation of classical mechanics, force is treated as a fundamental concept and mass is "defined" as the ratio of force to acceleration; on the other hand the <u>unit</u> of force is defined in terms of the previously chosen or defined units of mass and acceleration. Since the unit in which a concept is measured is not to be identified with the concept itself, this procedure does not imply any circularity of logic.

In connection with (a) above, concerning standards, certain points deserve closer attention. It is usual to say, for example, that the <u>unit</u> of mass in the SI system is defined to be the mass of a certain cylinder of platinum-iridium kept at the International Bureau of Weights and Measures at Sevres. Such a statement is open to criticism on two main grounds. First, "...an initial standard is not a unit... Nothing less than the complete specification of a scale will satisfactorily define a unit. If we merely refer to a particular object to be used as an initial standard for the fundamental measurement of some quantity, we have not defined a unit for the measurement of that quantity. In addition, we must describe the fundamental measuring procedure, and nominate the principle of correlation"

/(Ellis.....

(Ellis 1966). (By "principle of correlation" Ellis means the method of assigning numerals to the members of what he calls the "extended system of standards", which in the case of mass is the set of masses built up by finding known multiples and submultiples of the standard mass through comparison of masses with an equal-arm balance.) Second, a definition of a standard is unmistakably an operational definition - the whole point of having one is to establish a direct link between a measurement carried out in a specified situation or on a specified object and the standard it defines - and as such it is in doubt as a definition. Bunge (1967) has suggested the name "referition" for a definition of a standard, on the grounds that it specifies a relation between a "concept" (e.g. the concept of "1 kilogramme") and a physical object or situation (e.g. a cylinder of platinum-iridium) which is the referent of the concept.

With regard to the first of these criticisms, we may note that the pupil in the senior school is sufficiently familiar with the idea of measurement to have an intuitive grasp of the "fundamental measurement procedure" and "principle of correlation" used in setting up a scale of length. He should not have too much difficulty in understanding the application of these ideas to the setting up of scales of mass and other quantities; indeed it is an essential part of his training in physics that he should gain this understanding. Thus the criticism regarding the incompleteness of the definition of a unit in terms of a standard would be met to a sufficient extent by the instruction which the pupil would normally receive. Concerning the other criticism, it may perhaps be agreed that the distinction between a "referition" and a definition is a subtlety of terminology which need not trouble the elementary teacher. His account of his subject will not be complete unless he includes these

/"referitions".....

"referitions", and if he calls them anything at all, he may as well call them definitions.

Finally, it is worth returning for a moment to the doubts evidently felt about "operational definitions" by the author of the O.F.S. Teachers' Guide, since it is mainly in reference to definitions of units that he uses this term. We have seen that he describes his definition of a volt as "only" operational. As he says, the definition is not exactly phrased as a scientist would prefer it. However, it does bring together the right units (the volt, the coulomb and the joule); the clumsiness of phrasing really results from the gallant attempt of the author of the Guide to define the volt as a unit of EMF, a concept which it is not possible to define properly in the context of the syllabus concerned. In fact, an acceptable definition of the volt would express it either in terms of the joule and the coulomb or the watt and the ampere. A definition of this sort may be "only" operational - whether it is or not we shall consider in a moment - but it is difficult to see how one might improve on it. Consider the second of the above implied definitions of the volt. More fully expressed, it amounts to the statement that "In the SI system, the unit of potential difference, to be called the volt, is a watt per ampere;" i.e. the volt has been defined in terms of the watt and the ampere. We can therefore claim that the volt is constitutively defined. In applying this definition we would carry out instrumental operations; e.g. we could calibrate a voltmeter by measuring the rate of heat dissipation in a resistor carrying a known current and then making use of the fact that the measured rate of heat dissipation in watts divided by the measured current in amperes yields a potential difference measured in volts. But the unit of potential difference is no more operationally

/defined.....

defined than the concept itself; in the context of circuit theory, for example, one might choose to define potential difference as the rate of energy transfer per unit current, and this is certainly a constitutive definition.

3.3 Physical laws and the teacher

It is impossible to proceed very far in the quantitative teaching of physics without mentioning or discussing some physical "law" or other. The nature of the subject seems to require that a pupil's progress through it must be punctuated by statements of such laws: in the past, he was invariably required to commit these statements to memory. In old-fashioned physics courses, a pupil might have been required to "verify" these laws in the laboratory; nowadays, he is more likely to be expected to "discover" them for himself, aided by a certain amount of stage-setting from the teacher. Because laws bulk so large even in an elementary physics course, it is important for pupils to gain some understanding of their nature.

It would be natural for a teacher to wonder whether there exists any reliable general description of what constitutes a physical law. The most helpful simple description known to the author is that given by Rogers (1960): "It is..... a relationship, or description of behaviour which has been discovered, and which seems very general and appeals to us as simple and important." This can be regarded as an unexceptionable statement of what a physical law is, embracing all known examples, provided that one is careful to amplify what is meant by a "description of behaviour that has been discovered." In the case of many laws, this statement can be taken almost literally. Hooke's Law, for example, can be said to describe the way in which certain materials "behave" when stresses are applied

/to.....

to them, if by "behaviour" in this instance one means the manner in which certain measurable properties of the materials depend on the applied stresses. This law was "discovered" in the sense that it took a particular experimenter (Robert Hooke) to perceive this relationship between the load applied to a spring and the resulting extension. Whether or not Hooke guessed the correct form of this relationship before putting it to experimental test, we do not know: that he did "verify" it to his own satisfaction is certain (cf Magie 1935). Laws like Ohm's concerning the properties of certain electrical conductors, Boyle's concerning gases, Kepler's concerning planetary motion and even Coulomb's concerning the force between static electric charges may be said to be in the same category as Hooke's Law, in that their "discovery" seems to have been closely associated with the analysis of experimental or observational data and they constitute an almost direct "description of behaviour." Such laws (with the exception of Coulomb's) are often called "empirical laws" (cf Constant(1963): "Such a relationship, based purely on experimental evidence, is called an <u>empirical</u> relationship, or an <u>empirical</u> law."), to distinguish them from "fundamental laws". <u>Empirical</u> laws apply only to certain materials or systems and have a restricted range of validity, whereas <u>fundamental</u> laws have much more general applicability and their range of validity is, or may be postulated to be, much wider than that of the typical empirical law. It is for this reason that Coulomb's Law is placed in the category of fundamental rather than empirical laws, even though its "discovery" was associated with experimental observation. It will be noticed that the examples of empirical laws mentioned above, besides being "descriptions of behaviour which have been discovered", also meet Rogers's

/other.....

other criteria of generality (in the sense of being "fairly" rather than "very" general), simplicity and importance.

One of the tasks of the theoretical physicist is to "explain" empirical laws, i.e. to derive them from the fundamentals of a particular theory (or of several theories). Sometimes a "law" may be derived from the fundamentals of a theory prior to its investigation in the laboratory. Constant (1963) places such laws into the special category of "derived laws". Actually, it is not easy to find straightforward examples of "derived laws"; a great many physical laws seem to have been "empirically" discovered first and assimilated into satisfactory physical theories afterwards. One clear-cut example of a derived law appears to be the Maxwell-Boltzmann Law of distribution of speeds of gas molecules, which was originally derived from the assumptions of kinetic theory and not investigated experimentally until many years after its discoverers had died. Clearly, "derived laws" of this kind are similar to empirical laws in the manner in which they conform to Rogers's description of physical laws.

The case of "fundamental laws" is of particular interest. Newton's Laws of Motion "describe" the motion of bodies, and were discovered by Newton and others by dint of some hard thought about and careful observation of motion, but attempts to "verify" them, as opposed to merely illustrating them, run into the difficulty that one does not seem to be able to avoid assuming somewhere or other some aspect of what one is trying to prove. One may search the literature in vain for an account of some definitive verification of Newton's Laws of Motion comparable to the experiment of Plimpton and Lawton (1936) on Coulomb's Law. Thus the "description of behaviour" furnished by these laws is not quite of the same kind as that exemplified by empirical laws. It may still justly be called a "description of

/behaviour".....

behaviour", in that the observed benaviour of macroscopic systems is consistent with Newton's Laws of Motion, but the description is no longer an unambiguous experimentally established relationship between two or more physical concepts. Newton's Law of Gravitation can again be held to be a "description of behaviour" - it describes the motion of planets and satellites over a wide range of validity - and Newton "discovered" it, in the sense that he was the first to show that the observed motions of heavenly bodies were consistent with it. But it has never been subjected to what could properly be called experimental test - as Lindsay (1968) has pointed out, "...the Cavendish experiment ..... is a method of determining G on the assumption that Newton's Law holds" and not a "verification" of the Law - and it is unlikely that it ever will be.

The laws of classical electromagnetism in their most generally applicable form require perhaps even more cautious interpretation of the terms "description of behaviour" and "discovery" than do the previous examples. Coulomb's Law in the form referred to above is not true in the general case: it has to be replaced by Gauss's Law, which relates the surface integral of the electric field vector over a closed surface to the total charge enclosed within the surface. This has not been subjected to a direct experimental test of general validity. Maxwell (1881) must apparently be given the credit for implicitly proposing that this relation, which is equivalent to Coulomb's Law for non-accelerating charges, should be applied generally. (No explicit statement will be found in Maxwell's "Treatise on Electricity and Magnetism" that this equation is now being given much wider application than its derivation in a purely electrostatic context would suggest. But Maxwell lists it among the "General Equations of the Electromagnetic Field" in article 619 and implicitly uses it in arriving at his

/expression.....

expression for the "displacement current" in articles 610 and 611). This is certainly "discovery" of a kind, but it is obviously different from the simple inference from experiment which appears to have characterized the discovery of some laws. It is tempting to regard Faraday's Law of Electromagnetic Induction as "inferred from experiment" and therefore a direct "description of behaviour", but one should remember that the accuracy of Faraday's experiments in this connection was of the order of perhaps 1 part in 20 (cf Tricker 1966), and there have been no recent attempts to "verify" it to a higher degree of accuracy. Not only has Maxwell's "Law of Magnetoelectric Induction" (this is Constant's (1963) term for it), which relates the "circulation" of the magnetic field vector to the rate of change of electric flux, never been the subject of direct experimental check, but it does not seem possible to devise any simple experimental arrangement to carry out such a check (cf Section 7.21.2). It is impossible now to be certain of how Maxwell arrived at the discovery of this Law, but the following remark is typical of his published arguments concerning the need for it: "... the extreme difficulty of reconciling the laws of electromagnetism with the existence of electric currents which are not closed is one reason among many why we must admit the existence of transient currents due to the variation of displacement [i.e. the $\vec{D}$ vector]. Their importance will be seen when we come to the electromagnetic theory of light" (Maxwell 1881, article 607). (There is no indication in Maxwell's published work that he was influenced by considerations concerning the symmetry of the fundamental equations in arriving at this law (cf Bork 1963).) Thus the Law was postulated in order to patch up an inconsistency in the theory, and not in any way inferred from experiment. If we

/may.....

may stretch the word "discovery" to include "invention" or "postulation", then Maxwell certainly "discovered" the Law of Magnetoelectric Induction. The remaining Maxwell equation, which states that the net magnetic flux through any closed surface is always zero, can be held to be an inference from experiment and thus a fairly direct description of behaviour, in that free magnetic poles, whose existence would violate this "law", have never been detected (c.f. Section 7.15.2); the demonstration that an appropriately-defined field vector $\vec{B}$ must have zero divergence is a straight-forward exercise when it is assumed that "the algebraic sum of the magnetism of [any magnetic] particle is zero "(Maxwell 1881, article 402). But at the time that Maxwell asserted that $\vec{B}$ was necessarily divergence-free ("the solenoidal condition which is always satisfied by the components of the magnetic induction" as he remarks in article 403), the attempts to detect free magnetic poles were not being conducted on their present scale, if indeed they were made at all, so that this Law has, at least to some extent, the character of a postulate.

Enough has now been said to show that although fundamental laws are clearly different in character from empirical or derived laws, they can still be accommodated within Rogers's simple description of physical laws in general. This means that the teacher's burden in telling his pupils about laws is not too heavy. In a typical present-day course, the first example of a physical law that a pupil will encounter is likely to be a law of the empirical type, such as Hooke's Law, which can be "inferred from experiment". (Cf the Nuffield Physics Course, where provision is made for pupils to investigate the behaviour of cheap springs for themselves.) Subject to certain conditions which we discuss below, the teacher can comment on the emergence of

/this.....

this relation from the experimental data in much the same terms as those used by Rogers, and so introduce the pupil to the term "law". When at a later stage the pupil encounters fundamental laws, (e.g. Newton's Laws of Motion and Gravitation and the Laws of Conservation of Momentum and Energy), the teacher can draw attention to the fact that while those are still "laws", they are in a different category from Hooke's and Ohm's Laws. In some expositions of physics, all such "fundamental laws" are called "principles" to distinguish them from the other kinds of law (cf Shortley and Williams 1965, in which Newton's Laws of Motion and Gravitation become "Principles" and Archimedes's Principle is degraded to a mere "Relation".). We do not think that one need go quite as far as this in order to emphasize that there are different kinds of physical law, although there can be no objection to teachers or textbook authors proceeding in this way if they think it likely to be advantageous. What is important is that the pupil should appreciate that the laws now under discussion are of much more general application than those he first met, and that the step of assuming that they have a wide range of validity is taken on the basis of faith in the outcome rather than on unshakable direct evidence. It is important that, wherever possible, pupils and/or the teacher should carry out experiments which <u>illustrate</u> these laws, but the teacher should stress that these experiments are merely suggestive or illustrative of the underlying laws and must not be taken to be "proofs" of their validity. Given this kind of preparation, pupils might be able to accept the introduction in its proper place of Maxwell's "Law of Magnetoelectric Induction", for which it will not be possible to offer them any experimental illustration at all. No doubt we need more research into the problem of how pupils respond to the introduction

/of.....

of such postulates in the context of the physical sciences, but it is at least reasonable to suggest that with the approach suggested here the pupil need not be too alarmed by the appearance of a law which is obviously postulational in character.

Finally, there is a further point concerning even the simplest laws which the teacher should help his pupils to appreciate. It is commonly said of physical laws that they are "inferred from experience". This is true inasmuch as these laws derive from our attempts to describe experience, and in the case of such laws as Hooke's or Boyle's or the laws of reflection and refraction of light, the inference from experience may seem very direct indeed. But the process of inference is more complex than it may seem at first. Before Hooke's Law can be inferred from a set of measurements, the appropriate measurements must be carried out, and this requires a decision on the part of the experimenter concerning what to measure. That is, the experimenter must decide which concepts he is to study the relation between, and this means that the right concepts must already be available. (Cf Bunge 1967: "..... before we can join concepts to build [law] statements, the concepts must be there".) Hooke's Law as applied to the stretching of a spring concerns the concepts of "extension" and "load"; Ohm's Law deals with the relatively sophisticated concepts of current and potential difference, and the laws of reflection and refraction of light are formulated in terms of the concept of the light ray, which the pupil is likely to meet first in the context of a particular theory of light, i.e. that which assumes that it travels in straight lines. These examples show that even the establishment of familiar empirical laws takes place against a background of some assumed conceptual structure. The teacher should do his best to

/ensure.....

ensure that this aspect of the emergence of a physical law from data obtained in the laboratory is not lost on the pupil. When the pupil is busy "discovering" Hooke's (or any other) Law in the laboratory, his attention should be directed to the decisions he has made in setting up the experiment and in interpreting the data he obtains. He should be questioned about the confidence with which he assigns a particular mathematical relationship to a pair or group of physical concepts and the range over which he expects this relationship to hold. All this should help him to regard even "empirical laws" as more than mere "a posteriori summaries of data" (the phrase comes from Bunge (1967), who makes plain his opposition to the view of physical laws it represents).

3.4 Measurement and Units

The sooner a pupil starts measuring things, the better, as far as his secondary-school physics is concerned; and by the time he has reached Standard Six he is surely old enough to gain some grasp of the meaning of units of measurement, and the standards on which they are based. It may be that his early measuring in Standard Six may be placed in the context of finding out about different forms of matter, as in the Nuffield Physics Course, but this does not alter the fact that he will have to face the question of units and standards in order to have a proper idea of what he is doing. Here we will treat the matter as a subject in itself; our discussion could easily be placed in a context of the kind mentioned above if the need arose.

What system or systems of units should be used? With the general adoption of Système Internationale (SI) units by the scientific community of the world, there is less

/point.....

point than ever in drawing careful distinctions between MKS and CGS systems in elementary teaching; and with South Africa itself committed to metrication in the near future, with SI as its approved system of units, one need have no qualms of conscience in neglecting the British Engineering system in one's teaching in favour of SI.

3.5 The Concept of Length and its units and standards

The heading is not intended to imply that there is any need for a teacher to dwell on the concept of length as such. The furthest that need be gone in this direction is to stress how closely the concept is bound up in our minds with the way in which we measure it. Even when we think of something as being long, our thoughts seem to lead naturally to wondering how long it is - and the moment we start making estimates of that kind, we are mentally carrying out crude measurements of length. It would do pupils no harm to have this brought home to them - and it would help them to feel at ease when they meet this kind of thing again, as they will repeatedly be doing in physics.

Introduction of the metre and its multiples and sub-multiples is not the kind of problem that affords the teacher any difficulty, though his presentation of this material could be considerably livened up by the use of examples drawn from everyday life and from the macroscopic and microscopic worlds to illustrate different orders of magnitude of length (cf PSSC 1960, 1965: Chapter 3, "Space and its Measures") . As for the standard metre, what might otherwise be rather a dull matter could lead to a useful, open-ended discussion if approached in the right way. If, that is to say, the subject could be discussed under the headings of (i) the need for a standard (ii) the nature of a standard, with the pupils being invited to express their views on both points, the results might well be beneficial. This is hardly the moment

/to.....

to introduce the idea of a wavelength standard, in terms of which the standard metre is now defined; perhaps it is still true that "early mention of [the new definition] to any but the most sophisticated pupils can only lead to mystification and confusion" (Helm 1963). It is recommended that the modern definition of the standard metre should be kept in mind by the teacher and stored up for future use; it can be made known to pupils after they have learned something about light waves. (Were a school ever to be fortunate enough to be equipped with both a laser and a Michelson interferometer, it would be very easy for the teacher to illustrate the relation between length measurements and interferometry. Straight-line fringes obtained with the interferometer using laser light are easily projected, so that the passage of the fringes past a reference point as one of the mirrors is moved can readily be seen.) In the meantime, reference can be made to the old standard metre, whose specification (the distance between two scratches on a platinum - iridium bar kept at the International Bureau of Weights and Measures at Sevres, near Paris) can readily be grasped. For those who want it, an excellent account of the history of the standard metre is available (Barrell 1962), while some useful material along these lines will also be found in Feather's book on "Mass, Length and Time" (Feather 1959). Feather makes the point that "the precise definition of standards of length did not in fact precede the accumulation of knowledge", which is a useful reminder not to <u>over</u>-emphasize the need for standards.

3.6 <u>The concept of time and its measurement</u>

"Time" is a topic which South African science syllabuses have kept well clear of in the past; it is only in the new syllabuses that it has been allowed explicit treat-

/ment.....

ment. It may be that the reasons were mainly practical: that time measurements in a school laboratory should preferably be carried out with a stopwatch, and stopwatches are expensive. But since "time" was implicitly included in the pure physics syllabuses which did treat kinematics and dynamics to some extent, this argument does not seem altogether valid. Perhaps it was the nature of the topic itself which was the chief stumbling-block. Time is not the kind of concept one can attempt to treat by laying down a definition and leaving it there. As in so many other things, the PSSC led the way here by giving time and its measurement a prominent place early in their physics course. They showed that, far from presenting any formidable problems, the study of time could be absorbing and exciting if approached in the right way.

What <u>is</u> the "right way"? It is suggested that this might take the form of a two-pronged attack, the one prong being the performance of time measurements by pupils and the other a consideration of the fundamental principles of time measurement. As with length, this stress on measurement is important. It can help to dispel the air of mystery (and mysticism) which for some still surrounds the concept of time, and it can pave the way for an understanding of the approach to time to be found in special relativity. With the latter subject percolating lower and lower down in university physics syllabuses so that many universities now teach it in the first year, it is to be hoped that at the least, what is taught at school will not hinder the prospective student in his appreciation of the ideas of relativity.

The basis of the measurement of time is repetitive motion. Everyday experience comes to our aid by providing

/us.....

us with a host of examples with which to illustrate this point. We need look no further than our own bodies for a mechanism which seems to beat "time" more or less "regularly", namely, the heart. Or, in looking around us and confining our attention for the moment to natural phenomena, we observe the rising and setting of the sun, the passage of the seasons, the motions of other heavenly bodies: here are repetitive motions on a grand scale. Turning our attention once again to human beings, we find that there are many persons, usually those who have obvious musical talents, who can maintain a regular rhythmic pattern by waving a stick in the air, banging a stick on a drum, striking the keys of the piano or by various other means. Finally there are the many mechanical contrivances which incorporate repetitive motion and which may, indeed, be intended for use as clocks. All this should be brought home to pupils, who could in due course be encouraged to investigate simple examples of repetitive motion for themselves, e.g. simple pendulums or objects hanging from vertical springs and executing oscillations. We have referred to the alleged element of "circularity" in such experiments in our earlier section (2.4) on black boxes, the circularity lying in the use of (say) stopwatches to measure the period of repetitive motion. We (and the Nuffield Guide) have implied earlier that this logical difficulty is not a matter of great consequence; but it can in fact be bypassed, and in the process one could possibly bring one's pupils very close to an idea of what time measurement is all about. One might, for example, ask "What do we mean when we say that our heartbeats [or an unrhythmical drummer's strokes] aren't quite regular?" If the answer is given that one can check this against a watch, doubt can equally well be cast upon this procedure. But instead of prematurely extending the discussion at this

/point.....

point to a consideration of standard clocks, one can suggest an experiment with a simple pendulum and an object hanging from a spring. One's heartbeats can be compared in "regularity" with the small-amplitude oscillations of the pendulum and with the oscillations of the hanging object, and the oscillations of pendulum and hanging object can be compared with each other. Here is a foundation for a wider discussion of the checking of well-made clocks against one another, against the apparent motion of the sun and stars across the heavens and against atomic and molecular "vibrations". It is not suggested that this kind of discussion need be entered into at Standard 6 level; the pupil first needs to capitalize a little more on his intuitive grasp of time before he can derive benefit from the approach suggested above. But if he does not finally face this issue, his appreciation of this most important concept will lack a vital dimension.

Much fun can be had with time measurement, which will also advance the pupil towards the state of sophistication in which he can understand the discussion suggested above. We need not enter into detail here about the numerous experiments that can be done on direct and indirect methods of measuring time; suffice it to say that short time intervals are, for obvious reasons, particularly exciting to measure and that both simple and elaborate apparatus are available for this purpose. Under the former we may mention the hand stroboscope made famous by the P.S.S.C. (1960, 1965); under the latter, electric stop clocks, stroboscope lamps, cathode-ray oscilloscopes and quartz-oscillators-plus-scalers. (In Section 2.4 we have considered the logical problems attendant upon the use of these "black boxes".) South African teachers should not neglect to mention the caesium clock now installed at the Republic Observatory, Johannesburg, or the "pips" broadcast from it at one-second intervals at

/the.....

the frequency of 5.0 MHz.

Should a teacher try to define time for his pupils? Definitions are available that would be intelligible to physicists, e.g. "time is the independent variable in the equations of mechanics" (Margenau 1950) - a constitutive definition - or "time is that which is measured with a clock" - an operational"definition". But the first of these definitions is not really intelligible to the budding physicist, while the second, even if we momentarily leave aside the controversial aspect of operational definitions, merely shifts the burden on to the meaning of the word "clock" - which is where we (or the teacher) come in. The trouble is that there is far more to the concept of time than can possibly be conveyed by any glib definition. It is clear that the provision of a definition would in this case not solve any problems. For this reason, the answer to the question posed at the beginning of this paragraph is in the negative.

Finally, and again with especial reference to the future education of any budding physicists in the class, what about "absolute" and (for want of a better name) "relative" time? The prevailing practice in these matters is to let sleeping dogs lie - to make the implicit assumption that time is absolute and leave it at that. (One presumes that no teacher would inflict Newton's resounding phrases about time, or some modern version of them, on his pupils.) In defence of this there are two powerful arguments: (i) in the pupil's experience, time is to all intents and purposes absolute; (ii) the suggestion that time might be other than absolute might prove confusing and disturbing to the average class. Reluctantly, one feels that one has to admit the probable validity of point (ii). There are serious practical difficulties to contend with here. Given the present structure

/of.....

of physics or physical science syllabuses, there is little time in which to discuss the "relativity" of time in a way which will be both worthwhile and intelligible to the average senior school class. And even if the time were available, there would be very few teachers who could take advantage of it. We are therefore inclined to be timid and suggest that the most that the conscientious teacher could reasonably do in this connection would be to suggest to his brighter pupils that they <u>are</u> assuming the absoluteness of time and that this assumption might not be valid in all circumstances. (With the questions of whether relativity should be introduced into the school syllabus and, if so, how it should be taught, we deal in Chapter 11.)

To sum up: for the presentation of the concept of time we suggest (a) an introduction based on simple experiments on time measurement with clocks and/or stopwatches, to be followed by (b) more advanced experiments in which clocks of a less conventional kind could be used to measure short time intervals; and finally (c) a closer look at the meaning of time measurement, in which the concept of a clock is critically examined.

3.7 Mass

3.7.1 <u>Introduction</u> One of the simplest quantitative experiments a pupil can do early in his career is to "weigh" something on an equal-arm balance or some near-equivalent device. In this way, he meets the concept of mass. Of course he has been familiar with it in a way or with its fairly close relative, weight, for quite some time beforehand. He knows that many commodities he buys are priced according to their weight and he has seen people "weighing" things (in some such cases, it may actually have been the mass that

/was.....

was being determined). Now he meets it in the laboratory and the classroom, and for his future well-being in studying mechanics it is necessary that this concept should be handled with some care. We have already derided the practice of "defining" mass as "quantity of matter" and leaving it at that. Our task here is to suggest a way of improving on this. We shall tackle it in two parts; in the first, (see below) we shall justify in detail our rejection of "quantity of matter" and discuss methods of treating mass for Standard Six or Standard Seven pupils; in the second (see Sections 4.7.10-12) we shall approach mass more rigorously from within the context of dynamics.

3.7.2 Mass and "quantity of matter"

It is all very well for the present writer and many others besides to deride "quantity of matter" as a definition of mass; the fact remains that this definition or description still turns up, even in textbooks dedicated to a "modern" approach and intended for university use . (Cf Cotts and Detenbeck 1966; Harnwell and Legge 1967; Freier 1965). Even in a book almost entirely devoted to the exposure of errors in the traditional presentation of school physics, we find it stated that "Mass can be defined as the quantity of matter in a body" (Warren 1965); in the immediately preceding sentence the author tells us that the definition of mass is difficult! We shall deal with Dr. Warren's defence of this definition in due course; we may remark here that the fact that this is the only definition of mass to appear in such a book, is symptomatic of the power this tradition still exerts (cf Helm 1966). No doubt there are still science teachers in South Africa who know of no other definition of mass than this "wholly unnecessary absurdity"

/(Lindsay.....

(Lindsay and Margenau 1936). One remembers the genuine sense of surprise and wonderment with which a correspondent to the South African science teachers' journal "Spectrum" greeted the appearance of a "new concept of mass" [he meant "inertial mass"] in the PSSC physics course (Owen 1963). Therefore it seems sensible to devote some space to the problem here, in the hope that some clarification will result.

Although the concept of quantity of matter did not originate with Newton (see Jammer 1960), it is no doubt his use of the term which was primarily responsible for its continuing popularity. As is well known, Newton defines "motion" (momentum) as the product of "quantity of matter" and velocity, and in this and other passages in his "Principia" the term "quantity of matter" appears to occupy the same place in Newton's vocabulary as "mass" does in ours. Definition 1 of tne Principia reads, in Motte's translation, "The quantity of matter is the measure of the same, arising from its density and bulk conjointly", i.e. it is the product of density and volume. The interpretation of this definition has been disputed ever since its publication (cf. the summary given in Jammer, op. cit). On the face of it, it is circular; how can one define mass ("quantity of matter") in terms of volume and density, when density is defined in terms of mass and volume? But it seems that Newton, while his concept of density is similar to ours, did not define it in the way now accepted as usual; and as various authors have pointed out, among them Margenau (1950), Warren (1965) and Brown (1965), it is possible to "define" density operationally with the aid of experimental procedures involving Archimedes's Principle and without any reference to mass or volume. In this way we can "save" Newton's definition from

/circularity.....

circularity, but this does nothing to justify us in using "quantity of matter" as a definition of mass instead of a synonym for it. Did Newton himself attach a deeper meaning to this phrase than is apparent from his definition of it? Jammer (op.cit.) maintains that he did - that Newton's Principia shows that for him, "quantity of matter" was a fundamental characteristic of a body, which determined its gravitational attraction for other bodies and to which its inertial mass ("vis inertia" in Newton's original) was proportional. But neither Newton nor his "interpreters" are able to describe this "fundamental characteristic" for us in terms that would nowadays be regarded as convincing. In this connection, the author of the Nuffield Guide (IV) suggests that Newton's use of "quantity of matter" "was probably only a teaching device to make something clearer to people who were struggling with a new and unfamiliar idea. We might try that phrase with our pupils too." Although plausible, the first of these statements does not seem to accord with what Newton actually wrote about the concept. (We shall deal with the second statement in due course). We hasten to point out that we are not making a fatuous attempt to discredit any part of Newton's towering achievement. It is simply suggested here that what may have been a natural and perhaps intuitively obvious way of regarding mass in Newton's day need not seem so at the present time. Newton's Laws have now been with us for nearly three hundred years, and the experience accumulated over that period nowadays enables scientists to have a more precise understanding of certain aspects of mechanics than was possible for anyone living in the seventeenth century. In our view it would be unwise not to take advantage of this gain in understanding if it can benefit our teaching.

/Is.....

Is it a gain in understanding and could it benefit our teaching? In trying to answer these questions, let us look at the argument used by some to justify the prominence they give to "quantity of matter" in teaching mass. In a reasonably sophisticated form, it goes something like this (cf PSSC 1960, 1965): "Can we find some sort of measure for matter as we have for time and space? Volume suggests itself at first - the overall volume of a number of pieces of metal is not changed if we cut some of the pieces in half, and the more metal we have, the more volume it takes up. But this runs into obvious difficulties with gases, to say nothing of the everyday phenomenon of expansion and contraction of matter. Fortunately we find that there is a property of matter we can measure with a balance, which we call mass, which does seem to do the job for us. Take a piece of metal, find its mass; knock it about to change its shape, find its mass again - same answer. Melt it down, find its mass again - same answer. Take another piece of metal whose mass is the same as that of the first, find the mass of the two pieces together - answer: twice the mass of either. So we have twice the mass for twice the amount of stuff, three times the mass for three times the amount of stuff; so mass is proportional to quantity of matter and can be taken as a measure of it." In defence of this line of argument, we can say that as far as it goes it seems to be pretty sound reasoning, although one might object that the intuitive notion of "amount of stuff" or "quantity of matter" is not treated very searchingly. But suppose we take it further, and now say: "Here is a piece of iron and here is a piece of lead. I have measured their masses on a chemical balance, and I find they are the same. According to the idea of mass we have just arrived at, I must conclude that the two lumps of metal have the same quantity of matter in them. But what do I mean by this statement?

/I.....

I think I can see what is meant when it is said that two identical pieces of lead have twice the quantity of matter of one of the pieces. But it seems that the only meaning I can give to tne statement that tne lumps of lead and iron have the same quantity of matter in them is that they have the same mass, where mass is that interesting property of matter we measure with a chemical balance. We have clearly reached tne stage where we are defining quantity of matter in terms of mass, rather than the other way round."
Cf. tne following independent statements on this point, published in the same year: (i) "Those authorities who still feel compelled to talk about 'quantity of matter' in discussing mass use the phrase essentially as a synonym of mass, that is, they occasionally use three words instead of one to describe the same concept without, by this means, making it any clearer" (Helm 1963). (ii) ".....If we look closely at the expression 'amount of material', we see that it has no meaning independent of the concept of mass. Since we now nave a perfectly clear procedure for measuring mass [i.e. the dynamical method], we can discard tne concept of 'amount of material' as an unnecessary and somewhat inexact synonym for mass." (Taylor 1963).
(Mach (1902) appears to be making essentially the same point when he refers to the difficulties one encounters when trying to apply the concept of "quantity of matter" to chemically heterogeneous bodies.)

This might be tnought to dispose of tne problem, were it not that there are some who attempt to revive the concept of "quantity of matter" with reference to our present-day knowledge of microscopic matter. Thus, according to Bunge, ".....atomic physics teaches that the (proper) mass of a system is proportional to tne number of heavy 'particles'

/that.....

that make it up, for which reason mass is regarded as a measure of the quantity of matter in the system" (Bunge 1967). But "atomic physics" is somewhat misrepresented here; in the words of Constant, "..... the mass of a body is very nearly proportional to the number of heavy elementary particles (neutrons and protons) contained in the body" (Constant 1963) In our view, it would be a mistake to build too much on this near-proportionality, particularly in elementary teaching. Maybe the departures from exact proportionality are small, but they are of profound importance; much too important, in fact, to be obscured by undue emphasis on "quantity of matter." Significantly, none of the "quantity of matter" enthusiasts have, to our knowledge, attempted to explain what they mean when they say (as they presumably must) that the "quantity of matter" in the neutron is greater than that in the proton.

There is one more point to be considered before we deal with the question of the possible "benefit to our teaching" resulting from enhanced understanding of the concept of mass. This is the question of the light shed on the concept of mass by special relativity theory. We take up the relationship of mass and energy later; at this point we wish to consider what, if anything, the relativistic view of mass can tell us that might be of relevance to our present discussion. It has sometimes been said (e.g. Jammer 1960, Helm 1963) that the well-known relativistic mass-velocity relation is in itself enough to make nonsense of the concept of "quantity of matter"; how can "quantity of matter" be supposed to increase with velocity? But perhaps one is not being quite fair in taking this line. We arrive at the formula $m = \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}$ for the "relativistic mass" by taking the

/relativistic....

relativistic expression for momentum, $p = \frac{m_o v}{\sqrt{1-\frac{v^2}{c^2}}}$, and dividing it by the velocity, in accordance with the classical definition of momentum. But this is not the only way of approaching the problem; we could equally well choose to define relativistic momentum as the product of the proper (rest) mass m_o and the "proper velocity" $\frac{v}{\sqrt{1-\frac{v^2}{c^2}}}$, in which case we obtain a mass that does not increase with velocity. Some time ago, O'Leary (1947a) recommended dropping the concept of "relativistic mass" and following the alternative approach suggested above. In certain recently-published treatments of special relativity (Smith 1965, Taylor and Wheeler 1966, Helliwell 1966, Brehme 1968) other authors have taken up this cry and contended that it is preferable to confine the meaning of the term "mass" to the rest mass of a system and to avoid the use of the term "relativistic mass" ($\equiv \frac{m_o}{\sqrt{1-\frac{v^2}{c^2}}}$) altogether. According to them, the concept of "relativistic mass" has several disadvantages:

(i) by writing momentum in relativity as "mv" one conceals the factor $\sqrt{1-\frac{v^2}{c^2}}$ and may forget it is there (Helliwell);

(ii) one may mistakenly think that to calculate kinetic energy in relativistic mechanics, one need simply substitute the relativistic mass for "m" in the classical formula $\frac{1}{2} mv^2$ (Helliwell);

(iii) the idea of relativistic mass makes increase of energy of a particle with velocity or momentum appear to be connected with some change in the internal structure of a particle, whereas the increase of energy with velocity has its roots in the geometrical properties of space-time (Taylor and Wheeler);

(iv) since energy is <u>always</u> equal to m (relativistic mass)

/times.....

times c^2, it is not good economy with words or concepts to call m the mass, m_o the rest mass and mc^2 the energy (Smith);

(v) it seems desirable to have the word "mass" refer to an intrinsic property of the particle and not to a property of its motion (Smith);

(vi) calling $\frac{m_o}{\sqrt{1-\frac{v^2}{c^2}}}$ "mass" applies this name to the <u>time</u> component of the energy-momentum 4-vector, when we have already (in the guise of "rest mass") given this name to a very different concept, the <u>magnitude</u> of this 4-vector (Taylor and Wheeler).

In addition, Lyngesen (1967) has recently pointed out that the use of the word "mass" for $\frac{m_o}{\sqrt{1-\frac{v^2}{c^2}}}$ is not in accordance with current usage among elementary particle physicists, who are in the habit of using "mass" for "rest mass".

As against these disadvantages, Helliwell and Resnick (1968) have pointed out that there are certain pedagogic advantages in using the concept of relativisitc mass: the expressions for momentum and total energy take on simple forms in terms of relativistic mass, and one can give a plausible explanation for the existence of an upper limit to the speed of a body - its inertia increases as it goes faster and so it becomes more and more difficult to increase its velocity still further. Both Helliwell and Resnick conclude that it is really a matter of taste whether one uses "relativistic mass" or sticks exclusively to "rest mass" as the only "mass" - either approach, if consistently and correctly used, must give the right answers (cf. also Katz 1964). In this connection it is worthwhile asking what is actually measured in a determination of the "mass" of a rapidly-moving particle. The answer is that "mass"

/is.....

is never measured directly; that the momentum and energy of the particle are measured and a value for "mass" inferred from use of the appropriate formulae, or the momentum of the particle is compared with that of another particle of known mass and having the same velocity (Segrè 1964). Thus the value that may be assigned to "mass" in such a determination depends on how one chooses to interpret a formula; the choice is dictated by the experimenter, not the experiment.

In the face of all this, one has to admit that the argument that the increase of "relativistic mass" with velocity is in itself enough to discredit the notion of mass as "quantity of matter", loses something of its force. On the other hand, the very fact that the interpretation of "mass" in relativity seems to be a matter of taste can hardly be said to give much support to any association of mass with "quantity of matter." But let us assume for the sake of argument that some of the objections mentioned above to "relativistic mass", especially those labelled (iii) and (vi), have fundamental validity, and that we really should reserve the term "mass" for "rest mass" only. How does this alter the situation, if at all? It is true that by this means one avoids the seeming embarrassment of an increase of mass with velocity. But there are other difficulties to be faced: rest mass is not conserved in inelastic collisions, the total rest mass of a system of particles is, in general, less than the sum of the rest masses of the individual particles, the rest mass of a system increases if the system is heated - none of this encourages one to set too much store on an interpretation of rest mass as "quantity of matter."

To sum up: we have shown that neither in classical nor in modern physics does the idea of mass as "quantity of matter" stand up to critical examination. We conclude with Lindsay and Margenau that this idea is indeed a "wholly unnecessary absurdity."

/3.7.3.....

3.7.3 How should mass be introduced?

After this essay in what some will undoubtedly (though mistakenly, we think) regard as windmill-tilting, we turn with some relief to the problem of how one might teach mass. Neither the importance nor the difficulty of the concept can be overestimated. If it has taken centuries for man to evolve the concept and achieve some degree of clarity about it, and if some respected teachers still show confusion when presenting it, then there is obvious need for caution. The chief short-term goal of teaching "mass" at school is to enable pupils to appreciate its intimate association with inertia and, allied with this, its use in quantitative dynamics. A secondary aim would be to bring out the role played by mass in the phenomenon of gravitation. If a teacher can achieve as much as this with the average pupil, he will have done well. How can it be managed, and, in particular, how should one start?

It is useful to look first at some of the methods in current use. We have already noted that in the PSSC course, (PSSC 1960, 1965) the initial emphasis is on the alleged association of mass with "quantity of matter". In fairness, one must point out their use of this approach is as good as one can hope for, and that there is no question of their defining mass as "quantity of matter": the only definition of mass that appears in the chapter under discussion is "Gravitational mass is what we measure on the balance". And their purpose in introducing mass this way has an interesting motivation; it precedes a detailed introductory discussion of molecules and atoms, behind which lies the law of conservation of mass, with which is equated (as a good approximation), a law of conservation of matter. As they remark at the end of the chapter: "The conservation of mass (matter) is basic to the idea that chemical reactions are rearrange-

/ments.....

ments of atoms. It underlines the whole of the discussion of this chapter." In the physics course outlined in the "Physics is Fun" books by Jim Jardine (1964, 1965, 1966, 1967), mass is also introduced (in Book One) as a measure of quantity of matter, with the balance as the measuring instrument. Molecules and atoms have already made their appearance in the course, but no attempt is made at this point to link the discussion of mass with the constituents of matter. In a later chapter, the phenomenon of inertia is studied, and inertia is used as a basis of mass comparison. Since the underlying logic is of some interest and relevance, we outline it briefly here. Mr. Jardine invites his readers to take a pair of large cans, fill one with sand or water and leave the other one empty, suspend them from the highest point available, and try to push each can in turn. He then asks them to answer a number of pertinent questions which draw attention to all the important points that need be noted for the moment. Finally, however, he asks "If you were in a spaceship where the two cans were weightless, how do you think you might tell which had the greater mass?" Of course, the expected answer is "try pushing each can in turn and see which is easier to push", or words to that effect. The reason why one approves this method of mass comparison is that the mass of a body is a concept which is associated with the inertia of the body, so that if one compares the inertias of two bodies, one must be comparing their masses. But consider how the situation must appear to a pupil following Jardine's course. He has been introduced to mass as "something-to-do-with-quantity-of-matter-which-we-can-measure-with-gravitational-attraction". So at this stage he thinks of mass as "quantity of matter". Now he does Jardine's experiment with the cans, which shows him that the difference in inertia of the cans is related to their

/difference.....

difference in mass or (as far as he is concerned) the difference in their "quantity of matter". To enable him to appreciate the implications of this more fully, he is asked to imagine the same experiment being carried out in a spaceship. This, it is hoped, will make him realize that the comparison of the "quantities of matter" of different bodies can always be accomplished by comparing their inertias, even if a balance cannot be used.

In our view, Jardine's approach gets the logic of the problem the wrong way round. If mass is fundamentally a measure of inertia, then this idea ought to be the <u>starting-point</u> of one's discussion of mass. One can show, if one wishes, that the inertia of two similar bodies taken together is greater than the inertia of either body, so that the "quantity of matter" in a chemically homogeneous body is directly dependent <u>on</u> its mass; but it is seriously confusing to reverse the procedure by starting from a postulated association between mass and an unexplained "quantity of matter" and then showing that inertia is dependent on mass "defined" as quantity of matter. Insistence on this latter approach can lead to the kind of trap into which Jardine himself falls when, in a later volume (1966) of his course, he refers to inertia as a "property of mass." Near the end of the chapter in which he gives the above approach to mass, Jardine states that "The greater the mass of a body, the greater is its inertia. In fact mass is a measure of inertia." For him, these statements are a summary of a relationship between inertia and mass that is supposedly a deduction from experiment. It would have been better if they had been treated instead as a summary of what mass <u>is</u>.

The importance of mass is recognised in the Nuffield Course by giving it repeated attention: "In our programme,

/the.....

the concept of mass receives more and more attention from year to year". (Nuffield Guide IV). In the first year, pupils do some weighing, but no great fuss is made about what exactly is being measured this way. The issue of "mass" versus "weight" is allowed to lie dormant until someone forces it by asking "What is the name of the number that we get?" when, says the Nuffield Guide, "the best answer is MASS - with a promise of explanation later." In Year II the concept is taken no further, the attitude at this stage still being that "we should try to use 'mass' in the proper places without much comment on it." In Year III mass becomes the object of more serious, though still fairly informal study. Some simple experiments with "trolley and tickertape" are used to give pupils some "feel" for the concept of inertia and of the relation between force, mass and acceleration. Another notion may also be introduced: "... mass may have been described as a measure of how much stuff is piled together when we try to accelerate several trolleys". In Year IV, the emphasis is on inertia, with numerous experiments and demonstrations to bring home the idea, and the quantitative relation between (inertial) mass, force and acceleration is experimentally established. (Incidentally, Book IV of the Nuffield Physics Guide makes it clear by implication that the phrase "amount of stuff" is used essentially as a synonym for mass. Cf. the following passage: "The pull of the Earth's field is proportional to the amount of matter-to-be-pulled. The mass to be accelerated is proportional to the amount of matter-to-be-accelerated. We have no guarantee... that these two kinds of 'amounts of matter' are equal, or proportional, for all different materials.")

In a syllabus drawn up by the Science Masters' Association and The Association of Women Science Teachers of Great Britain (Science Masters' Association 1961), mass is

/apparently.....

apparently intended to be introduced by way of its association with inertia. The relevant "note" in the syllabus reads: "Mass is a property of a body, and depends on inertia, which is the property of continuing at rest or with constant speed in a straight line unless a force is applied." Among approaches in use in South Africa, we may mention the introductory approach followed by Brommert in his new textbook for South African Schools (Brommert 1968). He first tells his readers, "You have learned to recognise small and large masses. You may even have established your own scale of masses by having learned to classify stones according to how hard you can safely kick them. You have measured, very crudely but nevertheless measured, mass." One could argue that this is a rather terse introduction to a (presumably) unfamiliar physical term. At any rate, after this brief reference mass is not dealt with again until the fifth chapter of the book, which is entirely devoted to the topic. Here the author explains that his main concern for the moment is not to "give pseudo definitions of mass" but to tell readers something about it, the problem of possible definition being left to a later stage. "Telling something about it" is chiefly taken up with mass in its gravitational aspect, with the main emphasis being placed on techniques of measurement. (As Brommert's book is likely to become a standard reference work for South African teachers of physical science, it is possible that some at least of his suggestions may become widely adopted.) Finally, it is of interest to mention the approach to mass given in a book on relativity written for a lay audience by Bondi (1965a). (Originally, the material of this book appeared in a series of articles published in the "Illustrated London News".) How does a prominent scientist with a gift for communication tackle the problem of explaining mass to the uninitiated?

/In.....

In Bondi's case, he goes straight for the inertial basis of mass: "As for mass, we are all familiar with the fact that the same spring extended the same way will move some objects much less rapidly than others, and we call the objects that move sluggishly massive, while the ones that move rapidly are called light." (In the next sentence, this idea is much more precisely but less concisely expressed.)

These are some of the ways in which the concept of mass is introduced. Are they and other possible ones all equally acceptable? A few years ago, the present writer expressed himself as follows on the matter: "Different approaches are, of course, possible and probably equally valid. Some teachers may prefer to give some picture of mass by introducing it as 'quantity of matter', treating this as a concept to be refined on and ultimately refined away and replaced by something better. Others again may wish to give mass operational significance from the start by discussing it in terms of the inertia of a body, pointing out that it is easier to throw certain objects forward than others and explaining this as a difference in inertia between the objects... Individual teachers will be the best judges of what should be served to their classes as the first course of this particular meal." (Helm 1963). No doubt many teachers were very happy to agree with these views. The writer himself, however, is no longer so sure that he can endorse them. The trouble is that too casual an attitude to the underlying logic can easily lead to the kind of confusion which could prove a real block to understanding. We have seen this happen in a particularly glaring manner in Jardine's approach above, in which mass is defined by implication with reference to quantity of matter and gravitational attraction, and it is then suggested that the inertia of a body depends on its mass, instead of the other way round. This mistake is not

/made.....

made in the Nuffield and PSSC approaches; although the latter starts with mass as "quantity of matter" and has it measured on a balance, when dynamics is studied inertial mass is introduced quite independently of the previous ideas of mass.

There are some who might wonder what all the fuss is about. "Surely", one can hear them saying, "there is no real difficulty here. We all know that the balance is an excellent instrument for measuring the masses of macroscopic objects. Why can we not be happy with this and accept it from the start, treating mass as defined in terms of operations involving gravitational forces? Is there any real need to define mass afresh in terms of some other, less easily graspable operations, which only give us back the same answer in the end?" We deal with a more fully-developed version of this argument in Section 4.7.10. For the moment we stress that to proceed in the manner just suggested obscures a most significant fact of nature, viz. the fact that all freely-falling bodies in the same region of space have the same acceleration. At a time when physicists are still actively interested in establishing this result to the highest possible degree of accuracy (cf. Dicke 1961, Roll et al 1964), it seems more than a little silly to assume it without any kind of proof at the beginning of one's physics course.

Ideally, one would wish to teach mass from its inertial aspects, developing it quantitatively within the context of dynamics and then establishing the relation "W = mg" by experiment. This is undoubtedly the way in which it should be done in the senior secondary school years (Standards Nine and Ten), and we shall in due course examine possible ways of treating mass along these lines. But, with regret, we have to conclude that this is not a realistic way of

/introducing.....

introducing the concept in a typical school course in physics. It seems generally agreed that pupils need a certain maturity in order to achieve more than an entirely superficial understanding of Newtonian mechanics, and the fact is that some knowledge of mass is needed before it can be assumed that this maturity has been attained. It is needed as a background to elementary work on the measurement of force and the calibration of spring-balances, which will in turn be needed for an introduction to energy and work; it is needed also in elementary calorimetry. The easy way out is that already in use for many years: start with mass as quantity-of-matter measured with a balance, forget about inertia altogether and damn the consequences. This no longer merits serious consideration, as the foregoing arguments should have shown. Some sort of compromise is evidently necessary. One possible procedure is the following:-

(i) Start by introducing the pupil informally to the association of the name "mass" with the property of inertia (not necessarily to be dignified with this name at this stage) (cf. the quotation from the "Teachers' Guide for Physics" given earlier in this sub-section). Point out that we can devise a way of measuring this property directly (e.g. by a sophisticated version of Brommert's difficulty-or-ease-of-kicking method), but that this will take us into too deep waters for the moment, so we wait until later to discuss it.

(ii) Suggest that there seems to be a correspondence between the ease with which we can set a body moving and its heaviness - the ease with which we can lift it. Point out that we can compare the "heaviness" of things quite simply - by using an equal-arm balance. Analyse the process of comparison of heaviness of objects with an equal-arm balance. That is, take two objects of the same material and of as

/nearly.....

nearly the same dimensions as possible and suspend them at equal distances from and on opposite sides of the centre of a metre-stick pivoted about this point. Since the two objects are as nearly identical as possible, they must have the same heaviness; the experiment shows that when two objects of the same heaviness are suspended at equal distances from and on opposite sides of the centre of the metre stick, the latter is in equilibrium in a horizontal position. So we have the basis of a general principle for comparing the heaviness of two objects: if <u>any</u> two objects balance each other on an equal-arm balance, their heavinesses are the same.

(iii) Discuss the possibility of comparing <u>masses</u> by comparing heavinesses. Suggest that since there does seem to be a definite correspondence between the inertia and the heaviness of a body, it is perhaps reasonable to suppose that if two bodies have the same heaviness, their masses must be the same. Make it clear that this relationship between heaviness and mass is not to be taken as self-evident - that it will have to be put upon a sound footing in due course.

(iv) At this point, or earlier in the presentation, a fuller account can be given of what this comparison with an equal-arm balance is actually concerned with. It is really a comparison of forces - of weights, in fact; so the distinction between mass and weight has now to be drawn, and the point has to be made that in using the balance we assume that when the balance tells us that the weights of two things are the same, their masses must be the same as well. We should raise a warning flag (if we may be allowed to use Nuffield terminology) and emphasize that this is an <u>assumption</u>, proof of which will be forthcoming later.

(v) The way is now clear for further discussion of the principle of "weighing" with the equal-arm balance. The

/idea.....

idea of a standard of mass can be introduced, and the means whereby, in principle, a set of "weights" could be manufactured and compared with a standard kilogram can be explained: two objects which balance each other and, when combined, balance a mass of 1 kilogram have masses of 0.5 kg each.

This approach succeeds in the main preliminary objective of an introductory presentation of mass: it establishes a link in the pupil's mind between mass and inertia, and it gives a method of measuring mass. While the latter depends on an unproved assumption, the assumption is clearly stated and the pupil has the assurance that the matter will be taken up again in its proper place. No attempt is made to lean on "quantity of matter" at any stage of the argument. In essence, this approach is very close to that of Brommert described above. It is not free from objection: in particular, it is regrettable that although mass is introduced via inertia, the method of its measurement which the pupil first encounters depends upon gravitational effects and, moreover, upon an assumed proportionality between weight and mass. Unfortunately there does not seem to be any obvious way of avoiding this difficulty if mass is to be introduced in the junior secondary school and if, as we suggest, pupils are to be made to realize from the beginning that the term "mass" is to be associated with inertia. They need to be able to measure mass, and the only method they can be expected to understand at this early stage is the gravitational one. This places a considerable responsibility upon the teacher, who will have to work hard to ensure that his pupils retain their grip on the concept of mass as something associated with inertia while their method of measuring mass makes use of the "heaviness" of the objects concerned.

/A.....

A possible alternative to this approach is to concentrate entirely on the gravitational aspect in an introductory presentation, so that inertia is simply not discussed at all. When Newtonian mechanics is discussed, the term for which the symbol "m" is used in the equation "F = ma" can at first be identified with a "measure of the inertia of the body"; when the behaviour of bodies in free fall is discussed, the identification of this concept with the previously-introduced "mass" can be effected. Summarized, the discussion might take this form: "We have shown that there is a property of matter associated with its inertia, which we can measure by dividing the net force on a body by its acceleration. We learnt in earlier standards that there is a property of matter called mass, which is proportional to its weight - the force with which the earth pulls on it. Since freely falling bodies fall with the same acceleration in the same place, it must be true that the weight of a body must also be proportional to the quantity we associate with its inertial property; when we put all this together, we see that the quantity associated with the inertia of a body is in fact proportional to the mass of the body." This approach is logically consistent; it has the merit of dealing at length with one aspect of mass before going on to deal with the other, but it inverts the usual order of doing things, and one suspects that few teachers would be happy with it. Somehow it seems neater and more natural to argue from the facts of free fall that the equal-arm balance really does compare <u>masses</u> (in the inertial sense) than to argue from these facts that, say, an experiment in which we compare "inertias" by a dynamical method really does compare masses (in the gravitational sense). Incidentally, it might be wondered just now the approach just described differs from that of Jardine, concerning which we were somewhat critical. The difference is this: that in the above

/approach.....

approach, the pupil would know exactly where he was at any stage, from the point of view of assumptions made and definitions used; whereas Jardine's approach does not convey the same certainty. Starting with an implied definition of mass in terms of weight, he later shifts his ground to mass implicitly defined in terms of inertia, and concludes from free fall experiments that they "[justify] our earlier assumption that we can use weight as a measure of mass" (Jardine 1966). Assumption or definition? At first it was one, then the other.

So far we have avoided explicit reference to "inertial mass" and "gravitational mass". We have done this for two reasons: first, because no one would dream of using these terms in an introductory treatment of mass; second, because there are some authorities who maintain that to speak of two kinds of mass is to make a simple matter seem needlessly complicated. The latter argument must be seriously considered, and it seems better to deal with it when a more sophisticated approach to mass is under discussion. (See Section 4.11.2).

With the introduction of weight as an acceptable measure of mass, the way is open for a more extended qualitative discussion of gravitation, especially the apparent dependence of gravitational pull on [inertial] mass. In the now superseded Joint Matriculation Board Physics syllabus of the early- and mid-nineteen-sixties, there was actually a syllabus topic for the junior secondary school entitled "mass discussed from the gravitational standpoint", and the corresponding section of the Teachers' Guide suggested as an "interim definition of mass" the statement that "the mass of a body is that property of it that determines its gravitational attraction for other bodies". As the above discussion has made clear, we now take the view that it would after all not

/help.....

help very much to discuss mass in this way in the early stages, and certainly not to define it in terms of its gravitational aspects. (The above definition is, of course, too loosely worded for comfort; the mass of a body, as opposed to a particle, is not the only property of it that determines its gravitational attraction for other bodies; its shape and distribution of density play their part also). It is, in fact, the intention here definitely not to try to define mass. Some feeling for its nature and a knowledge of one way of measuring it will suffice for the moment. We will take up later the problem of whether mass can be defined at all.

A practical point of some importance arises in connection with the treatment of the technique of measuring mass with the aid of the balance. Ten years ago the equal-arm balance was still the principal instrument for measuring mass in the laboratories in South African schools, as it no doubt was in their counterparts in other countries. Today the situation is vastly different; science teachers have turned eagerly and with obvious relief to the simpler forms of balance now available from many suppliers. "Simpler," of course, means "simpler to operate" and not "simpler in principle"; and this is precisely the problem. While the direct-reading type of balance makes life a lot simpler for both teacher and pupil - the latter does not have to spend a protracted time in making a mass measurement and then make sure to add up the masses correctly, the former does not have to strain ~~any~~ every nerve to persuade him to do it properly - this gain is achieved at the expense of some loss in obviousness in principle of operation; the new types of balance are "black boxes" to some slight extent. The point is easily got round, however; all that is needed is for pupils

/to.....

to convince themselves initially of the soundness of the calibration of the balances by using them to measure known masses (e.g. the "weights" from an old-fashioned weight box). For the rest, we heartily approve this new approach to "weigning" in the elementary laboratory; it saves a great deal of time and trouble, while still yielding a degree of accuracy which is more than adequate for any mass measurement carried out in the school physics laboratory. (Those who want to do highly accurate weighing for whatever purpose can be referred to the chemistry laboratory, where they will almost certainly find the proper instruments.) The time has certainly come for lengthy treatments of the operation of the chemical balance, including discussion of sensitivity and stability, to disappear from syllabuses and textbooks, if they have not already done so. There are so many other matters which are both more exciting and more important that this loss would not be noticed at all.

3.7.4 Mass and the teaching of chemistry

We make bold to say that in spite of the difficulties it presents, the proposed treatment of mass we have outlined is a satisfactory way of introducing the concept to a physics pupil and would be found acceptable by physicists. Would it be acceptable to chemists also? Or would they be disturbed by its firm rejection of an association between mass and "quantity of matter"? We saw above that in the initial approach to mass in the PSSC course, mass was linked with "quantity of matter" so that "conservation of mass" could become "conservation of matter", and this in turn could be taken as "strongly suggesting that chemical change is a process of rearranging component parts." This, of course, was Lavoisier's view of the facts. Nowadays, while acknowledging the validity of the link between conservation

/of.....

mass and the idea of "rearrangement of component parts", one is not bound to base one's supporting argument on any initial assumption that mass is a measure of an ill-defined "quantity of matter"; that is, the association of mass with "quantity of matter" (in whatever sense) does not have to be treated as fundamental. Instead the teacher can start by taking from physics the fact that there is a property of matter called mass which is associated with inertia and for which there is a method of measurement. Physics provides two further experimental facts about mass that are relevant, viz:

(i) that the total mass of n similar particles is equal to n times the mass of a single particle; and

(ii) that if the total mass of a collection of dissimilar particles is measured, it is found to be equal to the sum of the masses of the individual particles.

(As we have pointed out above, these relations are in fact only approximate, but as far as their chemical applications are concerned we may take them to be exact.) The teacher can then draw attention to the additional experimental fact from chemistry that mass is conserved in a chemical reaction, and suggest that this should be interpreted in terms of some theory of the composition of the matter, making use of the above experimental facts about mass. A possible theory is that matter is made up of "ultimate particles" (Dalton's phrase) which are neither created nor destroyed in a chemical reaction, but simply rearranged. The "additive" property of mass, embodied in facts (i) and (ii) above, means that it is a simple consequence of this theory that mass must be conserved in a chemical reaction. It may not be the only possible theory, but it certainly fits the facts. The teacher and his pupils can then go on to extend this reasoning to encompass the law of definite and multiple

/proportions. .....

proportions.

Once the atomic theory is accepted, fact (i) above justifies one in asserting that finding the mass of, say, a lump of iron is a method of counting the number of atoms in it or, more generally, that finding the mass of any homogeneous substance is a method of counting the number of "molecules" in it. (Cf Sears 1964). If it be claimed that this amounts to acknowledging the relationship between mass and "quantity of matter" (in a limited sense), we have no quarrel with this; by "quantity of matter" is now meant something precise, viz. "number of atoms" or "number of molecules." But clearly this relationship between mass and number of atoms or molecules need not and ought not to be connected in any way with one's initial thinking about mass. We have in fact shown that the proportionality between the mass and the number of atoms or molecules of a homogeneous substance is a plausible deduction from certain experimental facts concerning mass and this, in our view, is how it ought to be treated, even in chemistry. Since our primary concern in these pages is not with the clarification of chemistry teaching we will not press this point further; but in answer to our original question, it seems to us that there is no good reason why the approach to mass outlined in 3.7.3 need not be acceptable to chemists, who could build on it to suit their own purposes in the manner just described.

3.7.5 Experimental work

Finally we consider some demonstrations and experiments that could be done to reinforce the very important ideas introduced here, namely association of mass with inertia, and the measurement of mass by means of a balance. Preferably, the concept of inertia as such should receive some attention in class before it is suggested to pupils

/that.....

that a measurable physical property can be associated with inertia. We will assume that this background has already been filled in by the teacher (we deal with it ourselves in discussing force in the next section) and consider only the association of mass with inertia. The task is to develop the idea of "relative-ease-of-starting-or-stopping-different-things." Obviously experiments chosen to illustrate this should be ones in which friction plays a minor part. There are a number of possibilities; we have already mentioned the suspended tin cans, one full, the other empty (Jardine 1964, Nuffield Guide IV), Brommert's "relative-ease-of-kickability" of stones idea (Brommert 1968), and the comparison of the ease with which different objects can be thrown forward (Helm 1963). In any such demonstrations, it is preferable that the pupil should not have any reason beforehand for thinking that one of the objects used is heavier than the other. These preliminary comparisons should concentrate on differences in inertia only. We may add to the previous short list the following suggestions:-

(i) Take a pair of balls of similar external appearance and size and invite pupils to set them rolling.

(ii) Place a pair of identical-looking closed tin cans next to each other and the edge of the table and ask a pupil to snatch them simultaneously towards him.

(iii) Do demonstration (i) with a pair of similar toy cars or trolleys, one of them weighted internally with lead.

(iv) Carry out a more sophisticated version of (iii) but this time have the cars or trolleys pushed in turn by a spring compressed to the same extent, or exploded apart by a spring. Perhaps one could say to pupils beforehand "It would be a good idea to give both cars the same push. But can we be sure that two successive pushes with our fingers have the same strength? Can you suggest a way of getting

/something.....

something to push with the same strength on the two cars at once, or on two cars successively?" This demonstration represents a step away from a purely anthropomorphic approach to the problem, and as such it might be a good idea at this stage.

Having come as far as this, some teachers might wish to press on and have their pupils measure mass with the PSSC-type inertial balance (cf. Section 4.7.12). This might seem a good idea - it affords a reasonably precise method of measuring mass without reference to gravitation. Unfortunately the technique is a little sophisticated for the full value of the method to be appreciated at this stage of the pupil's understanding of mass. He has first to obtain a calibration curve by timing the period of oscillation of the balance for several different numbers of similar masses on the balance; then he measures the unknown mass by placing the object in question on the pan, finding the new period of oscillation, and reading off the mass from the graph. Possibly a bright group of pupils might cope with this, but it seems more realistic to suppose that the average junior secondary school pupil is not fully ready for this method of measuring mass.

Once pupils have gained a semi-quantitative feeling for the idea of inertia, the correspondence between it and "lightness" or "heaviness" ought to present no difficulty. There remains the gaining of familiarity with the techniques of measuring mass with a balance. If, as we have implied, the chemical balance should play no more than an introductory role here, there is no good reason why the pupil should have to spend a lot of time in the laboratory learning how to "weigh". The technique of using pointer-type or triple-beam direct-reading balances can be very rapidly acquired and the point need not be laboured. What else can be done about "weighing"at this stage, besides teaching pupils how

/to.....

to do it? The traditional early application of weighing is the determination of the densities of various materials. The Nuffield Guide (I) makes out an exceedingly strong case for not dealing with density too early in a pupil's career and, although their discussion is concerned with eleven-(plus)-year-olds, we can perhaps take the hint and not place too much emphasis on density and its determination in this particular context. A quick review of the method of determining the density of a regular solid, and a measurement of the density of air (cf. the Nuffield Guide I for an excellent method of doing this) will suffice. For the rest, it is desirable that the pupils have some fun with weighing. If by the time they reach the secondary school they have not had the experience of using a home-made microbalance to weigh extremely light objects, this omission should certainly be rectified, and the present part of the course is a good place to do it. (The P.S.S.C. laboratory Guide and the Nuffield Physics Guide to Experiments I both give details of the construction of the cool-drink-straw type of microbalance). As the Nuffield Guide points out, there are possibilities here for looking forward towards other aspects of the course- use of the microbalance for the measurement of electrical forces or for close observation of the evaporation of liquids are only two of the ideas that might be stimulated by work with this instrument.

3.8 Force

3.8.1 Introduction There are many reasons why the concept of force should start to impress itself on a pupil's consciousness at a very early stage in a physics course. First and most fundamental - according to one viewpoint, we can sum up the whole of present-day physics in the four types of interaction between particles currently known:

/gravitational,.....

gravitational, electromagnetic, strong and weak, the latter two being associated with the nuclear domain. (Cf. Feynman et al 1963). Second and almost as fundamental - in a sense, forces are but one step removed from fields of force, and the concept of a field is of crucial importance in contemporary physics. Third and more mundane - forces are part of our everyday experience and it is comparatively easy to establish a link between this experience and the beginnings of a scientific concept of force. Fourth and highly practical - the energy concept is needed early in any modern physical science or physics course, and it cannot be given proper quantitative meaning without the aid of a quantitatively-established force. Last for our present purposes - it offers an opportunity at an early stage for suggesting the unity of physics.

Fortunately the concept of force does not seem to present us with such awkward problems in presentation as does the concept of mass. Although it would be far from true to say that there is unanimity in the scientific world about the meaning and status of the concept of force (we shall take up the controversial aspects of the subject later), no one seriously contests the validity or suitability of the usual anthropomorphic introduction of force as a "push or pull". We may not be happy with this description as a <u>definition</u> although, as Margenau has pointed out, it can be defended as such if pushes and pulls can be assumed to be measured (Margenau 1950); but this does not really matter at this stage. What is wanted initially is not a precise definition, but a statement about the concept that contains no glaring absurdities and can serve as a basis for discussion of force and for development of a way of measuring it. "Push or pull" is satisfactory on all these counts.

/3.8.2 .....

3.8.2 Introducing force

To enable pupils to draw maximum benefit from an introductory treatment of force, it is suggested that such a treatment should cover the following points:-

(i) preliminary exploration of the concept of "push and pull";

(ii) study of the effects of forces (including Newton's First Law);

(iii) study of different kinds of forces;

(iv) calibration of a spring balance in force units and measurement of forces;

(v) pairs of forces and Newton's Third Law.

Point (i) can be speedily handled by means of very simple "force-feeling" experiments (the Nuffield Guide (I) describes this section as "fifteen minutes of play"). Rubber bands, strings and springs can be used to provide pulls for pupils to feel. Compression of a spring does produce a push of the spring on the agent compressing it, but this requires the pupil to decide whether what he is feeling is his push on the spring or the spring's push on him, and this might raise the spectre of Newton's Third Law a little too early for comfort. The Nuffield Guide recommends the repulsive force between a pair of bar magnets to illustrate pushes; as this at once draws attention to action-at-a-distance forces, it would be a doubly valuable demonstration.

A weight held in one's hand can be used to demonstrate another action-at-a-distance push or pull. The Nuffield Guide suggests it as a means of feeling "the Earth's pull", though here the pupil is actually feeling the push of the weight on his hand; it requires an act of imagination on his part to connect the pull of the earth on the weight with the push of the weight on his hand. Association of the name "force" with pushes and pulls is virtually all that is

/needed.....

needed to clinch tnis stage of the argument.

Jardine (1964) poses the following question to the pupil near the beginning of his introductory chapter on force: "How would you recognize a force if you met one?" This sort of question could form an excellent link between the introduction of pushes and pulls and the study of what they do. Is a force to be identified as such only if we feel it? Discussion of the stretching of a spring by a weight hung from it or the effect of a magnet held in the hand on a nearby steel pin or needle can point the way towards the identification of the presence of a force by what it does to things, not just by feel. This is a highly important step in the direction of abstraction from experience, but it is a step which pupils seem able to take easily enough. The study of the effects of forces should encompass stretching (with springs or rubber bands), twisting (with thick rubber tubing), turning (with a simple lever arrangement) and most important of all, the effect on motion. In our view, the study of the first three of these effects need not be too protracted at this stage; a quantitative study of the turning effects of forces is obviously best postponed until force units have been introduced, and the other two need not be explored in depth either. It is the relationship between force and motion on which the stress now should be laid.

"The relationship between force and motion" is, of course, Newton's First Law in disguise. The main points to be made here are that a body has <u>inertia</u>, that is, it tends to continue in its state of rest or uniform motion in a straight line, and that a body will be accelerated if subjected to a resultant force. In regard to the first point, we suggest strongly that this be approached via the pupil's everyday experience, in particular of things that

/have.....

have happened to <u>him</u>. There are many things that will have happened to the average "European" pupil, boy or girl, that are readily explicable in terms of "inertia", even though they may not have obvious reference to a tendency to continue in "uniform motion in a straight line". We may agree with the historians of science that it was really quite natural for Aristotle and his contemporaries to think that circular motion was the natural motion of celestial objects, that a force was needed to maintain motion, and that heavier bodies must fall faster than lighter ones (cf. Toulmin and Goodfield 1960); but this need not mislead us into having our pupils initially follow Aristotle's footsteps when interpreting their own experiences. The pupil's attention can be directed towards those of his experiences which strongly support the idea of inertia; given the right kind of guidance, he will soon find himself able to interpret other experiences in terms of the effects of force and inertia. If the attitudes of some first-year university students may be taken as typical, some pupils will be Aristotelian in their outlook on mechanical phenomena; they should be educated towards the alternative view without delay. In the jet age, pupils will have travelled on or in bicycles, motor cars, buses, trains and aeroplanes. If they have had to stand in a crowded train or bus, they will know very well what happens to them if the brakes are applied suddenly, or if the vehicle gets off to a jerky start; if they have ever jumped off a moving bus, they will have some idea of the dangers attached to this procedure; they will have some idea of the function of safety-belts; and they will be familiar with the principle that the faster a vehicle is moving, the harder it is to stop it by applying the brakes. These are just some of the points that could be made in the process of extracting inertia from experience. Then the

/pupil.....

pupil can witness some of the now-celebrated demonstrations of near-frictionless motion and be invited to compare what he sees on the demonstration bench with his experiences when free-wheeling on his bicycle on a level road. The problem arises here of at what point the idea of friction should be introduced. We suggest that even though the word is probably part of the pupil's vocabulary already, it should not be used when these demonstrations are first shown. The motion of a dry ice puck, air puck or glider on a linear air track should be taken as what it seems to be, viz. uniform motion in a straight line. When the problem of the free-wheeling bicycle or car or toy is discussed, it can be hinted that a force is responsible for the retardation and the culprit identified as "friction".

Demonstrations of the effect of force on motion can now be given, preferably under conditions in which friction plays little or negligible part. Tentatively, the relation between force and acceleration can be introduced here. As yet, acceleration is not a precisely-defined concept for the pupil; but for the purposes of the moment it does not have to be. The pupil already knows the word; he associates it with what happens when a motorist "puts his foot down", in other words with "going faster". (Admittedly he may already have been subjected to confusing influences about it, such as the statement that "everybody knows it is possible to go faster downhill without accelerating", but these need not be serious obstacles in the path of proper understanding). Now he can be invited to study the connection between "going faster" or "going slower" and the application of pushes or pulls. The demonstrations he sees, the little experiments he may try and recollection of his own experiences will all help. Regarding the last-mentioned

/point.....

point, he will know that a passenger in a motor-car travelling on a smooth, straight, level road only "feels something" when the car is being accelerated or decelerated; that, for example, a passenger feels the car seat pushing against his back when the car is being accelerated. He will also have heard of the effect of large accelerations on the occupants of space vehicles. On this evidence he should be prepared to accept that pushes or pulls "cause" accelerations or decelerations - that they can slow things down or speed them up. But this study of the relation between force and acceleration should not be confined to problems of change of speed. Among the demonstrations or experiments he sees done should be ones in which forces change the direction of motion of objects, without necessarily changing their speed. (Cf. Jardine 1964, in which this kind of demonstration is done in a game of "blow football" with a table tennis ball and drinking straws). It is not essential for the teacher to describe what happens to these objects as "acceleration" - that can come later. But if he leaves his pupils with the strong impression that forces can change direction of motion as well as speed, he will have laid a good foundation for later understanding of acceleration as a vector quantity.

This work on force and motion also provides a good opportunity for the introduction of the idea of a net or resultant force acting on something. Elementary instances of the failure of a push or pull to produce any motion at all can be explained as the effect of opposing pushes and pulls acting on the body in question.

3.8.3 <u>The measurement of force</u>

After this more sophisticated discussion of force and its effects, attention can be given to the problem of <u>measuring</u> force. How splendid it would be if the newton

/could.....

could be introduced in a manner simple, direct and unambiguously related to its fundamental definition. Much beating about the bush and creation of possible confusion could thereby be eliminated. Unfortunately this is simply impracticable at this stage of the pupil's knowledge; to understand the newton in terms of its fundamental definition, he would have to have some experience of kinematics, for which the syllabus, realistically enough, makes no allowance at this point. So recourse has to be had to some interim approach; none are entirely satisfactory but the better ones, if carefully taught, need not give rise to trouble later on.

The basic principle of all such interim methods of measuring force is the use of a spring balance calibrated by hanging weights on it. The purist is instantly suspicious of this, and with good reason; he fears a premature assumption of proportionality of weight and (inertial) mass, with the risk of circularities in the logic when the subject is studied in a more advanced way later on. Unfortunately he is unable to propose anything better. We must therefore accept the limitations of this approach and do our best within them. The safest procedure from the point of view of logic is one that is, in fact, described in Book II of the Nuffield Guide and in the first-year university physics textbooks of Sears and Zemansky (1960,1964). The present writer described it in the first draft of his Teachers' Guide (Helm 1963), then suppressed it on the grounds of its possibly being too subtle for the age-group for which it was intended. Emboldened by the Nuffield precedent, he is now happy to endorse their recommendation. The method can be described as follows:-

Take a number of brass blocks or cubes having as nearly

/the.....

the same dimensions as possible and place them in turn on the scale pan of a spring balance, noting the position of the pointer of the balance in each case. [This preliminary step should confirm the expectation of the class that the spring balance will be extended by tne same amount each time.] Then place in turn two, three, four and five of the blocks in the scale pan, and once again note the position of the pointer in each case. Regard any one of the blocks used in the experiment as a standard object having a weight of one unit of force. Then make a force scale for the balance by marking the successive positions of the pointer noted above as 1, 2, 3, 4 and 5 units of force respectively. Any pull on the balance (provided it does not extend the spring beyond the calibrated part of its scale) can now be measured in terms of the units of this scale.

Of this method of calibration, the Nuffield Guide(II) comments: "We certainly should not raise any question with children about tne assumption behind this calibration process, yet tne teacher should keep it in mind: we assume that four equal lumps hung together on a spring pull it down with four times as big a force as one lump. We have to make some assumption like this in constructing a scale of measurement for a thing like force." To this we would add the following observations: that with thirteen- or fourteen-year-old children it might not be unreasonable to mention the assumption behind this calibration process, even though we might not wish to raise any question about it; and that as described here, this calibration procedure does not involve any premature assumption of a general constant of proportionality between weight and mass. What it does assume is the additive property of weight for similar objects; and if we add to this the assumption of additivity of mass, we naturally find

/that.....

that weight and mass bear a fixed ratio for such objects. But this is no proof that if we were to take any other set of objects (made of aluminium, say) that weight and mass would bear the same ratio for them. This fact has to be established by independent means.

If no attempt is made to make the pieces of metal used in this experiment conform to any particular specifications of size or mass, the units of the force scale of the balance are arbitrary. This is no bad thing; these temporary units could even be given some special name, and the spring balance could be used to measure the weight of some object in terms of these units. The limitations of such a force scale, especially the dependence of its calibration on <u>where</u> this is carried out, could be discussed with pupils. This could lead straight to the introduction of a newton; the need for a "unit of force which would stay the same and be the same everywhere" (Nuffield Guide II) could be stressed, and the newton could be identified as the official unit. There still remains the question of how one should define it at this stage. One answer is to attempt no definition at all - to shelve the matter altogether for the moment. This is not as cowardly as it may seem; one could reasonably argue that since the correct definition cannot yet be given, it is of little real use to produce a temporary definition which might possibly be a source of confusion. Class work on the measurement of forces could be done entirely in terms of the arbitrary force-scale established by the preliminary calibration. Another possibility is to introduce gravitational units of force prior to any discussion of the newton, and then to express the newton roughly in terms of them. Essentially this is the approach adopted both by the Nuffield course and by the new syllabus for standards Six, Seven and Eight in South Africa. A better approach than either of

/these.....

these, in our opinion, is to manufacture a set of "standard objects" each of which would have a weight of 1 newton to a good approximation (their mass would have to be roughly 102 grams each). Then pupils could be told something like this: "This object has a weight of 1 newton. I can check this by hanging it on a spring balance that has been properly calibrated in terms of the correct newtons, although actually I don't have to go to such lengths - there are much simpler ways of knowing that this has a weight of 1 newton, which you'll hear about later. Of course, if this object has a weight of exactly one newton here, its weight won't be exactly one newton somewhere else. That means that if we want to calibrate our spring balance for very accurate work, we must be careful to do it in the right place. Luckily for our own purposes we need not be so fussy."
And so the discussion could proceed, leading smoothly on to the measurement of forces in newtons by the pupils. There seems a great deal to be said for this approach; it brings in the right units of force early on; it really removes any need to introduce gravitational units of force, with their acknowledged potential for confusion and the need to define them in terms of an assumed proportionality between weight and mass; and it cannot be said to confuse pupils in the way it is defined, since this "definition" in terms of simple operations need never be committed to print or memory. We seriously advocate it as a superior alternative to those usually offered.

With a spring balance calibrated in newtons, the teacher is able to suggest to his pupils that they may feel a force of one newton for themselves, by the simple act of pulling on the spring balance so that it reads one newton. This will help pupils not only to appreciate how strong a force

/of.....

of one newton is, but also to gain familiarity with this fundamental unit of force. The measurement of the weight of various objects in newtons will help teacher and pupil to keep the concepts of weight and mass well separated; if the pupil becomes used to measuring mass in kilograms or grams and weight in units with an altogether different name, he is not so likely to fall prey to the confusion which teachers have to labour so hard to prevent.

3.8.4 Newton's Third Law

The revised Joint Matriculation Board syllabus for pure physics included as one of its topics for the junior years, "Newton's Third Law and simple illustrations of its application". Newton's Third Law is generally conceded to be a puzzling topic for beginners, and indeed more advanced students sometimes stumble when pondering instances of its application. Professor Jay Orear remarks in the Preface to his book "Fundamental Physics" (Orear 1961, 1967) that liberal arts students taking a course in elementary physics "have more trouble in understanding Newton's Third Law than they do in understanding Fermi energy or charge conjugation invariance". This kind of experience might make one hesitate before trying to introduce even a simple treatment of this topic into pre-matriculation science. Yet we feel that, provided certain precautions are observed, the attempt would be well worth making. The fact of nature embodied in the Third Law, viz. that forces are always to be found in pairs, is fundamental to any serious study of forces and should surely be brought into the open as soon as possible. Any honest discussion of the use of the spring balance to apply a known force to an object cannot sidestep the point that the spring balance really measures the force that is exerted on it - that is how we usually calibrate it - and

/not.....

not the pull that it exerts on something else; in fact we implicitly invoke Newton's Third Law when we use the spring balance in the manner indicated. What of the situation when we push against a door with our hand - is it our own push that we feel, or that of the door against our hand? The latter answer is, of course, correct, and the moment we pose the question we are compelled to accept the reality of "forces in pairs".

Everybody has probably heard Newton's Third Law stated in the direct translation from Newton's original Latin, viz. "Action and reaction are equal and opposite." Here is one instance of what may be familiar to the pupils being of little or no use to the teacher. One may state with some confidence that this rendering of the Third Law forms a most unsuitable starting-point for a discussion of this topic; it surely conveys very little to any but those who already know what it means. In the Space Age, it is better to begin with the principle of the rocket or, if this is felt to be too advanced a starting-point, to take up the problem of how to get off the surface of a frozen lake if implements for obtaining a grip on the surface are not available. Jardine (1964) suggests that pupils try to approximate to this situation by carrying out the experiment of standing on roller skates or a trolley and throwing various objects. This seems an excellent way of clearing the ground for a discussion of the rocket principle. The author's own experiences in talking to scholars at a preparatory school on the one hand and intelligent laymen on the other suggest that these ideas can be appreciated if presented in simple terms.

Various questions arise in connection with the presentation of Newton's Third Law at an elementary level:-

(i) Should the <u>equality</u> of "action" and "reaction" be stressed at this stage?

/(ii).....

(ii) If it is, how does one justify it?
(iii) Are there any other examples besides those already mentioned that can usefully be cited here as illustrations of the operation of the Third Law?

To (i), we might answer that it is not <u>essential</u> for the <u>equality</u> of action and reaction to be brought home to the pupil at this stage; it is more important for him to be aware of the existence of force pairs than of the equality of the forces composing them. But the issue of the relationship between their magnitudes is bound to come up, and if it does there is no point in concealing the fact that the magnitudes are equal. This immediately brings up question (ii): how do we justify it? The author can remember reading as a schoolchild the correct explanation of the principle of the rocket and being puzzled, not by the connection between Newton's Third Law and the propulsion of the rocket - this seemed logical enough - but instead by the Law itself. It did not seem obvious to him, and he wondered whether some scientific or "common-sense" argument lay behind it which he for some reason was unable to see. Later he learnt that there was no question of any theoretical or even direct experimental proof of the Law; one accepted it because with its aid, one was able to give a satisfactory account of Nature. Perhaps it would be just as well for pupils to be brought face to face with this situation early in their careers, and so come to realize that there are physical laws or axioms which cannot be deduced from some previously-stated theory. The conscientious pupil may take some convincing that he need not, indeed cannot be expected to provide a justification for Newton's Third Law, but he will have learnt something of value when he is finally convinced.

On the various alleged experimental "proofs" of the Third Law involving the use of spring balances, we need not

/comment.....

comment in detail; we have already referred to the fact that this Law is implicitly assumed in the customary applications of the spring balance. Rogers (1960) has suggested that "we can show experiments that illustrate the idea of ACTION = REACTION even if they do not prove it", but after reading his lucid account of these experiments and their logical flaws one cannot help wondering whether it would not be better to omit such "demonstrations" altogether. Rogers also discusses another fallacious proof of the Law, viz. the object-resting-on-table "proof". Since it is likely to remain a potential source of confusion, and since it is still possible to find in textbooks such statements as "If a man stands on a level floor, his weight acting vertically downwards is the action of the man on the floor" (Stead 1948), we think it worthwhile to include some discussion of this "proof" here. While it is certainly true that the Law applies to an object at rest on a table as well as to almost everything else, the fact that the object is in equilibrium does not prove the law: it means rather that the force exerted by the table <u>on the object</u> is equal to the force exerted by gravity <u>on the object</u>. The forces just mentioned act on the <u>same</u> object, and so do not constitute an action-and-reaction pair in the sense of Newton's Third Law and cannot be used to prove it. The action and reaction pairs in this case are (a) the gravitational pull of the earth on the object and the gravitational pull of the object on the earth; (b) the force exerted by the object on the table and the force exerted by the table on the object. (The forces in (a) are gravitational, those in (b) elastic). Certainly in the special case of the object being in equilibrium on the table, the forces in (a) are numerically equal to those in (b), but this is not an illustration of the Law under discussion. Suppose that the object, instead of

/resting.....

resting on a table, is on the floor of an accelerating lift. Since the object is now accelerating, like everything else in the lift, the force exerted on it by the floor cannot be equal to its weight; it is still true, however, that the force exerted by the object on the lift floor is equal to the force exerted by the floor on the object. (Cf. the statement quoted above from Stead).

It is unlikely that a teacher who understands the above discussion of the object on the table would deliberately introduce this example in the course of his teaching of Newton's Third Law, for he will probably feel that its complexity is too daunting for a beginner. Yet an enterprising pupil is almost certain to raise it, and we should consider how the teacher might best answer it. Perhaps it is most important of all for him to seize the opportunity to stress the fact that action-reaction forces generally act on <u>different</u> bodies, whereas the equilibrium state of the object on the table shows the equality and oppositeness of two forces acting on the same body. (We say "generally" because some pupil might point out that if he pinches himself, the action and reaction forces act on different parts of the same body, namely his own.) To the pupil who asks "How does the table 'know' how strong an upward force it must exert on the object resting on it?", the answer can be given that the surface of the table is slightly depressed by the object lying on it, just as a bed "gives" slightly when one lies on it, or a trampoline when one stands on it; the deformation of the table surface will continue until the restoring upward force exerted by it is just enough to counterbalance the weight of the object.

In discussing the object on the table, we have given a partial answer to question (iii) above, which was concerned with examples of the applications of the Third Law. Naturally there are many examples that could be given - the

/importance.....

importance of the "reaction forces" in such familiar activities as walking or climbing, the recoil of a gun, the reaction of the water on the hose from which it is emerging, to name only a few - but it is probably unwise to pile on too many at this early stage. As for more complicated examples, such as the problem of the horse pulling a cart or an engine pulling a train, these are best left out unless the teacher is really confident of his ability to sort out for his pupils which forces act on which bodies. This is the crux of the matter: as Whitmore (1967) has pointed out, pupils and students seem to experience difficulty at first in considering parts of a system in isolation, and while this is a difficulty they will have to overcome, it seems prudent not to attempt too much in this direction at first.

Should the pupil be given a formal statement of the Third Law at this stage? We have already implied that it is certainly unwise to give Newton's original statement; we feel also that a <u>formal</u> statement is unnecessary, but that after having heard, experienced and pondered the ideas implicit in the Third Law, the pupil should devise his own informal statement of it, along the lines of "If I push on something with a certain force, that something will push back on me with an equal force". This will help to fix these important ideas in his mind. Later he can be guided towards producing a more formal version, e.g. "When body A exerts a force on body B, then body B simultaneously exerts an equal and opposite force on body A".

3.8.5 Summary of the introductory treatment of force

We recommend that in an introductory treatment of force, this concept should be informally "defined" as a push or a pull; that besides the more obvious examples of pushes and pulls, gravitational, magnetic and electric forces should

/also.....

also be demonstrated; that pupils should be made aware of the qualitative relationship between force and acceleration; that the quantitative nature of force should be made clear with the aid of a spring balance, calibrated in newtons by the method described above; and that the occurrence of forces in pairs should be pointed out and simple examples given of the application of Newton's Third Law.

3.8.6 The "turning effect" of forces

Both the South African junior secondary school science syllabus and the Nuffield Physics Project prescribe the experimental study of the turning effect of a force, leading to the establishment of the Law of the Lever (or "Principle of Moments"). One cannot reasonably object to the inclusion of this topic in an elementary physics course, as it provides a good opportunity for pupils to "discover" a simply-stated but important law through straightforward experimental work. It needs no special comment here, as no logical pitfalls lie in wait for the teacher who presents it. Later (Section 4.13) we take up the question of whether the Law of the Lever should ultimately be regarded as a "derived" or a "fundamental" Law.

3.9 Elementary notions of energy and work

3.9.1 Energy in present-day physics teaching

It is a noteworthy feature of modern courses in school physics that they allow a prominent place to the concept of energy. The P.S.S.C. textbook devotes lengthy chapters to work and kinetic energy, potential energy, heat, molecular motion and conservation of energy, and energy and motion of charges in electric fields. The textbook for a Canadian physics course closely resembling the PSSC course is called "Matter and Energy" (Maclachlan and McNeill 1964a). A spokesman for the Nuffield Physics Project has said that

/"the.....

"the topic of energy pervades the whole course" (Lewis 1965); a glance through the teachers' guides for this course shows that energy is explicitly treated in Year I ("A first look at energy"), Year II ("Energy changes, power"), and Year IV ("Universal conservation of energy"), and that it is used in support of much of the other material of the course. The reasons why energy should be given this kind of preferential treatment in elementary courses are obvious enough. It enables quite complex phenomena to be discussed in comparatively simple terms; it is found to be exceedingly useful in other sciences besides physics (e.g. chemistry and biology); and the Principle of Conservation of Energy is a fundamental physical principle of apparently unchallenged validity. What is new in the treatment of energy to be found in these new courses is the emphasis placed on establishing energy as a quantitative concept. In former days it was common practice in South African teaching of physical science to make free use of the concept of energy without at any stage showing how it could be measured. Needless to say, this approach to the problem added to rather than lessened the pupil's confusion, and one can only marvel that it took so long for its deficiencies to become officially recognized. The new attitude towards the teaching of energy should do much towards promoting a real understanding of the nature of physical concepts and of principles of conservation in physics.

It must not be supposed that the prominence given to energy in elementary physics teaching has been greeted with enthusiasm in all quarters. Thus both G.B Brown (1965) and Armstrong (1965a) have implied that there is something akin to mumbo-jumbo in present-day usage of the term "energy" and that in the new approaches to school physics energy is being made to appear what it is not, viz. something perilously like

/the.....

the "caloric" of the earlier theories of heat. We shall embark on a full-scale analysis and defence of the "orthodox" view of energy later in this thesis (Section 4.15); here we will content ourselves with remarking that in the hands of the careful teacher, there need be nothing corrupting about the modern way of presenting energy.

3.9.2 The introductory treatment of energy

A "modern way" of introducing energy that has found wide acceptance is that originally described by the P.S.S.C. and by Rogers in their respective textbooks, and which is now to be found in the Nuffield Physics Course (1966), Brommert (1968) and Maclachlan and McNeill(1964a), to mention only a few recent accounts. In this approach, one begins by talking about the difference between those jobs that require fuel and those that do not. Jobs in the former category are distinguished by the fact that when they are performed, the forces involved move their points of application in their own direction (in Rogers's phrase, "the forces move along"); for example, the lifting of a weight requires the application of a force having an upwards component. The association of such jobs with the consumption of fuel leads to the supposition that something might be transferred from the fuel to the system on which the job is done; this "something" is tentatively called "energy". ("At this stage 'energy' is a term that is used in describing activities that are accomplished or jobs that are done, and in describing the consumption of fuel needed for the job"- Maclachlan and McNeill 1964 b). This statement is far too vague to serve as a definition; it is an interim description intended to help the pupil feel his way towards an understanding of energy.

The quantitative development of the subject is begun by considering a job whose quantitative aspects are easily

/described. .....

described. An obvious example is the lifting of a load. Let us define a "unit job" in this connection as the lifting of a weight of 1 newton by 1 metre, and say that unit amount of fuel is required for the performance of this unit job. Then a job of lifting a load of 5 newtons by 1 metre is equal to 5 unit jobs (think of the job being done by lifting 1 newton at a time) and a job of lifting 1 newton by 3 metres is equal to 3 unit jobs, requiring 3 units of fuel (think of the load of 1 newton as being lifted 1 metre at a time). Therefore a job of raising a load weighing 5 newtons by 3 metres should require 15 units of fuel. This type of reasoning suggests that the amount of fuel used up in this kind of job is proportional to the product of the force applied and the distance moved by the point of application of the force in its own direction (if the load is lifted at constant velocity, which is implicitly assumed above, an upwards force equal in magnitude to the weight of the load must be applied to it). If the amount of fuel used for any job is taken as a tentative measure of the energy required for it, then the preceding argument suggests that the energy required for a particular fuel-consuming job is directly proportional to the product "force times distance moved along" for the job. "Force times distance moved along" is thus a measure of the amount of energy made available by the fuel for a particular job or, as we prefer to say nowadays, the amount of energy <u>transferred</u> from the fuel.

One advantage of this method of presentation is that it shows prior to any formal definition of "work" that the quantity "force times distance moved along" has physical interest and significance. In elementary teaching, this is much to be preferred to the practice of <u>initiating</u> a treatment of work and energy with a bald statement of the formal definition of "work".

3.9.3 The introduction of work

There are various ways in which the concept of work could be introduced following a discussion of the above type. For example, one might simply say at this point, "We've seen that fuel-consuming jobs are jobs in which 'forces move along', and that the fuel used up is proportional to the product 'force times distance moved along'. We give a special name to this important quantity: we say that when a force 'moves along', it does work, and we say that the amount of work done by the force is equal to the product of the (magnitude of) the force and the distance its point of application moves in the direction of the force." This is a very direct way of introducing "work", but there is still a slight element of formality about it which might be found unsuitable in an elementary treatment. It is perhaps for this reason that in the Nuffield physics course, the measurement of energy changes is preceded by an extended qualitative discussion of different "forms" of energy.

By means of a number of simple qualitative experiments (some of them "thought experiments"), the pupil is made aware of the idea that in a "useful job" (one requiring fuel), energy is transferred from something to something else. For example, we need "fuel" to do a job like stretching a spring; but the stretched spring can itself do a job when returning to its normal length. So if we think of energy as "that which is stored in fuel and which can enable us to do a useful job", the action of stretching a spring seems to have transferred energy to a spring. With the aid of these rudimentary ideas about conservation of energy, it is possible to assign a more definite meaning to "work" for the pupil; work is a measure of the energy transferred in a "useful job". As the Nuffield Guide(II) puts it, "Work F.s. tells you how much you've taken out of one kind of energy

/store.....

store and put into another". In pursuing this point, the Guide draws an analogy between work as a measure of energy transfer and cheques or postal orders as a "measure" of the amount of money transferred from one person to another. This helps to make clear the important idea that work is not itself a form of energy, just as cheques or postal orders are not money in themselves.

The Nuffield Guide takes pains to stress an important point about this treatment of work and energy, viz. that it sidesteps the difficulties associated with the terms "work done on" and "work done by" and the accompanying use of plus and minus signs for the amount of work done. One simply has to take care to specify that the work is the transfer of energy from some particular form to some other particular form. Thus when one lifts up a suitcase, one transfers some of one's store of chemical energy ("breakfast energy" in the Nuffield phrase) to potential energy of the suitcase (strictly speaking, potential energy of the earth-suitcase system). By implication, work is always positive in this treatment. This does indeed seem a reasonable approach to the problem in the early stages of a treatment of work and energy; it is only later that the pupil need concern himself with the full implications of "$W = Fs\cos\theta$".

3.9.4 "Derived" units, including those of work

Following on the quantitative development of "work" outlined above, there is no reason why one should not proceed directly to the correct SI unit of work, viz. the newton-metre or joule. The introduction of this unit raises a question of some importance: how, if at all, is the teacher to justify the form which the unit of a "derived" quantity like work takes on? Should he take it as essentially self-evident that when a concept like work is

/compounded.....

compounded from other concepts, the unit for the derived concept should be compounded in a similar way from the units of the other concepts? The answer seems to be that this is virtually the only thing he can do short of venturing into the realms of dimensional analysis. The pupil who knows and has formed some understanding of a unit like "mile per hour" has already acquired some feeling for the treatment of units in this way, and it should be quite easy for him to extend this principle to the "multiplication" of the newton and the metre to form the newton-metre. Constant (1963) goes so far as to formulate a general rule for the compounding of units: "When physical quantities are multiplied or divided together, or raised to the n^{th} power, their corresponding units are respectively multiplied, divided, or raised to the n^{th} power, like algebraic quantities." The idea that units can be multiplied, divided or raised to the n^{th} power is likely to make the hair of the sophisticated dimensional analyst stand on end; in order not to shock him too deeply we should perhaps describe the compounding of the newton-metre from the newton and the metre by saying that "it is as though we multiply together the newton and the metre". Such a practice is unlikely to embed any serious misconceptions in the minds of pupils. It is important for them to regard this way of building up units as natural, and the less fuss that is made about it the better. We think it unnecessary ever to dictate to a class such a rule as Constant's, quoted above; by repeatedly meeting instances of the formation of units of derived quantities, pupils will become acquainted with the general principle without having to have it spelled out for them.

What is the underlying significance of such complex unit names as the newton-metre (or, more fully, the $kg\text{-}m^2/s^2$)? It is worth considering this question for a moment to see whether, in fact, the justification for

/compounding.....

compounding units in this way can be rendered in a manner which would be intelligible to the pupil at any stage of his school career. A recent discussion of the matter is that of Ellis (1966), who suggests that the significance of the complex unit names is that they indicate how "dependent scales" (i.e. scales derived from other scales) rest on "independent scales" (i.e. those not so derived, such as the commonly-used scales of length, time and mass). This may not seem to be a particularly profound statement, but Ellis reveals its significance more fully by discussing what happens to dependent scales if one decides to change the independent scales on which they are based by multiplication by conversion factors. For example, if the mass, length and time scales are changed to similar scales by the conversion factors p,q, and r respectively, then the scale of <u>force</u> derived from them must be changed by the conversion factor pqr^{-2}. The fundamental name of the unit of force (e.g. $m\text{-}kg/s^2$ in SI) thus indicates the form of this conversion factor. As Ellis points out, this information is important because relations between physical quantities are conventionally "expressed with respect to classes of similar scales". ("Similar scales" are scales for a given quantity which are related to one another by constant conversion factors.)

We doubt whether attempts to translate the above reasoning into terms understandable to the average pupil at whatever stage would be worth the effort spent on them. The principles under discussion here are of fundamental importance to the theory of measurement; but it is true in general that some understanding of these principles can be acquired more or less intuitively in the course of witnessing repeated applications of them. Early attempts to raise this understanding from the intuitive level could well produce confusion where none previously existed. In cases where the

/process.....

process of setting up a scale of measurement is not intuitively obvious (e.g. time or temperature), it is necessary to deal with this problem explicitly, since failure to do so would be likely to "embed serious misconceptions in the minds of pupils".

The renaming of the "newton-metre" as the joule illustrates another common practice in regard to physical units - the substitution of a single word, usually derived from a proper name, for a name compounded of several words. When such cases are encountered (e.g. "watt" for joule per second, "volt" for joule per coulomb, "farad" for coulomb per volt), it is important that the more fundamental form of the unit should be given first. Pupils should also be introduced to the standard ways of abbreviating units, as laid down by the Commission on Symbols, Units and Nomenclature of the International Union of Pure and Applied Physics (I.U.P.A.P. 1965), and now embodied in the "Basic Guide to the Metric System in South Africa" (South African Bureau of Standards 1968).

3.9.5 Towards conservation of energy

There are various possible extensions of the above account of energy that might be considered for use in an elementary treatment. In the Nuffield course some attention is given to machines, with stress being laid on the role of machines as "force multipliers" and on the fact that their energy output never exceeds the energy supplied to them. The reasons advanced by the Nuffield Guide I for the inclusion of machines are the following: "Children enjoy learning...descriptions and explanations of various machines; and they feel they are becoming knowledgeable engineers. Such teaching establishes links between physics and the ordinary world as well as engineering". This is true enough, and if there is sufficient room in the course, one

/would.....

would agree that pupils should enjoy the benefits of this kind of discussion. Of course it would not be necessary to spend time in explicit definition of such quantities as mechanical advantage, velocity ratio and efficiency, or of the relation between them; a few simple examples of machines could be discussed, such as the seesaw and a simple system of pulleys, their mechanical advantages (which need not be so named) could be experimentally determined and calculations carried out on the energy transferred to and by the machine. The latter point will provide further support for the idea of energy conservation.

Another extension of "energy" to be found as early as the end of Year One in the Nuffield course is an introductory treatment of radioactivity. The intention here is to show by means of cloud-chamber tracks that atoms of certain substances occasionally shoot out little pieces of themselves and that these "little pieces" (alpha-particles) clearly have energy associated with them. The inclusion of this introduction to "atomic energy" (as it is called in the Nuffield Guides) presupposes that pupils will previously have been given some grounds for believing in atoms, that they will have some idea of the sizes of atoms (or at any rate that they are much too small to be seen), and that the operation of the cloud chamber should not be entirely mysterious to them. Perhaps it is this last point which is most important. Any honest teaching of this topic will have to try to convince pupils that the tracks seen in the cloud chamber are made up of droplets that mark out the paths followed by the alpha-particles and that the alpha-particles are not themselves visible. This means that some explanation will have to be given of the formation of the alpha-particle track. If the introduction of radioactivity has been preceded by some account of elementary electrostatics, then it

/would.....

would be permissible to tell pupils something about the formation of ions in the chamber by the passage of alpha-particles; the use of smoke particles or ions to promote droplet formation should be demonstrated to pupils.

There is no doubt that radioactivity can be most usefully introduced if treated in the manner described above, both for its own sake and because of its revelation of sources of energy within the atom. (It is in fact reasonable to speak of this energy as "atomic energy" at this stage, since the pupil cannot yet be given firm evidence for the existence of the atomic nucleus.)

The treatment of energy so far outlined has stopped short of any direct, quantitative demonstration of the conservation of energy. We feel very strongly with the planners of Nuffield physics both that such a demonstration should be left to a later stage in the course, and that prior to that stage energy conservation should be merely hinted at and not simply asserted as an incontrovertible fact.

Compare what happens in the chapter "Work-work-work" of Book One of "Physics is Fun" by Jardine (1964). In the third paragraph of this chapter, the author says: "You cannot make energy or cause it to disappear; you can only change it into a different form. Scientists call this fact the conservation of energy and textbooks usually say, "Energy cannot be created or destroyed'". In a book one of the most attractive features of which is the constant questioning of the reader on points that are raised, it is astonishing to find the author suddenly flinging down a statement likely to provoke a big question from the reader, to which he gives no suggestion of an answer, viz. "How do you know that energy cannot be created or destroyed?" It is greatly to the credit of the Nuffield course that it treats the demonstration of conservation of energy as one of its climatic

/moments.....

moments. Surely it is right that so important and fundamental a principle should not be made the subject of mere hand-waving by the teacher and tacit acceptance by the pupil. In our later and fuller consideration of energy we shall outline arguments and demonstrations by which the meaning and truth of the conservation of energy can be brought home to pupils.

3.10 Pressure

3.10.1 The need for the concept of pressure in elementary physics courses

The concept of pressure, especially fluid pressure, is indispensable to any physics course. Even if we treat pressure in liquids as a side-issue, we need the concept of pressure to be able to talk sensibly about the behaviour and properties of gases; and we do need to talk about these, both because we find it advantageous to develop an elementary kinetic theory of gases in our teaching, and because we shall want to discuss constant-volume thermometers in more advanced teaching of the concept of temperature. As with other concepts we introduce, we want to explain the nature of the concept and discuss practical ways of measuring it.

From certain remarks in the Nuffield Guide (I), ("....this is a region where a difficult concept and unruly apparatus combine to make the essence of the topic too hard for pupils of this age"), we may infer that "pressure" is not easily taught to pupils of average age around twelve. To pupils two or three years older (say in Standard Seven or Eight in South African schools), the concept itself ought to present much less difficulty, and the difficulties encountered in the Nuffield course in trying to relate the concept to practical methods of its measurement should

/likewise.....

likewise be minimized. Such pupils ought to be able to follow the fairly simple chain of reasoning which leads from the definition of pressure through simple ideas about fluid pressure to the principle of the U-tube manometer. This is, in fact, the sequence of ideas recommended in the new syllabuses for junior secondary schools.

3.10.2 The nature and definition of pressure

Since a Standard Seven or Eight pupil can be taken to have gained some familiarity with force as a quantitative concept, it should not be too difficult to lead him on from this to a consideration of the effect of a force applied not at a point, but over an area. In particular he can be shown the effect of the same force applied to different areas of some suitable material. On this basis the concept of pressure can now be defined in the proper way. Some text-book authors (and, no doubt, the teachers who slavishly follow them) are surprisingly careless about so simple a matter as the correct fundamental definition of pressure. Consider the following examples from South African textbooks:

(a) "Pressure per unit area = $\frac{\text{Thrust}}{\text{area}}$ " (Gordon and Neser, no date);

(b) "Die krag wat op 'n eenheidoppervlakte uitgeoefen word, word die druk genoem. Druk is krag per eenheidoppervlakte". (The force exerted on unit area is called the pressure. Pressure is force per unit area.) (Prinsloo et al, 1965).

Comment on the first of these quotations is superfluous. The second quotation illustrates a much more common mistake, by no means confined to South African textbooks, viz. that of failure to specify that pressure is normal force per unit area. It is true that the context usually makes it clear that this is what is meant; but surely there is no harm in making what is implicit explicit in this case, especially

/as.....

as it requires the addition of only one word to a very simple definition. Anotner mistake to be avoided in tne early stages of the teacning of pressure is exemplified by this quotation: "Tne pressure exerted by a liquid is perpendicular to tne surface with which it is in contact" (Gordon and Neser, no date). In tnis statement, "pressure" is being used synonymously witn "force"; in fact, were "force" to be substituted for "pressure", tne statement would be unexceptionable. As it stands, it is misleading because it implies tnat pressure, like force, is a vector quantity. By raising this objection we do not mean to imply that tne pupil must be aware of the distinction between vector and scalar quantities at tnis stage. Rather do we suggest that the teacher should keep tnis distinction in mind so as not to sow tne seeds of future confusion in tne mind of the pupil.

Some time ago, a minor controversy on the subject of "Does Pressure Have Direction?" flared up in the correspondence columns of the American Journal of Physics (Summers 1946, 1949; Barton 1949; Stickley 1950; Dempster 1951). The final contribution to it by Dempster clears up the matter satisfactorily. He points out that in advanced treatments the pressure in a static fluid is in fact regarded as a scalar quantity, defined by the equation $\vec{F} = p\vec{A}$ (in the limit as A tends to zero), where p is the pressure and $\vec{A}$ is a vector whose magnitude is equal to that of the area on which tne force $\vec{F}$ acts and whose direction is that of the outward- or inward-drawn normal to the area. He proposes the following versions of the usual textbook statements concerning the pressure in a static fluid: "In the case of a fluid at rest, the force exerted by the fluid upon any surface is perpendicular to that surface; at any given point, the force exerted upon an element of surface is independent

/of.....

of the orientation of that surface." (The second part of this statement is intended to replace "the pressure in a fluid at rest is the same in all directions at any given point".) Even this is not fully correct; we should insert the phrase "the magnitude of" in front of "force" in the second part of Dempster's proposed revision.

There is no reason why the pupil should not be told the correct SI units for pressure at this point, viz. newtons/metre^2 (N/m^2). Although these will not be the units he will use in his initial studies of pressure, he should nevertheless know about them, even if only because they are the units implied by the definition of pressure.

3.10.3 The measurement of pressure

In arriving at a means of designing or calibrating a pressure gauge, the most urgent need is to show that the pressure at a given level in a liquid depends on the height of liquid above that level. A standard demonstration for this purpose is the one in which a tin can is punctured at various distances from its base and is then filled with water. The velocity with which the water emerges from a hole is taken as an indication of the pressure of the water at the level of the hole. Described in these terms, the demonstration seems a reasonable means of achieving its intended purpose. However, there is a snag; the horizontal range of the jet is usually taken as a naive measure of the velocity of emergence of the water, and this range will naturally depend on the height of the can above the surface on which the jet is allowed to impinge. If this surface is simply that of the trough in which the base of the can rests, the water emerging from a hole near the base of the can will have a very short range, comparable to that of the water emerging from a hole near the top of the can.

/In.....

In surprisingly many textbooks, these readily verifiable facts are ignored and diagrams are given showing the can resting in the trough or on a table and water from the lowest hole travelling the furthest. Teachers could use this as an illustration that textbooks do not always tell the truth! (The "Water Can Paradox" has been discussed by a number of authors, e.g. Paldy (1963) and Biser (1966). The latter gives an amusing survey of the treatment of the phenomenon in sixteen different textbooks and describes an experimental study of it.)

In an alternative "demonstration" of the dependence of fluid pressure on depth, which appears in at least two South African textbooks, a glass tube is attached to a manometer and lowered to various depths in a liquid in a tall gas jar. Manometer readings are taken for different depths of the tube mouth beneath the liquid surface, and the corresponding differences in level between the free liquid surface and the surface inside the tube are also measured. From these readings it is concluded that the pressure inside a liquid increases uniformly with depth. (In the form described, the experiment appears in Prinsloo et al (1965).) Gordon and Neser describe essentially the same experiment, reaching the conclusion that "the pressure is proportional to the depth below the free surface" (sic), and in the O.F.S. Teachers' Guide for Standards 6, 7 and 8 teachers are recommended to "Investigate [the] relation between pressure and depth, using a mercury manometer".) Here is a "Zanzibar" experiment (Rogers 1960) if there ever was one. In order to use the manometer in the way in which it is used in this experiment, it has to be assumed that the pressure in a liquid increases uniformly with depth beneath the liquid surface - which is exactly what the experiment sets out to prove. If the experiment were carried out with

/water.....

water in the gas jar and mercury in the manometer, then the most it can be said to establish is that if pressure increases uniformly with depth in mercury, it must do the same in water. But this is hardly worth the time spent in taking readings, tabulating them neatly and then drawing a graph! If there were some independent means of calibrating the manometer, this could save the experiment, but there seems to be none that could conveniently be used in school.

The demonstration with the punctured can is acceptable, provided that it is carried out correctly. But it cannot be used to find the quantitative relation between pressure and depth below the surface. Jardine (1964) suggests the use of a previously-calibrated Bourdon gauge to investigate this relationship; this raises the question of how the Bourdon gauge was calibrated. One could calibrate it by making use of the principle of the dead-weight tester (the author is indebted to Professor Gledhill for reminding him of this possibility). That is, one could attach the gauge to a vertically-mounted cylinder and piston and apply a known weight to the latter, so that known increments of pressure would be applied to the air inside the Bourdon gauge. It is true that the absolute pressure would not be known in this case, but this does not matter: the Bourdon gauge can be set to read excess pressure (over atmospheric), and it is the variation of this quantity with depth below the surface of the liquid in which the chief interest lies.

The alternative to this procedure is to derive the relation between pressure and depth theoretically for pupils and then to have them do some simple experiments which give it qualitative support. The proof amounts to showing that the weight of a column of liquid of uniform cross-section is proportional to the height of the column and should

/present.....

present no difficulty. It cannot at this stage be extended to produce the formula $P = P_o + \rho gh$, because the pupil knows nothing quantitative yet about acceleration; and it should not, in our opinion, be used to produce a variant of this formula in which pressure would be given by gravitational units of force per unit area. The advantage of the latter procedure would be that the density of the liquid would appear explicitly in the formula for liquid pressure. In spite of this, we think it best to keep gravitational units of force out of the way if we possibly can. The pupil should certainly be aware of the dependence on density, so that he would realize that it would be more sensible in some situations to use, say, a water instead of a mercury manometer to measure pressure difference; but most important for him at this stage is the knowledge that by means of a manometer (or a suitably calibrated Bourdon gauge) he can measure pressure in so many centimetres of mercury (or whatever liquid is in the manometer). He does not yet need absolute units for pressure, but he does need to have a clear understanding of the statements that "the difference in pressure between the two columns of the manometer is (x) cm of mercury" or "the pressure of the air in this container is 80 cm of mercury".

Here or earlier, pupils should do a simple experiment to show that, at a given level in the liquid, the pressure is independent of the direction of the force exerted by the liquid (or, as it is usually expressed, "the pressure is the same in all directions"). The simplest such experiment is an extension of the "tin can" experiment described above (cf. Nuffield Guide I). Again several holes are knocked into the can, but this time they are spaced horizontally round the tin. The experiment should be done with a can

/of.....

of undistorted shape and with one which has been knocked into irregular shape, so that the water would come out of the holes in directions other than horizontal.

The one application of the facts of fluid pressure that pupils should certainly know would be the mercury barometer. For pupils really to understand how it works, the explanation of its action should be as full as is consistent with their knowledge of the concept of pressure and of the laws of fluid pressure. If the barometer is tilted over, pupils should be helped to understand why the difference in levels is still the same; they do not yet know about components of forces, but they should be able to understand that part of the weight of the mercury column must now be supported by the walls of the tube. It would be unwise to burden pupils with the "hydrostatic paradox" or other subtle consequences of the principles of hydrostatics. Pascal's principle is the only other fact concerned with fluid pressure that they need know about, and it can be treated as a matter for simple demonstration with the aid of "Pascal's syringe" or similar apparatus. The application of this principle to hydraulic systems can be briefly discussed; if discussion of pressure follows that of energy, the "force amplification" of a hydraulic system can be discussed from the point of view of energy and compared with lever systems in this respect.

3.10.4 Fluid pressure from the molecular point of view

Demonstration of the existence of gaseous pressure can introduce a discussion of a model of a gas that would account both for the propensity of a gas to fill its container and its ability to exert a pressure. The kinetic model can then be brought in and its predictions discussed with the aid of a simple mechanical model for class demonstration. (Many teachers would prefer to take up this

/aspect.....

aspect in the context of a general study of the particle model of matter. Cf. Section 3.13.2 .).

Although it is not usual to discuss "molecular" models for the pressure exerted by liquids and solids, some pupil might well raise this point, here or later, and so we consider it briefly here (cf. Swartz 1966). It is of course true to say that the pressure exerted by a liquid results from its weight, but this alone cannot clarify matters on the molecular scale; the weight of a liquid is the force of gravity upon it, and we want to know how the liquid transmits this force to the surfaces in contact with it, or how one part of the liquid exerts this force on another. The answer lies in the forces between neighbouring molecules of the liquid. The mean separation of neighbouring molecules in a liquid is not much different from the "equilibrium" separation for such molecules (i.e. "r_o", the distance between the centres of adjacent molecules when the force between them is zero). When they are squeezed closer together than this, the intermolecular forces become repulsive. This is what happens when the liquid above a certain level presses down on the liquid below it, or when all the liquid in the container presses on the base. The molecules are pushed closer together, and so adjacent layers of liquid exert repulsive forces on each other; the liquid above pushes down on the liquid below, and the latter pushes back on the liquid above with an equal and opposite force. Any given layer within the liquid is in equilibrium under the action of the forces exerted on it, i.e. its own weight and the forces applied to it by the layers of liquid above and below it. A given molecule will be under restraint from repulsive forces from molecules above and below it. Molecules in the surface of the liquid are constrained by the weight of the atmosphere above them to take up equilibrium

/positions.....

positions at vertical mean distances less than r_o from the molecules below them. Again it is the short range repulsive forces between molecules which are the means whereby the atmosphere exerts this force upon the liquid surface. In this case these forces are exerted during collisions between the molecules of the atmosphere and those of the liquid.

As a model to aid this explanation of the transmission of pressure through a liquid, one can use the idea (or an actual mechanical model) of a lattice of balls joined together by springs. Application of a force to the top of such a lattice will result in the balls being pushed closer together; if the weight of the balls is appreciable, those near the bottom of the lattice will be markedly closer to one another than those near the top. In view of the complicated nature of the motions of molecules of a liquid, the above description must be regarded as a considerable oversimplification (cf. Section 5.4.18); yet it is essentially correct.

3.11 Temperature

3.11.1 Introduction

In the Teachers' Guide for Physics for the Joint Matriculation Board (Helm 1963), the present writer wrote "... We believe that the treatment of temperature to be found in many textbooks in common use is open to serious criticism on grounds of logical rigour... It is... because so many of the conventional treatments are so formulated that they cannot fail to confuse the pupil almost from the start that we suggest that a fresh beginning must be made." In writing these criticisms, the author had in mind the kind of treatment of temperature in which misleading statements

/were.....

were made about the relative "advantages" and "disadvantages" of mercury-in-glass and alcohol-in-glass thermometers, and fundamental confusions were evident about the ideal gas scale of temperature. In the years since the above remarks were written, there has been no general improvement in textbook treatments of temperature. While a number of first-year university textbooks and a few school textbooks offer unexceptionable treatments of this topic, and there are signs that the forthright criticisms of Rogers (1960) and Warren (1961, 1965) are having their effect, some new school textbooks betray the same kinds of confusion as the old ones did and others introduce new kinds of confusion of their own devising. In some cases the textbook authors have been following in the footsteps of the planners of syllabuses who have not thought out their methods of attack on temperature with the requisite care. Thus in our view there is still as much room for criticism of the teaching of temperature as ever, and as much need for the outlines of a correct approach to be carefully set out.

3.11.2 A critique of the introduction of temperature in terms of the kinetic theory

Perhaps the most widespread of the "new kinds of confusion" is the tendency to initiate a treatment of temperature with the assertion that the temperature of a substance depends on the average kinetic energy (or simply on the speed) of its molecules. As far as it goes, this statement is true enough. The objections to its use in introducing temperature are the following:-

(i) It is simply not possible to produce a reasonable argument in favour of the existence of a relation between temperature and average molecular kinetic energy without bringing in a string of assumptions about the nature of heat and matter. This is exactly what Jardine (1964) and Stollberg

/and.....

and Hill (1965) do; they begin with the assumption that heat is a form of energy and then proceed swiftly to the assumption that matter is made of molecules in ceaseless motion, on which basis they can then argue that temperature is related to molecular speed or average molecular kinetic energy. They both set forth this argument before embarking on any discussion of the measurement of temperature.

(ii) If one is going to define temperature, implicitly or explicitly, in terms of average molecular kinetic energy, then one ought to have a precise numerical relationship between temperature and average kinetic energy in mind. It is only for an ideal gas that it is possible to state such a relationship in a convenient form, i.e. one which will yield up temperature as it is customarily defined. And in any case it seems more than a little unrealistic in an elementary course to choose as a definition of temperature a relationship which can only be given operational meaning in rather special circumstances, e.g. in measuring the temperature of ions and electrons in the ionosphere or exosphere, where one measures the average kinetic energy of the particles directly, or in measuring the temperature of hot plasmas, where the temperature is inferred from measurements on the Doppler broadening of spectral lines.

(iii) Behind the present-day tendency in both syllabuses and textbooks to stress the relation between temperature and molecular kinetic energy is, one suspects, a desire to make the concept of temperature more meaningful. How nice it would be to be able to say that temperature is "really" the average kinetic energy of the molecules. But is it? We have indicated above that only for an ideal gas is the (absolute) temperature directly proportional to the average kinetic energy of the molecules; and what is one to say about the temperature of radiation - a most useful concept?

/Even.....

Even if the relation, absolute temperature ∝ average molecular kinetic energy were to have universal validity, we would not be justified in saying that "real temperature is the average kinetic energy of a molecule". As Rogers (1960) puts it, "we have not discovered [this relation] as a description of 'true' temperature; we manufactured it when we settled on the gas scale for convenience."

(iv) As is well known, it is possible to give a satisfactory "definition" of temperature having a wide range of applicability, with reference to macroscopic concepts only. (By "definition" in this case we mean a combination of a denotative definition of temperature in terms of thermal equilibrium and the definition of the Kelvin scale in terms of the properties of the Carnot cycle.) In view of this, it seems a dubious proceeding in an elementary course to define temperature in terms of microscopic concepts, when this definition depends on a particular theory of the nature of microscopic matter.

(v) Fashionable methods of "defining" temperature in terms of the kinetic theory in elementary courses are considerably oversimplified versions of the "real thing". To define temperature rigorously within the context of the kinetic theory of gases requires the following sequence of steps after the establishment of the initial assumptions of the theory (cf. Feynman et al.1963):

(a) Introduce the concept of thermal equilibrium.

(b) Define temperature denotatively in terms of thermal equilibrium.

(c) Show, using sophisticated kinetic theory, that when two ideal gases are in thermal equilibrium, the mean translational kinetic energies of their molecules must be the same.

(d) Define a temperature scale constitutively in terms of the mean kinetic energy of the molecules of an ideal gas.

The presence of step (c) in this sequence places this

/rigorous.....

rigorous treatment well beyond the reach of any school syllabus in physics. One wonders whether those who so eagerly introduce the concept of temperature in elementary treatments by way of the kinetic theory have any real idea of the extent of the oversimplification embodied in their approach. If they had, perhaps they would not be so ready to lay so insecure a foundation for the introduction of an important fundamental concept.

On account of these objections, we regard it as unwise, especially in an introductory treatment, to "define" or "explain" temperature in terms of the average (translational) kinetic energy of molecules. We feel that it is attempting too much too soon in the way of exploitation of the kinetic model of matter, to confront a pupil in the junior secondary school with the assertion that "temperature is a measure of the average kinetic energy of the molecules of a gas "before he has had any opportunity of making a proper quantitative study of gaseous behaviour, or even of making a preliminary study of thermal equilibrium or the manner in which a temperature scale is set up. By all means let pupils become acquainted with this "interpretation" of temperature, but let them do so at a moment which is truly opportune. It may be felt that this makes the treatment of temperature less exciting and less "modern" at school than it otherwise might have been. If this is indeed so, then perhaps after all it is still better to stick to the more correct and less exciting treatment, for this is less likely to leave troublesome misconceptions behind in the minds of the pupils.

3.11.3 A suggested introduction to temperature - Stage One

At the root of any correct presentation of temperature is the concept of thermal equilibrium. However, it appears to be the standard and approved practice to keep

/well.....

well clear of this concept in introductory treatments of temperature, if the textbooks are anything to go by. It is difficult to see why this should be so; there is nothing frighteningly difficult about the idea of thermal equilibrium and the way in which it is associated with temperature, as we hope to show. In the treatment of temperature which we now proceed to outline, thermal equilibrium will be accorded its proper place.

By way of introducing a discussion of temperature in an "exciting" way, one can begin by referring briefly to the pupil's earlier experience of temperature as something measured with a thermometer and related to "hotness" or "coldness", and then raise the question of what could possibly be meant by statements in the press about temperatures of millions of degrees in the sun, the stars and in the explosions of "atomic" and "hydrogen" bombs. When the physicist talks about such fantastically high temperatures, one could say, he must be discussing something that is well beyond the range of human sensation. What, then, can the physicist mean by "temperature"?

To answer this question, one can start by referring back to the humble mercury-in-glass thermometer and briefly describe its mode of calibration. Then one can discuss in detail the way in which we use it to measure temperature. Suppose it is used to measure the temperature of a beaker of warm water. To do this, we place the bulb of the thermometer in the beaker and stir the water a little; the mercury thread rises at first and then attains a steady reading. Before the thermometer was placed in the beaker, the mercury inside was in mechanical and chemical equilibrium; the mechanical forces on it were balanced, and it was not undergoing any chemical changes. The same was true of the water in the beaker. But after the thermometer was placed in the

/water.....

water, the length of the mercury thread changed. Had we sufficiently sensitive instruments, we would also have been able to detect a tiny change in the volume of the water in the beaker. Therefore, when we placed the thermometer in contact with the water in the beaker, they could not initially have been in equilibrium with each other, because each underwent changes. After a while, no further changes took place, and equilibrium was attained. This is evidently a special kind of equilibrium, different from mechanical and chemical equilibrium; we call it thermal equilibrium, and say that the thermometer and the water have attained thermal equilibrium.

(Of course, this is hardly a rigorous treatment of thermal equilibrium; to provide that we should have had to invoke the concept of a diathermic wall, and arrive at the statement that "when two systems are separated by a diathermic wall, the joint state of both systems that exists when all changes in the coordinates have ceased is called thermal equilibrium" (Sears and Zemansky 1964, Zemansky 1964). But perhaps it will be allowed that the failure to mention diathermic walls at this level of treatment is not a dangerous over-simplification, and that the failure to give a rigorous definition of thermal equilibrium is equally inoffensive.)

Pupils can next be invited to consider whether if two bodies are each in thermal equilibrium with a third, they are necessarily in thermal equilibrium with each other. Of course the answer is that they are, but it is important to make the point here that this answer can only be decisively given by experiment; we cannot take it as being obvious. (Cf. Zemansky (1964): "Woman A loves man C. Woman B loves man C. But does woman A love woman B?") There is no need at this stage to dignify this result with its official title

/"The.....

"The Zeroth Law of Thermodynamics". What is needed is to pursue the argument further so as to arrive at a definition of temperature. From the "Zeroth Law", it is possible for us to imagine a collection of bodies or systems so arranged that when any two of them are placed in contact, they will immediately be in thermal equilibrium. Such a collection or bodies must have some property in common. What is this property? Or, putting the question in slightly different form, what property of a body or system determines whether or not it will be in thermal equilibrium with another body or system when placed in contact with it? Evidently it is not its mass, its volume, its density or any of its physical properties we have studied so far. It is, in fact, a new property which we call its temperature. We can now define temperature as follows:- The temperature of a system is that property which determines whether or not it will be in thermal equilibrium with other systems. Accordingly, when two or more systems are in thermal equilibrium, they are said to have the same temperature.

3.11.4 The status of the "Zeroth Law"

The above discussion of thermal equilibrium and the "Zeroth Law" has been based on the orthodox view of this aspect of the subject. This view is essentially the same as that adopted by Maxwell in his "Theory of Heat" (cf. the tenth edition, 1891). Maxwell, after having given denotative definitions of heat and temperature, states a "Law of Equal Temperatures": "Bodies whose temperatures are equal to that of the same body have themselves equal temperatures". He points out that "this law is not a truism", draws attention to its empirical basis, and stresses that it is "the foundation of the whole science of thermometry". Nowadays, it is generally regarded as preferable to state the "Zeroth Law" in terms of thermal equilibrium rather than of tempera-

/ture.....

ture. (An exception is provided by Thomsen (1962), who thinks that the scalar nature of temperature ought to be clearly implied by the statement of the law, and accordingly states it as follows: "There exists a scalar quantity called temperature, which is a property of all thermodynamic systems (in equilibrium states), such that temperature equality is a necessary and sufficient condition for thermal equilibrium".) But the underlying attitude of such modern expositions of the subject as those of Zemansky (1957), Pippard (1957), Hatsopoulos and Keenan (1965) and Buchdahl (1966) remains the same as Maxwell's.

This "orthodox view" has recently been challenged by Redlich (1968), who asserts that "The 'zeroth law' is not the generalization of observations, it is not a necessary condition for the definition of temperature, and it is no law". (I am grateful to Professor J.A. Gledhill for calling my attention to this paper.) Redlich illustrates his argument with two examples. The first concerns an object A that is permeable for neutrons, an object B that absorbs neutrons, and an object C that radiates neutrons. "Not knowing anything of these radiation properties", he says, "we establish thermal equilibrium between A and B and between A and C. Then we find that B warms up on contact with C." In his other example, A,B and C are represented by water, benzene and carbon tetrachloride respectively: "the thermal equilibria AB and AC are easily established, but B and C produce a heat of mixing on contact." Redlich concludes that, far from being a law, the "so-called zeroth law is simply an example of a 'rule of order' imposed on us by any description of natural events", which in the examples cited and in similar situations leads us to look for an initially unsuspected "non-thermal mode of interaction" which is the cause of the failure to establish thermal

/equilibrium.

equilibrium.

Initially, this reasoning appears persuasive, but it seems to us to be open to certain objections. First, let us examine Redlich's examples more closely. It is immediately apparent that they violate a fundamental condition that is always, following Caratheodory (1909), laid down in the rigorous exposition of classical thermodynamics, viz. that there should be a "diathermic wall" between the systems which are supposed to interact thermally with each other. Such a wall would presumably be constructed so as to prevent the passage of matter from one system into the other, so that the situation depicted in Redlich's first example could not arise. Thermal contact by mixing, envisaged in the second example, is also ruled out by the requirement that a "diathermic wall" must be present. Redlich has anticipated this objection and seeks to counter it by referring disparagingly to the introduction of "conditions of isolation...that come as an afterthought. As introduced at the start, they are entirely indefinite. Restrictions imposed afterwards in particular cases have sapped the conceptual strength of thermodynamics". (He cites no evidence in support of the last statement.) But it is surely normal practice in formalizing the structure of a theory to set down with some care the conditions under which its fundamental laws may be claimed to hold. It may well be that some of these conditions may be "afterthoughts"; first attempts at setting up a theory are generally less careful than later ones.

Another objection that can be raised against Redlich's first example is that one would not expect a pair of systems to attain thermal equilibrium if there was a constant transfer of matter from one to the other. Redlich would presumably retort that in terms of the conditions he lays down,

/one.....

one is not supposed to know beforehand that there is such a transfer of matter, so that it would be the experimental observation that the Zeroth Law is apparently violated that would make one suspect that something odd is going on. As we have seen, this is his reason for rejecting the view that the Zeroth Law is a "law"; as soon as one encounters a situation in which it appears to be violated, one looks for and (Redlich implies) invariably finds a means of "rescuing" it; hence, according to Redlich, Popper's (1959, 1963) criterion of refutability for a physical law or theory cannot be met, because "the zeroth law can never be found to be invalid". Let us consider this idea in conjunction with Redlich's bald assertion that the Zeroth Law is not a generalization from experience. The latter assertion seems to be contradicted by the facts. It <u>is</u> a fact of experience that if a thermometer registers the same reading when placed successively in contact with a number of everyday objects, the latter objects will be in thermal equilibrium when placed in "contact" with one another. (We could always, if we wished, arrange for this experiment to be carried out subject to certain precautions which approximate to the use of adiabatic enclosures and diathermic walls required by the strict theory.) On the basis of this kind of experimental result, one acquires confidence in the general validity of the law under appropriate conditions, and finally gives it the status of a fundamental law. One does not cast doubt on that status when, in dealing with the kind of situation discussed by Redlich, one suggests that the apparent violation of the law is probably to be attributed to a failure to satisfy the conditions under which the law can be expected to hold. It is neither unusual nor unreasonable for a physical law to be used in this way. When beta-decay of nuclei was first studied experimentally, the phenomenon appeared to violate

/the.....

the Laws of Conservation of Energy and Angular Momentum. Some physicists, reluctant to believe that such a violation had really taken place, suggested that the missing energy and angular momentum were carried away by an undetected particle, the neutrino. Their stand in the matter was vindicated many years later by the detection of the neutrino by Reines and Cowan (1956). To the best of our knowledge, the insistence of physicists on adhering to these conservation laws was not viewed by them or others as a degradation of the status of these laws to "rules of order". In our view, the attitude one adopts towards the Zeroth Law in the analysis of Redlich's examples is essentially the same as that adopted by Pauli, Fermi and many others towards conservation of energy and momentum in respect of beta-decay. It is not that these laws "can never be found to be invalid", but that previous experience suggests that their range is very wide, to say the least. Therefore in situations in which they appear to be violated, it is natural first to try to trace the supposed violation back to some hitherto unsuspected factor.

Redlich's remark that the Zeroth Law "is not a necessary condition for the definition of temperature" is obviously true. There is nothing to prevent one from introducing and defining temperature before introducing the Zeroth Law if one so wishes. The concept of thermal equilibrium can be introduced and temperature defined immediately afterwards with respect to the thermal equilibrium of a single pair of systems. This can be done whether temperature is treated as a fundamental concept or whether it is treated as "derived", as it is in a formulation of thermodynamics in which entropy is treated as fundamental and temperature is derived from it. But the concept of temperature can only be regarded as useful within the context of a theory in which the

"Zeroth.....

Zeroth Law can be assumed to have wide validity. Hence a treatment of temperature in which the Zeroth Law is stated before temperature is introduced cannot seriously be said to be in error.

We conclude that Redlich does not appear to have substantiated his claim that the Zeroth Law is "no law". Certainly there is nothing in what he has to say which may be taken as in any way invalidating what we have called the "orthodox" view of temperature and the Zeroth Law, and elementary teachers need not feel in any way uneasy about the correctness of a treatment of temperature based on this view.

The thoughtful teacher may well wonder whether the Zeroth Law cannot be derived from the other laws of thermodynamics. It turns out that it can - provided that a suitable additional assumption is made. Turner (1961, 1962) has produced a proof of the Zeroth Law based on the First and Second Laws of Thermodynamics and on a "general assumption concerning thermal equilibrium" to the effect that "thermal equilibrium will result if any number of bodies are allowed to interact by any arbitrary set of thermal contacts." Hatsopoulous and Keenan (1965) prove the Zeroth Law from the Second Law and (implicitly) the First Law, and also the Gibbs principle of generalized inertia, a simplified statement of which is, in these authors' words, "that a sudden discontinuous alteration from a finite rate of change to a zero rate of change cannot occur without finite alteration in the 'circumstances'". They remark that "all attempts so far to prove the zeroth law from the second law alone have failed", and point out that the law applies only to systems whose entropies have continuous derivatives with respect to their energies and that it may fail for other systems. These proofs of the Zeroth Law are sophisticated and there is no possibility of their being used in an elementary treatment

/of.....

of the subject. For such a purpose it is sensible to treat the Zeroth Law as an independent, fundamental law.

3.11.5 A suggested introduction to temperature - Stages Two and Three The foregoing constitutes Stage One of the proposed introductory treatment of temperature. It has, as we see it, the following advantages:-

(i) It places proper emphasis on the role of thermal equilibrium;

(ii) it arrives at a definition of temperature which is entirely divorced from any anthropomorphic associations (after all, the pupil has been implicitly using the definition "temperature is hotness measured on some definite scale" for some time previously; he should now be ripe for something better); it thus helps to point the way towards a satisfactory answer to the question posed at the beginning, viz. "what does the physicist mean by temperature?"

In Stage Two, the problem of the measurement of temperature can be taken up. Here one can begin by referring again to the mercury-in-glass thermometer in the beaker of water. In the process of attainment of thermal equilibrium, the length of the mercury thread increases at first, then becomes steady at a new value; its becoming steady shows that thermal equilibrium has been attained. If the thermometer were placed in another beaker of water, the length of the mercury thread would have a different value when thermal equilibrium was reached. We conclude from the definition of temperature that the two beakers of water must be at different temperatures; and we see that the length of the mercury thread can be taken as an indicator of those temperatures - or rather, it is an indication of the temperature of the thermometer itself, which is the same as that of either beaker of water when thermal equilibrium has been attained. This introduces the idea of a thermometric

/property.....

property whose value is to be used to specify temperature. Then the remaining steps required for the establishment of a temperature scale should be outlined, viz. the definition of changes in temperature as being proportional to the changes in the value of the thermometric property, and the choice of a fixed point or fixed points to which arbitrary values for the temperature would be assigned. This last point raises a practical matter of some importance to the teacher. In the comments accompanying the New Common Basic Physical Science Syllabus (1965), he is advised as follows: "The Celsius gas scale is defined in terms of the triple point and this fact must be accepted. It is suggested that the Celsius scale be introduced by t °C = K - 273.15 . It should then be explained that standardising laboratories have measured the ice and steam points on this scale and found them to be 0.00°C and 100.00°C respectively."
We would suggest instead that the introduction of the Celsius scale by this means is neither desirable nor necessary at junior secondary school level. (We leave aside the question of what is meant by the Celsius gas scale.)
At this stage, the pupil could be given no satisfactory account of where the magic number 273.16K as the temperature of the triple point comes from. Also, he could have no idea yet of how a temperature like, say, 5K could possibly be measured. In fact, he is not yet ripe for a consideration of whether such temperatures have operational meaning and it would seem pedagogically unwise to thrust this upon him prematurely. Everyone knows that the Celsius scale temperature is now officially defined by t °C = K -273.15 ; but everyone also knows that for all practical purposes in the school laboratory (and indeed in many other laboratories) the Celsius scale is identical with a temperature scale

/established.....

established by using the ice and steam points as fixed points and dividing the temperature interval between them into 100 degrees. The teacher would therefore scarcely be practising wanton deceit if he introduced the idea of the Celsius scale with the aid of the ice and steam points; and he could always salve his conscience by waving a "warning flag" in the shape of a statement that this is not how the Celsius scale is officially defined nowadays, and that pupils will have the correct definition explained to them in due course. The transition from two fixed points to one fixed point is not likely to cause serious distress when the occasion arises. The most important points for pupils to appreciate about a fixed point are the necessity for it and the arbitrariness of the choice of value to be assigned to its temperature.

In Stage Three of this introductory treatment of temperature, the point should be made that one could construct a variety of different Celsius thermometers based on different thermometric properties, according to the principles previously outlined. Following on the heels of this, it should be stressed that it must not be expected that these different thermometers would agree with one another except at the fixed points. No treatment of temperature can be considered adequate if it does not make this extremely important point. The question should be put to pupils: "Which thermometer is right?", and it is essential to convince them that the correct answer to this question is "Whichever one we choose as standard". (In its otherwise rather brief consideration of the introduction of temperature, the Nuffield Guide (II) does bring this point out strongly: "Ask whether one can be sure these different types of thermometer will agree with one another. Warn pupils that we cannot expect that, but must choose one standard type.")

/3.11.6.....

3.11.6 The "advantages" of mercury - in - glass thermometric scales

In traditional treatments of temperature, the above kind of discussion of thermometric scales did not appear. Instead, consideration was given to the comparison of the advantages and disadvantages of mercury-in-glass and alcohol-in-glass thermometers. This type of comparison was open to serious criticism on a number of counts. Thus among other things it was said that "Since mercury freezes at -39°C, it is not good for extremely cold weather, while alcohol does not freeze until the temperature is -117°C. On the other hand, alcohol boils at 78°C. Thus for determining the temperature of steam, you would use mercury, which has a boiling point of 357°C." (The actual quotation comes from Stollberg and Hill (1965), but similar statements can be found in a great many school textbooks, including many published in South Africa.) These statements are true, of course; but one must question the wisdom of making them at this stage. To a pupil to whom the need for the choice of a standard thermometer has not been explained or even hinted at, what exactly does it mean that "alcohol freezes at -117°C"? If he is lucky, he will have been shown how to calibrate a mercury thermometer. He has been told that mercury freezes at -39°C and that alcohol boils at 78°C. How, then, is he to regard the statement concerning the freezing point of alcohol as verifiable? One cannot set up an "alcohol-in-glass" temperature scale by the method of two fixed points, so that there is no alcohol-in-glass scale to extrapolate into the region -39° to -117°C. Nor can the alcohol-in-glass thermometer be calibrated against the mercury thermometer in this region, for obvious reasons. Since the pupil knows nothing of the choice of standard scales

/of.....

of temperature or their range of applicability the bald statement that "mercury freezes at -117°C" is at best meaningless for him.

An "advantage" of mercury usually referred to in this connection is that "mercury has very uniform expansion, giving equal lengths of scale for equal temperature intervals", in contrast with alcohol which "does not expand quite as uniformly as mercury" (Tyler 1951). (Compare also the following, in a textbook published in South Africa as recently as 1965: "Kwik sit reelmatig uit by verhitting en krimp reelmatig in by afkoeling" (Mercury expands uniformly with heating and contracts uniformly with cooling) (Prinsloo, Hattingh and Koen 1965).). Of course, the teacher knows very well what these statements mean. The "temperature intervals" referred to in the quotation from Tyler are temperature intervals of the standard ideal gas or thermodynamic scale of temperature, and it is in terms of this scale that the expansion of mercury turns out to be uniform and that of alcohol not. But what can the pupil understand at this stage by the assertion that "mercury has very uniform expansion"? The only temperature scale he knows is based on the expansion of mercury, so that for him "mercury has uniform expansion" is either meaningless (cf. Rogers's forthright and amusing discussion of this point (1960)), or it implies that temperature, besides being a reading on a thermometer, as he has thus far been led to believe, is also a mysterious, ill-defined entity which can take on quantitative meaning without reference to a thermometer or to any theory with which he is acquainted as yet. Thus can arise the kind of confusion about the nature of temperature which can take years to eradicate.

It is true that confusion of this kind was prevalent even among eminent scientists until well into the nineteenth

/century.

century. As Mach remarks, "Until contemporary times... workers in this field seem more or less unconsciously to have sought for a natural measure of temperature, a real temperature, a kind of Platonic Idea of temperature, of which the temperature read from thermometers is only an incomplete and imperfect expression" (Mach 1896, quoted in Ellis 1966). He gives a number of quotations to illustrate this point, e.g. Lambert: "One hesitates as to whether the actual degrees of heat are really proportional to the degrees of expansion", and Dutton: "Mercury...seems to approach more nearly to a uniform expansion". But present-day teachers are no more bound to perpetuate these erroneous ideas in their treatment of temperature than they are to teach the caloric theory of heat without correction. We cannot deny that the pupil is to some extent conditioned towards a "Platonic Idea of temperature" by virtue of his experience since early in his schooldays of reading something called "the temperature" on thermometers. But again there is no good reason why a teacher should tacitly accept this state of affairs instead of seeking to improve on it.

In view of the confusion which can arise, it is best to omit all reference to mercury's uniformity of expansion and alcohol's lack of it in any introductory treatment of temperature. Later it would be sensible to point out that there is , in fact, good agreement between the thermodynamic and the mercury-in-glass scales over the range of validity of the latter, and that this agreement is an entirely fortuitous consequence of the properties of mercury.

There is really no good reason why one should spend time in class in discussing the advantages and disadvantages of mercury and alcohol as substances for use in thermometers. Omission of such a discussion will mean the loss of lists

/of.....

of points which could conveniently be memorized for examination purposes, but in these days when even examiners seem to be convinced of the need for examination reform, this loss is not too serious.

3.11.7 Conclusion

There would be no need to take this treatment of temperature any further for the moment. It has established the fundamental concept of thermal equilibrium and its association with temperature, and it has established the fundamental principles of the measurement of temperature. It has introduced the mercury-in-glass thermometer, the scale of which can serve as a perfectly satisfactory temporary standard for temperature measurement for any work to be done in the laboratory at this stage. It has not sought to build any link between temperature and the molecular model of matter; we believe we have justified this adequately in our introductory remarks in this section. Some might complain that if temperature is to be dealt with in this way, one has to begin it as a completely fresh topic; at least on the "temperature-as-average-kinetic-energy-of-molecules" approach the introduction of temperature follows smoothly on what has gone before. We do not believe this to be a serious objection; one might like to think of physics as a "unified" subject, but there are situations in which it might well be disadvantageous to try to stress this unity too early. But even if this is conceded to be a genuine drawback, we believe that the advantages of the above treatment still recommend it above some of the more popular ones.

3.12 Other topics connected with "heat"

3.12.1 Thermal expansion of solids and liquids

It has long been traditional practice to include quantitative treatment of this topic in school syllabuses in

/pure.....

pure physics. It was not included in the new South African Common Basic Syllabuses in Physics and Physical Science, but some examining bodies have seen fit to add it to their own syllabuses for these subjects. One wonders why this has been done. The quantitative treatment of the thermal expansion of solids and liquids can lead nowhere in a school science course. There seems no point in determining the coefficients of linear and cubical expansion of various substances, except to make the far from profound points that (a) they are quite small in the case of liquids and very small in the case of solids, and
(b) they are different for different materials. The second point at least can be established by means of a simple qualitative experiment which can be done in the primary school; indeed such experiments are already included in the new syllabuses for general science in the primary school. As for the first point, the data obtained from measurements of coefficients of thermal expansion cannot conveniently be utilized in the discussion of any theory of thermal expansion, since a quantitative theory of thermal expansion goes far beyond the reach of any school syllabus.

There seems to be an irresistible urge compelling both teachers and planners of syllabuses to try to explain the thermal expansion of solids and liquids in terms of the kinetic theory of matter. On the face of it, the explanation is very simple. As the temperature goes up, the "molecules" acquire more energy; they vibrate more vigorously and so need more room in which to move. (In this and similar contexts, the word "molecule" must be understood to denote molecules, atoms or ions, whichever may be appropriate. Brommert (1965) has suggested the use of the term "amion" as a collective name to embrace the terms atom, molecule and ion.)

/While.....

While this explanation is true in its way, it is misleading in that it implies that even if the intermolecular potential energy curve were completely symmetrical about the equilibrium molecular separation, thermal expansion would take place. The truth is otherwise. Because the intermolecular potential energy curve for neighbouring "molecules" in a lattice is, in general, asymmetric about the equilibrium separation, increasing vibrational energy gives rise to an increasing average separation between molecules (cf. Warren 1965, Halliday and Resnick 1966). Thus thermal expansion is, microscopically speaking, a far more complex phenomenon than might at first be supposed. The moral of the story is that teachers should restrain their urge to "explain" thermal expansion, and planners of school syllabuses should cease to require that teachers should explain it. In fact, we would strongly suggest that thermal expansion of solids and liquids, except in so far as it may contribute usefully to the teaching of temperature, should be left out of secondary school science syllabuses altogether. The qualitative facts of thermal expansion and their practical consequences can be taught in the primary school, when the annoying difficulty of explaining the phenomenon cannot arise. If pupils in the secondary school ask for an explanation of it, the answer can be given in much the usual terms, except that the teacher should wave his warning flag concerning the inadequacy of his explanation very vigorously indeed. Perhaps when he gets round later to talking about the shape of the intermolecular potential energy curve, he can tell his brighter pupils something more sophisticated about thermal expansion. But this is emphatically not the kind of topic on which examination questions ought to be set.

3.12.2 Heat and calorimetry

Here we are once again on traditional ground - we

/almost.....

almost said hallowed ground. For traditional calorimetry, despite the sneers to which it has been subjected by the spokesmen of new physics courses of schools, still lives on in many places. For instance, the new syllabuses for general science in junior secondary schools in South Africa still require determinations by the method of mixtures of the specific heats of solids and liquids and of latent heats of fusion and evaporation. Can we not dispense with this sort of thing? The celebrated "Nuffield" jibe (cf. Lewis 1965) about the amount of water transferred just compensating for the amount of heat lost to the surroundings points up the ridiculousness of much that is done in the name of traditional calorimetry; e.g. the solemn absurdity of weighing to hundredths or even thousandths of a gram and then measuring temperature differences to two or at best three significant figures, and making only the most rudimentary attempts at correction for heat losses. The progressive teacher is understandably impatient with those who would seek to preserve their allegiance to these experiments, and cannot really be blamed for wanting to do away with this kind of calorimetry altogether. For his sake and that of elementary teaching generally, we should study the question of what purpose is served by experiments on calorimetry and see if any short cuts are possible.

From the point of view of the elementary physics teacher, the most important purpose to be served by an introduction to calorimetry is the provision of a unit of "heat", which is a possible preliminary to a quantitative demonstration that "heat" is a form of energy. If it is agreed that such a demonstration is an essential part of an introductory physics course - and we shall proceed here on this assumption - then we can hardly do without the calorie or the kilocalorie, at least to start with. From the point of view of chemistry, it is

/important.....

important to know something about calorimetry in view of the amount of attention being paid to the measurement of "heat" in present-day teaching of chemistry. So some calorimetry is necessary; it remains to be decided how much and in what form.

There are various ways of introducing the idea of a unit of heat. A safe, though somewhat longwinded approach is to carry out a series of short experiments designed to show that the amount of "heat" (whatever "heat" may be) liberated or absorbed by a body or an amount of liquid when it changes temperature depends on its mass, the magnitude of its change of temperature, and the nature of its material. (This is the approach recommended for junior secondary schools in the Orange Free State.) Thus one can heat up a pair of similar beakers containing different amounts of water for the same length of time and with similar burners, inter-changing the burners half-way through (cf. Jardine 1964), and show that the smaller amount of water experiences the bigger rise in temperature. Different masses of metal heated to the same high temperature can be transferred to similar amounts of water in similar beakers, producing different temperature rises. Similar masses of different metals can be transferred to similar amounts of water, producing different temperature rises; and finally similar masses of the same metal can be heated to different temperatures and then transferred to similar amounts of water, producing different temperature rises.

It is questionable whether this rather lengthy build-up, which can at best only provide evidence that heat liberated or absorbed depends on the product (mass x temperature) for a given material, is worth the time and trouble spent on it. In the Nuffield course this preliminary experimentation is reduced to one simple experiment, in which pupils

/heat.....

heat 1 kilogram of water electrically in an aluminium saucepan and measure the temperature rise. (The electric heater is, as yet, a "black box" in this context, but this does not matter here; a Bunsen flame would be at least as much a "black box".) By way of justifying the method of measurement of heat, the teacher is simply asked to "Explain that for over 150 years scientists have been measuring heat given to water by multiplying the mass of water by the temperature rise", and to suggest to his pupils that it is reasonable that the amount of heat delivered to the water should depend on how much water there is as well as on the temperature rise. This is surely adequate. Tastes may differ about the "appeal to authority" apparently implied by the reference to what scientists have been measuring for over 150 years - some might prefer to stress that the choice of "mass x temperature rise" as a measure of heat delivered to water is in the nature of a plausible guess to be justified or otherwise by the consistency of measurements based upon it - but few could reasonably contest that this kind of experiment is a neat and effective way of starting off a treatment of heat. And a repeat of this experiment with a different mass of water (say 0.5 kg) in the saucepan could achieve several things: If the heating were carried out for the same period as before, it would show that the temperature rise was greater in this instance. This would lend support to the idea that heat delivered was proportional to mass times temperature rise, although the new value of this product would not be exactly the same as the earlier one. Attention could be drawn to the fact that no allowance has been made for the container of the water, or for possible loss of heat from water and container to the surroundings. It would not be worth the trouble to attempt to correct for these factors in an experiment like this, but pupils should be aware of them and realize that they would

/have.....

have to be taken into account in accurate work on the measurement of heat.

Either at this stage or after the introductory experiment, the calorie and kilocalorie can be introduced as (temporary) units of heat. Perhaps it is worth mentioning to pupils that there is no reason to suppose that exactly the same amount of heat is needed to raise the temperature of a given mass of water from, say, 90°C to 91°C as is needed to raise it from 20°C to 21°C. The point is not one of any practical importance for any experiment that might be done in the school laboratory, but it would be just as well for pupils to receive an early intimation of the fact that specific heat is not necessarily independent of temperature.

So far we have not said anything about a problem which seems to bulk large in the minds of most teachers, namely the need to make clear the difference between heat and temperature. It seems to us that this problem need not be a serious one if these concepts are approached in the manner suggested in these pages. According to this approach, temperature is introduced by way of the concept of thermal equilibrium which, be it noted, is <u>not</u> defined with reference to heat. Heat, on the other hand, is introduced as "something which we give to a body to raise its temperature" or "something which a body gives up when its temperature drops"; simple experiments of the kind described above show that when the same amount of heat is supplied to different masses of the same substance, different rises in temperature are produced. It is only when there is too much talk of "heat" in introducing temperature that the problem need become acute. Incidentally, the practice of attempting <u>at an early stage</u> to "explain" temperature in terms of average molecular kinetic energy is not, we think, likely to lead to a clarification of the difference between temperature and heat.
One method used to link temperature and average kinetic

/energy.....

energy of the molecules is to do a simple experiment on the increase of gaseous pressure with temperature at constant volume, and then to argue that in terms of the molecular explanation of pressure, the experiment shows that increased temperature is accompanied by increased molecular kinetic energy. This is true enough; but experiment and argument cannot by themselves clearly establish that "temperature" is associated with <u>average</u> molecular kinetic energy while "heat" is associated with <u>total</u> molecular energy. This kind of clarification can only come when a <u>quantitative</u> attempt has been made to link pressure with average kinetic energy; and at that stage it is probably better to replace "heat" with "internal energy" anyway.

The concept of specific heat or "specific heat capacity", as it is now officially called, is needed in an introductory physics course, even if only to make sense of the measurements by which one seeks to show that "heat" is a form of energy. Traditionally, specific heats would be determined by the method of mixtures in the elementary physics laboratory, but it is extremely desirable to get away from such a messy and clumsy method. A sensible procedure would be to try to adapt the principle of modern methods of measurement of specific heat to the school laboratory. This has been done in the Nuffield course by the inclusion of an experiment in which a block of aluminium is electrically heated for a certain length of time and its rise of temperature measured. This experiment rests on the assumption that the electric heater supplies the same amount of heat in the same period of time to whatever it heats; thus if it has been established in a previous experiment in which 1 kg of water was heated for 5 minutes that, say 5×10^3 calories were supplied by the heater, it is now assumed that the heater will supply 5×10^3 calories to the block of aluminium if it heats it

/for.....

II

for 5 minutes. The Nuffield Guide refers to this assumption as "a large unsupported assumption, which we should offer in an apologetic tone" and "the weak point of the whole experiment." Perhaps so; but it must not be thought on this account that this assumption makes the experiment inferior in principle to a "method of mixtures" experiment. In using the "method of mixtures", it is necessary to make a corresponding "heat conservation" assumption, namely that in the absence of heat losses to or gains from the surroundings, the heat lost by the components of the mixture that cool down is equal to the heat gained by the components that heat up.

In this Nuffield experiment, the block of aluminium has holes drilled in it to take the thermometer and the heater, and oil is put into the holes to provide good thermal contact. The result of the experiment is of course that the same amount of heat produces a bigger rise in temperature in the block of aluminium than in the same mass of water. This is now explained by saying that the amount of heat needed to heat 1 g (or 1 kg) of aluminium through 1 Celsius degree is less than 1 cal (or 1 kcal) - and so the idea of the specific heat of a substance can be introduced. The fact that the experiment in question cannot yield up an "accurate" value for the specific heat of aluminium is of no consequence here; it has helped to introduce "specific heat" and shown how, in principle, it can be measured, and that is enough. There is really no need to elaborate the concept of specific heat any further in school physics; certainly there is no point in having pupils spend time on measuring specific heats of various substances. "Thermal capacity" is a concept of marginal usefulness at this stage and could perhaps be introduced; if it is, no advantage would be gained by introducing "water equivalent" as well. Those aspects of specific heat

/and.....

and thermal capacity which make the concept of particular interest to physicists, viz. the comparison between predicted and observed values of molar thermal capacity, go beyond what can conveniently be covered in a school course.

Finally we come to "latent heat" (or "specific latent heat"). It is undoubtedly important for a pupil to know about the existence of this quantity, for this is something that he will be able to explain in terms of the kinetic picture of matter. But the existence of "latent heat" can be shown quite simply by means of the well-known "cooling-curve-of-naphthalene (or paraffin-wax)" experiment, and by the fact that when water boils its temperature stays constant. Whether the pupil should have to carry out the usual determinations of the latent heats of fusion of ice and evaporation of water is debatable, to say the least. Perhaps it can be argued that these experiments show that (i) heat absorbed or given up in a change of state is proportional to the mass of the substance changing state (ii) the latent heat of vaporisation of water is much bigger than the latent heat of fusion of ice. Point (i), which "commonsense" might lead one to expect, is "shown" by the fact that different pupils using different masses of water or ice obtain much the same values for the latent heat in cal/g. But the very definition of a latent heat is built upon this assumption, the quantity is of course defined before it is measured, and it seems to be common practice among teachers and textbook writers to take this point for granted anyway. As for point (ii), if it is really desired to demonstrate this experimentally, this can always be done in an approximate fashion by showing that it takes longer using the same (electrical) heating supply to boil away a given mass of water than to melt the same mass of ice.

/It.....

It is of course possible to take the view that as the calorie and kilocalorie are not accepted SI units of heat, one ought to try to dispense with them altogether. We do not think it is essential to go as far as this: one can introduce these units along the lines suggested above, and then abandon them after the relationship between heat and work has been studied experimentally. If it is decided to do without the calorie altogether, then the following teaching sequence is recommended:-

(i) Establish experimentally by one of the methods outlined above that heat delivered is proportional to mass times temperature rise for a given material.

(ii) Carry out an experiment to show quantitatively that "heat" is a form of energy. Spurgin (1969) suggests measuring the specific heat capacity of a copper block by heating it with a friction band. He says "If this can be done by a variety of pieces of apparatus, it will be illustrating the essence of Joule's work, that a certain amount of mechanical energy will always produce the same amount of heat, in terms of a certain rise of temperature for a certain amount of a particular material. This will appear from these experiments in the constancy of the specific heat capacity obtained, irrespective of the method used." We would prefer that at this stage the result should be expressed in the form "the work done in joules divided by the product of the mass and the temperature rise is a constant."

(iii) On the basis of the result established in (ii), suggest that it is legitimate to use the joule as a unit of heat, and introduce the concept of specific heat capacity with the unit joule per kilogramme degree Celsius (the joule per kilogramme kelvin can be kept for use later).

(iv) Measure the specific heat capacity of water in $J\ kg^{-1}\ {}^{o}C^{-1}$ by the method of mixtures, using a copper calorimeter and the result obtained in (iii) and (iv) (cf. Spurgin, op. cit.)

/(v).....

(v) If necessary, measure other specific heat capacities and specific latent heats using whatever method is desired.

To proceed in this way would require a revision of existing South African syllabuses. Those concerned with the drafting of these syllabuses ought certainly to consider such a revision. Obviously if they do decide to implement it they must arrange the new syllabus in such a way that pupils are prepared for this approach to "heat" by being given an adequate understanding of the measurement of work. This can be done within the general structure of the existing junior secondary school syllabus.

3.12.3 The transport of heat

This topic, treated qualitatively, was once a favourite in secondary school physical science syllabuses in South Africa. It seems now to have disappeared from the senior secondary school, but is still likely to turn up in the junior secondary school or even in the primary school. We feel that there is really no case for including a quantitative treatment of heat transfer in a syllabus which goes up to O-level or its equivalent or a year beyond. Inclusion of quantitative thermal conductivity would achieve little purpose beyond making it possible for careless examiners and teachers to set problems of the kind so savagely pilloried by Warren (1965). As for radiation, there would be no hope of being able to present a derivation of, say, Stefan's law or Planck's law in such a school course, and it would be a pity simply to hand these results out to pupils. (There is, of course, no hope of being able to derive them experimentally either.)

Pupils should certainly be aware in a general way of the three modes of heat transfer. We feel that there is no need to spend much time on this. Easy experiments to show some of the features of conduction and convention can

/be.....

be done in the primary school, which is surely the best place for them. The Nuffield course allots a fair amount of space to slightly more advanced experiments on heat transfer; these would be most useful should any teacher find that he can spare the time for them, but it seems better to treat this sort of thing as optional in the secondary school.

If this topic is treated in the secondary school, some explanation of the various modes of heat transfer will be called for. Convection presents no problems in this respect. With conduction a little more care needs to be exercised. A typical textbook explanation is the following:- "As een end van 'n metaalstaaf verhit word, is die ander end ook spoedig warm. Die molekules wat verwarm word, ossileer met groter en groter amplitudes. Deur botsings met buurmolekules word kinetiese energie oorgedra en hierdie molekules begin ook hewiger ossileer." (When one end of a metal rod is heated, the other end soon gets warm. The molecules that are heated oscillate with larger and larger amplitudes. Kinetic energy is transferred through collisions with neighbouring molecules, and these molecules also begin to oscillate more vigorously) (Prinsloo et al 1965). This explanation can be faulted on two counts. First, it is misleading to talk of molecules being heated; "heat" and "heating" are essentially macroscopic concepts. Second, it is incorrect to say that kinetic energy is transferred through collisions with neighbouring molecules. In metals (and the above account does speak of a metal rod), the "molecules" play a negligible part in thermal conduction. In non-metal solids and liquids, this conduction takes place by a process that can be described as a kind of wave motion (very high frequency acoustic waves, in fact) or, more precisely, as "phonon transport". (Popular accounts can be found in Sproull 1962, Warren 1965 and Ziman 1967.) The easiest way

/out.....

out of this difficulty is probably to use the account given in the Nuffield Guide II: "We describe conduction as some process of handing heat on from one bit of stuff to the next, like a message being handed along a line of pupils from neighbour to neignbour". Use of the phrase "some process" avoids the problems of going into detail about the "process" in question, while the picture of a message being handed along a line of pupils could, with some effort, be regarded as an analogy with wave motion!

Radiation needs even more care in description. Here are some statements about "radiant heat" from a South African textbook (Gordon and Neser, no date): "Radiant heat, light, X-rays, etc., are all fundamentally the same; they differ only in their wavelengths.....Radiant heat is only one form of the radiant energy which the sun emits, and differs from light in that its wavelength is longer, and that it exerts a heating effect on the skin, instead of stimulating the optic nerve.....Another name for heat waves is infra-red rays...Both light and heat waves, when absorbed by bodies on which they fall, are converted into heat energy. Heat waves possess the greater heating effect; light waves produce also the sensation of vision by their chemical effect on the retina of the eye". It will be evident that these statements betray a fundamental confusion about the nature of "radiant heat"; that is, the authors believe that what they are pleased to call "heat waves" have the greater heating effect. As the Nuffield Guide II puts it; "All the ordinary incandescent sources we know, including the sun, emit infra-red radiation in a much more powerful stream than visible light. The reason why a thermopile or thermistor exploring through the spectrum is heated so much more in the infra-red is not that infra-red consists of "heat radiation" but simply because there is more of it!"

/It.....

It is clear that if teachers are to explain the nature of so-called "radiant heat" correctly, there are certain points that they will simply have to make:

(i) that it is wrong to distinguish between "heat radiation" and other kinds of electromagnetic radiation, since all electromagnetic radiation will produce heating of a surface which absorbs it;

(ii) that radiation can only produce "heat" when it is absorbed. In recommending this approach, the Nuffield Guide refers to it half-apologetically as "dogmatic". But the above-quoted string of misstatements about "radiant heat" were all dogmatic-and wrong; a little dogma in this connection will not come amiss if it is right.

If the teacher is concerned about avoiding confusion on the part of pupils on the subject of radiant heat, then he will have to exercise care in his interpretation of some of the traditional experiments and demonstrations used to illustrate this topic. Take the experiments that are used to show that "radiant heat" can be reflected and focused like light. What these experiments show is that <u>infra-red radiation</u> can be reflected and focused like light; as we have seen, it is seriously in error to identify "radiant heat" with infra-red radiation alone. Thus in interpreting these experiments the teacher should make clear that they show that that part of the radiation from the source which does not excite our sensation of vision but which is nevertheless capable of being absorbed by our skin and so felt as heat, is in certain respects similar in its behaviour to visible light. It is also important to try to show in the course of these experiments on radiation that visible light contributes its share to what we feel as "radiant heat". This is not easy, as powerful sources of radiation such as

/electric.....

electric radiators pump out most of their energy in the infra-red region. Again the Nuffield Guide comes to the teacher's aid in suggesting an experiment in which a sheet of glass is used to cut out most of the infra-red radiation from an electric radiator, while of course still allowing the red light to get through. Here it is urged upon the pupil that because he can still <u>see</u> some radiation, the glass cannot have cut it all out; and he is asked to move closer and closer to the heater with the sheet of glass interposed between it and his cheek, which acts as detector, so that he will <u>feel</u> as well as see that some radiation is still coming through the glass. (The heater is mounted behind an asbestos sheet with a 1 inch-diameter hole in it).

3.13 <u>Atoms and molecules as the basis of matter: the kinetic theory of matter</u>

3.13.1 <u>Elementary physics and the particle structure of matter</u>

In Section 2.3, we noted that there is one school of thought which advocates that the introduction to physics should be through a consideration of atoms, molecules and electrons, and that there is a strong tendency in newly-developed courses in school physics to stress the atomic nature of matter as early as possible. We indicated our own lack of sympathy with a purely dogmatic introduction of atoms, molecules and electrons, but pointed out that there were ways in which these concepts could be introduced without requiring mere blind acceptance on the part of the pupil. Indeed, we feel it to be both realistic and sensible to foster an acquaintance with these concepts from an early stage in the pupil's study of science, provided that ample opportunity is given him of considering evidence that points toward the reality of atoms, molecules and electrons. We might not go as far as the Nuffield course in treating atoms as "familiar

/things.....

things from the beginning", but we would certainly think it desirable to let the pupil do experiments and hear arguments which suggest that matter might be made up of microscopic particles.

3.13.2 Experimental evidence in support of the particle model of matter

In the realm of physics, there are various kinds of experiment which might be used in support of the view that matter is made up of microscopic particles. None of them would be conclusive on its own; the interpretation of the results as being suggestive of, say, the existence of atoms will invariably be found to rest on some or other assumption, which should be clearly stated. But the cumulative effect of the study of a number of such experiments, the results of which are all consistent with the atomic view of matter, should strengthen the validity of this view for the pupil. These categories of experiment are (i) dilution experiments; (ii) diffusion experiments; (iii) Brownian motion; (iv) study of crystals; (v) measurement of the thickness of a monomolecular layer; (vi) the study of radioactive decay with the aid of an expansion cloud chamber. (As we are concerned for the moment with experiments suited to the junior secondary school, we leave out of account more advanced experiments such as Millikan's oil drop experiment).

(i) Dilution experiments (cf Lewis 1964) Experiments of this type show that matter can be spread out very thinly. In a typical experiment, colouring material like fluorescin or purple dye could be dissolved in water and the resulting solution could then be diluted more and more, say by a factor of ten each time. The colour would still be detectable even if the dilution were very great; for example, a concentration

/of.....

of 1 x 10^{-8} g of fluorescin per ml of water still gives a detectable uniform colour. It is clear that the best that this experiment can do (and this is admitted by Lewis, for example) is to show that if matter is made of "ultimate particles", then there is a certain upper limit to the size of these particles. This might seem to be a possible starting point for a consideration of the particulate nature of matter. However, it is reported in the Nuffield Guide (I) that teachers who have shown this sort of demonstration to their classes are "almost unanimous" in recommending its omission. The trouble appears to be that pupils are prone to misinterpret what the teacher has said and regard the demonstration as telling them "how small atoms are" or as providing evidence for the existence of atoms. It may be that pupils in the "Year I" stage of Nuffield work would be more likely to be confused in this way than the somewhat older pupils in Standard Six or Seven in South African schools. But perhaps we can accept the summing-up of the Nuffield Guide: "Now that we know so much about atoms, and can make good measurements, the experiment is only of historical interest - unless we can use it with very careful logic."

(ii) Diffusion experiments There are many experiments one can do to show diffusion of gases, liquids and solids, and certainly pupils should see examples of all three during their years at school. Lewis (1964) recommends the following simple experiments for liquid and solid diffusion respectively:-

To demonstrate diffusion in liquids, a tall vessel can be carefully filled with three layers of liquid of roughly equal volumes, the bottom layer consisting of a strong sugar solution, the middle layer of concentrated copper sulphate solution and the top layer of distilled water. Simultaneous upward and downward diffusion of the copper sulphate will take

/place.

place.

A hard-set gelatine (10% to 20% in water) is suitable for demonstrating that solid-like substances (in the sense that they are not fluid) can show diffusion. A closed boiling tube having gelatine in its lower half and potassium chromate solution in the upper half can be used to show the diffusion of gelatine in any desired direction. Lewis recommends that three such boiling tubes be made up and mounted respectively upright, inverted and horizontal. A fourth tube, to be used as control, should be completely filled with gelatine and mounted in any fashion. After a few hours diffusion of the gelatine in all directions should be apparent.

Of the various experiments to illustrate gaseous diffusion, the most spectacular is perhaps that in which bromine is allowed to diffuse first into air and then into vacuum. Provided bromine "pigs", suitably-shaped glass tubes and a rotary vacuum pump are available and provided the precautions appropriate for experiments involving bromine are observed, these experiments are easy to carry out and highly rewarding (Nuffield Guide IV). In the one case, the bromine diffuses slowly into the air; in the other, it appears to fill the vacuum instantaneously.

To be of real use in interpreting the nature of matter, these experiments on diffusion should be done when the pupils have some model for gaseous behaviour in mind. In suggesting or supporting such a model, demonstrations of Brownian motion are invaluable.

(iii) Brownian motion There are a number of ways of demonstrating Brownian motion. Smoke cells suitable for viewing under a microscope are obtainable commercially from various sources; these can be used to show the Brownian motion of the smoke particles as they come under irregular bombardment

/from.....

from the molecules of the air. To show this motion with particles in suspension, one can use a pinhead-sized droplet of colloidal graphite in a few ml of distilled water; the disadvantage of this arrangement is that it has to be viewed under appreciably higher magnification than the smoke cell. Although Brownian motion may seem much more definite in its implications than some of the other demonstrations and experiments discussed in this section, one should remember that even it only acquires proper meaning in the context of these other demonstrations and experiments.
As Warren (1965) has pointed out, "... all that a qualitative study of [Brownian motion] can show is that matter which is macroscopically at rest is microscopically in turbulent motion". The "turbulent motion" becomes meaningful if we relate it to the diffusion of matter and to the fact that liquid films have a limiting thickness.

(iv) The thickness of a monomolecular layer The importance that is nowadays attached to experiments in which the thickness of a monomolecular layer of liquid is measured may be gauged from the fact that the Nuffield Guide I devotes some eighteen pages to a discussion of the experiment, its background and its interpretation. The general principle of the method is the following:-

A drop of suitable liquid is allowed to form a layer on the surface of water in a large tray. From the area of the layer and the known volume of the amount of liquid in the layer, the thickness of the layer can be calculated. Layers formed by two and three drops are shown to have respectively twice and three times the area of a layer formed by a single drop. This indicates that a lower limit to the thickness of the layer of liquid has been reached, and this lower limit can be identified with one of the dimensions of the molecules of a liquid.

/There.....

There are currently at least two "standard" versions of this experiment in use. In one, due to the PSSC, the method is as described above, with a 0.5 per cent solution of oleic acid and ethyl or methyl alcohol being used as the liquid. This is dropped from a dropper, and the alcohol soon evaporates, leaving a layer of pure oleic acid behind.

The Nuffield Guide criticizes this version on the grounds that use of a solution of the layer-forming liquid leads to "much too complicated arithmetic". In place of this, it recommends the use of (undissolved) olive oil. The drop from which the layer on the water is formed is adjusted by the experimenter to a diameter of 0.5 mm. Only one drop is used in the experiment; that is to say the step is omitted in which it is shown by the use of more than one drop that a lower limit to the thickness of the layer has been reached. Instead the pupils are expected to be guided by the teacher's accompanying "patter" towards making Lord Rayleigh's guess that the layer must be one molecule thick if it has gone on spreading until it can spread no more. The Nuffield version of the experiment is also preceded by a quick sequence of demonstrations of surface tension phenomena.

Obviously each of these versions has its advantages and disadvantages. The Nuffield version is, as intended, arithmetically much simpler than the PSSC version. On the other hand, in the Nuffield version the pupil is required to make a fairly delicate adjustment of the size of his drop, an operation which involves the use of a scale, wire loop and magnifying glass, whereas in the PSSC version, in which much larger quantities of liquid are actually dropped on to the surface of the water, no delicate adjustment of drop size is necessary; the experimenter measures the number of drops in one millilitre, and then has a good enough measure of the

/volume.....

volume dropped on the water. The Nuffield criticism of the allegedly excessive amount of arithmetic involved in the PSSC version loses some of its force if the pupils concerned are thirteen- or fourteen-year olds instead of eleven- or twelve-year olds. Thus the South African teacher may feel that there is little to choose between the two versions but that the PSSC version is perhaps to be preferred because it is simpler to carry out. At any rate no deep question of principle is under debate here. The experiment is undoubtedly of great value, and pupils should do it themselves.

(v) The study of crystals A most fascinating field for qualitative investigation by pupils, and one which looks forward to far more profound investigations, is the elementary study of crystals and crystal growing. It is significant that both the PSSC and the Nuffield Physics give this activity (and it is essentially an activity) a prominent place. Indeed it virtually begins the Nuffield course, and one can see why the planners of that course were keen to introduce it so early - it furnishes an early opportunity for bringing in the idea of atoms in an unforced manner. There are three main lines of investigation in this preliminary look at crystals: looking at crystals (preferably of "familiar" materials) with the naked eye and through a microscope, growing crystals (and examining the finished product) and making models of crystals. In addition, it is possible to study the cleavage of crystals. All this is very simple in essence, yet at the same time can be very instructive. We would suggest that it is not necessary to back up this work with impressive-looking models of complicated crystal structures, or with electron or field-ion emission microscope photographs of atomic patterns. But if the teacher would like to supplement his teaching at this stage with this material, we would suggest that he do it in the spirit of wanting to whet his pupils' appetite for knowledge and

/understanding.....

understanding of how this kind of information can be obtained.

(vi) The study of radioactive decay with the expansion cloud chamber To some extent, we have already discussed this point under the heading of "energy". (See Section 3.9.5.)

Another elementary demonstration which can be regarded as suggestive of an atomic picture of matter is one showing contraction of volume when salt dissolves in water. This would seem to exhaust the repertoire of demonstrations suitable for this purpose in the junior secondary school. With one exception (the measurement of the thickness of a monomolecular layer) all these experiments have been qualitative. This is an excellent reason for waving the "warning flag" recommended by the Nuffield Guide: the teacher should try to convey to his pupils that however powerfully suggestive the body of evidence assembled in these simple experiments might be, it cannot be thought of as conclusive. It is only when a theory has been developed on the basis of the atomic or kinetic picture of matter and the numerical predictions of this theory have been verified by experiment, that one can begin to acquire a physicist's feeling of confidence in the correctness of the atomic view. Indeed, one can argue that such a feeling of confidence cannot be achieved without a consideration of certain quantitative chemical evidence. This is a problem we shall have to leave to the chemists; it is to be hoped that they would draw up their contributions to physical science syllabuses with due regard for the problem, and that they would not succumb to the temptation of wanting to take atoms and molecules fully for granted right from the beginning.

"The evidence for the existence of atoms is far too profound for young school children to understand", says Warren firmly in his book on the teaching of physics (1965). We may accept this proposition as self-evident, and at the

/same.....

same time feel entirely justified in presenting to these "young schoolchildren" the kind of evidence that has been outlined above. What Warren is apparently saying is that we cannot ask pupils to do experiments on diffusion, Brownian motion, crystal structure and monomolecular layers and then expect them to arrive unaided at the conclusions that matter consists of extremely small "ultimate particles" in ceaseless random motion, that in crystalline solids these "particles" are arranged in a regular lattice structure and that in a gas they are rushing about randomly inside the container, occasionally colliding with one another and with the walls. As he puts it, "Difficult scientific ideas cannot be deduced by childish reasoning from simply observable phenomena." Certainly they cannot. But it is surely not the belief of those who favour the early use of the above experiments that children can make such deductions without help. Clearly some very positive "teaching" will have to be done here in order to direct the thoughts of pupils along the right lines; but this is not a valid argument against the introduction of this topic. What is important is that the teacher should not be too dogmatic in his approach and so deprive pupils of the opportunity of exercising critical thought. For them, learning about atoms should not be a matter of mere conditioning to the process of acceptance of what is laid down by Authority.

3.13.3 The integration of the particle picture of matter with the teaching of the entire syllabus

Consideration must now be given to the problem of how to integrate this "modern" material with the rest of what is considered essential in a school physics syllabus. One way of dealing with the problem is to decide that it is not worth solving - that material concerned with molecules

/and.....

and atoms and their constituents should be treated under a separate heading as "modern physics" (or some such title). It is safe to say that no teacher today would be happy with this approach. Even the poorer teacher or textbook author finds it convenient to use ideas derived from the microscopic world to explain macroscopic phenomena; and it would be just as well to regularize this kind of thing. The extreme alternative to the segregation of everything that could be classed as "modern physics" under a separate heading, is the type of physics course recommended by the OEEC special committee (Brown and Clarke 1960), to which we have already referred. This starts with "The particle nature of matter. General ideas of the differences between atoms, molecules and electrons", and then, after a preliminary study of the fundamentals of mechanics, leads on to "The simple properties of gases and vapours; ideas of temperature and pressure as phenomena of particles in motion: states of matter and change of state in terms of aggregations of atoms and molecules; chemical bonds", and in due course to "The structure of atoms; electrons as carriers of electricity; electrostatics; thermionics; electromagnetism; magnetism;" with much more to follow. Some of the important facts concerning atoms, molecules and electrons are thus stated at the outset, and much of what follows is deduced from or built up with the aid of these facts. In Section 2.3, we have given our reasons for the rejection of this kind of approach. We evidently need some middle way in which we can somehow use the atomic picture of matter as a kind of unifying feature of the course without any premature disclosure of the facts concerning microscopic matter, and without violating those principles which we consider to be essential to the proper teaching of fundamental physical concepts. We should not expect that the proposed "middle way" would necessarily abolish the

/traditional.....

traditional subdivisions of physics; there is no compelling reason why it should, since our aim is the comparatively modest one of trying to integrate "modern" material within our teaching of the fundamentals of the subject.

A possible "middle way" is to begin with some of the experiments discussed in this section (e.g. the study of crystals) in order to introduce the hypothesis that matter is built of very tiny particles which, for the moment, can be thought of as indivisible. In our view, this is all that is needed for an introduction; it is no use confronting the pupil with too many ideas about the microscopic behaviour of matter at this early stage. The problem of how big atoms might be can then be discussed; this can be made the "excuse" for a general treatment of the measurement of length, which could pave the way for the experiment on the thickness of a monomolecular layer. This could then raise the problem of how atoms "stick together" in solids and liquids, and of what they could be expected to be doing if they were constituents of a gas. Discussion of these points brings in the concepts of "force" and "motion", which could then be treated along conventional lines. Of course, this initial treatment of force would not take the pupil any nearer towards understanding the nature of the forces between atoms or molecules in matter in bulk; but it would be possible to establish quite simply the idea that if atoms are pulled too far apart, there is an attractive force between them, and that if they are pushed too close together, there must be a repulsive force between them. This can be done by means of the usual simple argument from the fact that matter in the solid state obviously resists extension or compression. (It is very important for what follows to stress the idea that if matter consists of atoms (or molecules), then the force between neighbouring atoms must be repulsive when the separation is very small and attractive

/when.....

when it is bigger. As Warren (1961, 1965) in particular has pointed out, serious misconceptions can be created in the minds of pupils by failure to indicate the role of the short-range repulsions between atoms or molecules. The teacher's task is not complicated unduly by this requirement; if he uses a model of a spring joining two balls to represent the intermolecular force, he could suggest that the spring was slack when the molecules were at their normal separation, so that it is compressed when the molecules are closer together than this and stretched when they are further apart.)

Beyond this point, the use of the theme of "finding out about atoms" as a means of motivating the study of certain other areas of physics may seem to be somewhat forced - if it has not already become so! Thus it is difficult to see what "atomic excuse" we can invent for finding out about energy, unless it be that pupils have heard about atomic energy and would like to know more about it. But this kind of problem would only arise if we stick rigidly to the premise that our main business is finding about atoms, and that all else is therefore incidental. The "atomic motif" can always be re-introduced at appropriate moments, sometimes in advance of the experimental evidence it is intended to illuminate or elucidate, sometimes after it. Thus the simple, qualitative model of a crystalline solid or of a liquid could be used, in conjunction with ideas about energy, to predict the existence of latent heats of change of phase; or one could use a simple model of an ideal gas to predict Boyle's Law. Alternatively, one can interpret the results of "Joule-type" experiments on the relation between "heat" and work in terms of what happens to the molecules composing matter; the transmission of sound waves through matter in its condensed or gaseous phases can likewise be given a simple interpretation, as can the electrical behaviour of

/matter.

matter. Thermionic emission can be introduced much earlier in the course than is customarily the case, so that instead of forming part of an isolated section called "electronics" or "electron physics", it can be used to provide evidence of the nature of the particles responsible for the conduction of electricity in metals. Other examples of this approach will be found later in these pages.

Perhaps purists will argue that although we are careful in this kind of teaching to stress that we are using models to aid us in our interpretation of phenomena, we are to some extent cheating, because we deliberately use only those models which are known to correspond to some degree with the truth as we now know it. To this objection we may reply that there simply is not time to consider all possible models which might fit some or all of the facts and then go on to discover why these models should be rejected; that nevertheless in some instances (e.g. the study of the nature of light) it is advantageous to discuss in some detail the merits of rival models or theories, and that this is in fact done; and that in the limited time available it is sensible to keep returning to the atomic-kinetic model of matter, since so wide a range of phenomena can be explained in terms of it. We stress again that this does not (or at any rate should not) mean that this model is to be taken for granted; the teacher should at all times encourage a critical attitude towards it, since its validity will be all the more appreciated if it shows itself to be proof against criticism.

3.14 Electricity

3.14.1 Electricity in the revised syllabuses

One of the ways in which elementary physics teaching can claim to have advanced in recent years, particularly in South African schools, is in the improved approach to the

/teaching.....

teaching of electricity. In former years, electricity was excluded altogether from the junior secondary school years, and in the senior school it was expected to be taught without reference to essential mechanical concepts. Considerable attention was devoted to current electricity, explained entirely in terms of the water analogy, with Ohm's law forming the climax of this part of the course. Electromagnetism and electromagnetic induction were dealt with in an entirely qualitative and often inaccurate manner. Electrostatics was treated as an afterthought and received scant attention. The introduction of the new Joint Matriculation Board syllabuses in Physics and Physical Science and the subsequent Common Basic Syllabuses in these subjects completely altered the situation. It was realised that the extremely great importance of electricity in modern science and technology, and the difficulties encountered in teaching it at an elementary level, made it imperative for much more time to be allocated to it in the secondary school. At the same time it was also realised that the phenomenological aspects of the subject could be dealt with in the junior secondary school, leaving more time available for the teaching of electrical concepts in the senior school. The result was that much of what was formerly regarded as Matriculation-level work in electricity is now to be found in South African syllabuses for the junior secondary school. This has placed a heavy responsibility on the shoulders of the teachers of pupils in Standards Six, Seven and Eight, who are now faced with the task of providing an adequate foundation for the difficult and crucially important work in electricity in the senior years. There is a real danger that this foundation will be one of confusion and bewilderment if certain guiding principles, which we shall try to outline in due course, are not heeded.

/3.14.2 .....

3.14.2 Electricity via electrostatics

A significant feature of these new syllabus revisions is the increased amount of electrostatics included in both the junior and the senior divisions of the syllabus. This reflects an interesting change in attitude towards electrostatics on the part of physics teachers all over the world. There was a time when electrostatics was regarded with suspicion as a subject associated with some of the oldest and most disreputable-looking pieces of apparatus in the laboratory store cupboard. Except on very dry days, one had to go to extraordinary lengths to ensure that one's demonstrations had a reasonable chance of working and the effort expended seemed out of all proportion to the results achieved. As recently as 1960, we find Professor Rogers prefacing an excellent chapter on electrostatics in his book, "Physics for the Inquiring Mind" with the following apologetic remarks: "The study of electricity at rest... used to bulk large in elementary physics. It was all that was known of electricity two centuries ago, and tradition dies hard. It makes a poor beginning for modern electric circuits, so we have avoided it. Now you need some knowledge of it for atomic physics. How much you see and learn of this part of physics will depend on apparatus, weather and instructor. On the whole, the less the better." The reason for the change of attitude towards electrostatics is actually indicated in part by a sentence from the above quotation (" - you need some knowledge of it for atomic physics"); it is more fully set out by Harrap (1966): "...if one is to teach what has become important since [1890], the nature of electrical forces is one of the fundamental necessities. The case for teaching electrostatics depends on this greater relevance to recent physics, but in many ways the topic has now become easier to teach in itself and easier to relate to current electricity". To this we might add that by starting one's

/teaching.....

teaching of electricity with electrostatics, one is in a position to develop the concept of potential difference from the concepts of electric field, electric intensity and potential energy in such a way that the pupil can gain a proper appreciation of the significance of potential difference in an electric circuit soon after he has begun to study current electricity. Experience has shown that it is satisfactory to teach electricity in this way at first-year university level, and it is now standard practice for textbooks in introductory university physics to follow this approach. (Cf. McCormick 1965; Halliday and Resnick 1966; Sears and Zemansky 1964; Weber et al 1959, to mention only a few). However, because of the way in which electricity is divided into elementary and advanced sections in South African school courses, it is not possible to follow through this neatly logical presentation in such courses.

The reasons why electrostatics has now become easier to teach than before are to be found in the easy availability of plastic materials which hold a charge far more readily than do the ebonite and glass rods of old, and in the development of electrostatic generators (e.g. model Van de Graaff generators) which are cheap, robust and reliable. (Perhaps this is just an example of teaching needs stimulating the design and manufacture of the right kind of teaching apparatus.) The work of the P.S.S.C. in the U.S.A. has also prompted the manufacture of cheap electroscopes, which can be most helpful in teaching electrostatics. The result of this is that it is now possible for pupils to do elementary experiments in electrostatics for themselves and have a very good chance of finding out what their teacher hopes they will find out. There is no longer the danger that a laboratory period spent in class work on electrostatics will degenerate into a pointless farce.

/3.14.3.....

3.14.3 Introductory electrostatics

(i) Charging by rubbing

In the junior secondary school, there seems to be no compelling reason why work on electricity should not begin with electrostatics. Indeed, this could follow quite smoothly on a discussion of forces: the attraction of a rubbed plastic ruler for small pith balls or pieces of paper could be used to illustrate action-at-a-distance forces. (It would be worthwhile strategy for the teacher to demonstrate early on that electrical forces have no apparent connection with magnetic ones). After this beginning, pupils could go on to use plastic rulers or rods of various kinds to establish that similar objects, similarly rubbed repel each other and that in some cases, different rubbed rods attract each other. (Jardine's (1964) suggestion of using plastic rulers or strips mounted on watchglasses in these experiments works extremely well; the repulsive or attractive forces are readily noticeable, even to quite a large class watching a lecturer performing these demonstrations on a lecture-bench.) The term "charge" should be introduced at this stage. We would wholeheartedly endorse the suggestion of the Nuffield Guide that the teacher should "explain" the concept of charge purely in terms of what the pupil has seen happening in his experiments. He rubbed one end of a rod and found that as a result of that operation, the rubbed end of the rod was able to attract small objects and repel other, similar rods rubbed in the same way. The physicist's way of summing this up is to say that a charge has been given to the rod by rubbing it - the process of rubbing it has given it this new, special property which we call a charge. It should also be established in the course of these experiments that the attractive or repulsive forces are bigger

/when.....

when the attracting or repelling bodies are closer together.

Some teachers will want, even at this early stage, to "explain" the operation of charging a rod in terms of the motion of electrons. Indeed they may be required to do so by their syllabus or by their "Teachers' Guide". Thus the Teachers' Guide for Physical Science for Standards Six, Seven and Eight in the Orange Free State (1967) says: "Explain charge by friction as the removal of electrons." There are certain points to be borne in mind here. Use of words like "electrons" and "ions" should be very guarded at this stage. The pupil may have done some experiments the results of which encourage the idea that there might be things called atoms. Now he has done experiments which show him that under certain conditions, matter can become "charged"; also, these experiments seem to show that there are only two kinds of charge. There are various ways in which he can interpret these facts. Having drawn the reasonable conclusion that "neutral" matter might contain "positive" and "negative" charge in equal amounts (whatever might be meant by "amounts of charge"), as well as possibly containing some neutral matter, he could say: "Perhaps atoms consist of positively-charged and negatively-charged particles, which normally stick together. Then when I rub certain things together, several things could happen: perhaps some positively-charged particles move over from the one object to the other, or perhaps some negatively-charge particles move over in the reverse direction, or perhaps both kinds move at once. How can I tell which is right?" Of course there is nothing in these experiments which can supply the answer, and we ought to make sure that pupils realize this. Many teachers would be impatient with this attitude; "Surely, one can hear them saying, "there is no need for this kind of

/agnosticism.

agnosticism. We all know not only that it is the negatively charged particles which move across, but that these particles are in fact electrons, and there is no point in concealing this knowledge from pupils, who will probably have heard of it anyway." (Cf. the Nuffield Guide (II) on this point: "...pupils have heard of electrons and will themselves tell us it is the electrons that move. So we should in practice move with the times - and accept this and say it, if pupils do not say it for us.") But is our knowledge of this point so secure? It is certainly a well-established piece of textbook "folklore" that it is electrons which move over in frictional or contact electrification, but do any of us know of any convincing experimental evidence for this? Recently Harper (1967) has published an account of extensive research carried out in investigation of this problem, and his conclusions will be decidedly upsetting to firm adherents to the "electronic" explanation of these phenomena. Harper finds that free electrons play no role in contact and frictional electrification of insulators; that good insulators like polyethylene or amber "charge up" as a result of local heating produced by vigorous rubbing, or by the acquisition of contaminating ions; that the amount of charge acquired by these insulators is two to three orders of magnitude less than that acquired by glass or quartz under similar conditions, but that their electrical behaviour is so pronounced because they can retain their charge so well.

We mention these results in order to make it plain that there is a real need for caution in explaining macroscopic phenomena in terms of microscopic matter. In particular, we need to be careful in what is said about electrons and what they do. Certainly let us take the results of some of our experiments in electrostatics as suggesting that the "ultimate particles" of matter may be divisible into posit -

/ively.....

ively- and negatively-charged components (although we should remember here that the chain of reasoning really begins with the statement "If there are atoms..."); but in referring to those components as (positive) ions and electrons we should point out that this is really just a matter of affixing names to things of whose existence we cannot as yet be certain, and of whose nature we are largely ignorant. Pupils wanting more information about, say, electrons can be asked to specify just what it is they think they ought to know about an electron. Perhaps they might say that they would want to know its mass, how big it is, how much charge it has or even what an electron is. This could start a useful discussion on the need to know these things, with promises from the teacher that pupils will be seeing and hearing a great deal more about electrons in the near future.

(ii) The unit of charge

The Orange Free State syllabus for Standard 6 requires the unit of charge to be mentioned to pupils as being the total charge on 6.25×10^{18} electrons. The desirability of this can be questioned. On the one hand, it is certainly desirable for pupils to be introduced as soon as possible to the idea that charge is something that can be measured. This can be done by means of simple reasoning from experiments on the attraction or repulsion between charged bodies; if body B repels body C more strongly than body A does when it is held the same distance away from C as A was, then B is presumably more strongly charged than A - we say it has more charge than A. Of course this is very crude - ideally we should be talking about particles, not bodies - but the oversimplification is not so gross as to disguise the essential truth. This is really all the pupil needs to know about the quantitative aspects of charge at this stage; within the context of an essentially qualitative introduction to

/electrostatics.....

electrostatics there is nothing he can do with a unit of charge. Thus, telling the pupil that the unit of the charge is the coulomb does not help him in any way; nor does the information that the coulomb is the total charge on 6.25×10^{18} electrons confer any real benefit upon him. This is not a definition of the coulomb, and the teacher giving this information would have to make this point clear. It would be much more valuable to tell the pupil, "The official unit of charge is about a million times [or whatever we believe the factor to be] the amount of charge on the dome of that Van de Graaff generator or on that balloon we have charged up". This will give him a better idea of how big a coulomb is than the reference to 6.25×10^{18} electrons.

(iii) Conductors and insulators

This preliminary work on electrostatics could safely go on to include the difference between conductors and insulators, an introduction to the use of the electroscope, charging by contact and by induction, and the use of the Van de Graaff generator. The difference between conductors and insulators can be demonstrated in a number of ways; it is particularly easy to do if the electroscope is used as an indicator of charge. Again there will virtually be a compulsion upon the teacher to "explain" the different electrical behaviour of conductors and insulators in terms of electron movement, and again he needs to be careful in giving that explanation. Here he has to remember that there are conducting materials besides metals, and that in certain of these there could be movement of positive as well as negative charge carriers, and that the latter do not always have to be electrons. It is perhaps worth stressing that when the charge carriers are in motion, they move within the body of the conducting material and not just over the surface.

On the other hand, it is not possible to give pupils at this stage a proof that the excess charge on an isolated conductor

/must.....

must reside on its surface (but see below), and no particular advantage would be gained by mentioning it.

(iv) The principle of the electroscope

A straightforward way of introducing the principle of the electroscope is to use a pair of (coated) pith balls suspended from the same point; the giving of charge to the balls causes them to stand apart. The actual electroscopes to be used can be of very simple design; for example, a metal rod with a knob attached to one end and a gold leaf to the other, mounted in a conical flask by means of a rubber stopper, will be found to answer all the requirements of the earlier part of the course. Warren (1965) has stressed the need for earthing the case of the electroscope to secure maximum sensitivity, and points out that this requirement is often accidentally satisfied if the electroscope has a glass case and is allowed to stand on a wooden bench, since "glass and wood are under common conditions of humidity and uncleanliness sufficiently good conductors for the purpose". Thus one does not (in these experiments, at any rate) have to worry about using electroscopes with metal cases which can be seen to be eartned. This is fortunate, as such a detail could only prove an annoying complication here. The electroscope can be used, among other things, to show sharing of charge (including "earthing" by sharing of charge with the human body); as an indicator of the sign of charge; to show that the air can be made conducting by various means; and in connection with experiments on charging by induction.

Before showing that air can, in certain circumstances, be made to conduct electricity, it is necessary to show that in normal circumstances it conducts to only a negligible extent. The behaviour of a charged electroscope that is left to itself will serve for this purpose. Then one can show that an open flame held near the electroscope will cause

/The.....

the latter to discharge. This is not conclusive evidence that the flame does something to the air; pupils could be invited to suggest other possibilities (e.g. that particles from the flame are being deposited on the knob of the electroscope (Gledhill 1968)), and to suggest ways of investigating and eliminating these. It should also be shown that if a radioactive source is held near the electroscope, the latter can be discharged. Again this is not conclusive evidence for ionization of the air by the source. If in both the above cases it is thought that the electroscope is being discharged by the accumulation on the knob of charged particles of appropriate sign ejected by the flame or source, this could be checked by giving the electroscope a charge opposite to its original charge and repeating the experiment.

We do not think it premature to use the electroscope at this juncture to show the ionization of air by flames or radioactive sources. As far as the latter demonstration is concerned, if we use it in conjunction with what is seen to happen in a cloud chamber when a radioactive source is inside it, this evidence can help to clarify what radioactivity is all about (and also the process of track formation in the cloud chamber). The Nuffield Guide frowns on the use of an electroscope for this purpose in Year I of the Nuffield course "because, while it is just as mysterious [as a Geiger counter], it fails to slow the individual events". But in the present context, the electroscope is not mysterious, so this objection falls away.

(v) Electrostatic induction

For introducing electrostatic induction, experiment 4 from Section 2 of "Modern Physical Science Report No.1" (Lewis 1964) serves very well indeed. In it, equal and opposite charges are induced on a pair of copper calorimeters mounted on insulating stands (e.g. wax blocks or polythene

/tiles.)

tiles). The calorimeters are placed in contact, a charged rod or strip is held near one of them and they are separated while the rod is kept in position. An electroscope is then used to test the sign of the excess charge remaining on each conductor. The conductors are then touched together and the electroscope is used to check if any charge remains on them. After this experiment has been performed and interpreted, it is possible for pupils to explain for themselves why the leaf (or leaves) of an electroscope should deflect when a charged object is brought near the electroscope knob. The logical next step is to give the electroscope a permanent charge by induction. The explanation often given of certain details in this process of charging of the electroscope leaves much to be desired (cf. the articles on "Charging an Electroscope by Induction", in "The Physics Teacher" 1965). Suppose a negatively-charged rod is held near the electroscope. Prior to the "earthing" of the electroscope, there will be an induced positive charge on the knob and an induced negative charge on the leaves. What happens when the experimenter's finger touches the knob is usually described thus: "The electrons flow via the body to the earth" (cf. Gordon and Neser, no date, Prinsloo et al 1965). We pass over the question of how this explanation ought to be modified if the experimenter is wearing shoes with clean rubber soles. What is more important is that it is by no means obvious to the average pupil or student that the electrons should do what they are declared to do. The major difficulty here is this: why is it that the electrons, which have been driven down to the leaves by the presence of the negatively charged rod, now decide to travel up again and then along the experimenter's finger to "earth"? The problem is solved if the necessary extra step in the description (and in the usual accompanying sequence of diagrams) is

/inserted,.....

inserted, viz. an indication of what happens as the experimenter's finger approaches the knob of the electroscope. The finger necessarily has positive charge induced on it also, so there is less induced positive charge on the knob, and some of the excess free electrons move away from the leaves. The closer the finger approaches the electroscope, the fewer excess electrons there will be on the leaves; when the finger finally touches the electroscope, the excess electrons on the body-electroscope system are all transferred to the experimenter's body - the positive charge induced on the finger pulls them across.

Since the subject of electrostatics was introduced by the phenomenon of charged objects attracting small objects, it would be well for pupils to gain some understanding of why a charged object can attract an uncharged pith ball, piece of paper or speck of dust. The explanation lies in charging by induction, but since the light objects just mentioned are made of insulating materials, something will have to be said about the mechanism by which they can acquire temporary charges by induction. This can be explained as a slight "pulling-apart" of the positively- and negatively-charged components of the material because of the presence of a nearby charged object. It would not be necessary here to pursue the behaviour of dielectric materials in electric fields any further.

(vi) The Van de Graaff generator and electric discharges in air

The Van de Graaff generator can be introduced here as a black box capable of giving charges to its own domes and to objects placed in contact with them. A useful experiment to carry out early in one's acquaintance with this device is the determination of the sign of the charge on the main dome. One of the principal delights provided by the generator is

/its.....

its ability to produce impressive sparks. This phenomenon will have to be discussed and explained; since pupils know notning of "potential difference" at tnis stage, the explanation must be given entirely in terms of electric forces.

Before we consider how this might best be done, we shall examine the question of the electrical "breakdown" of a gas more closely. Those called upon to explain the phenomenon would presumably like to be sure of the background, and some of the simple accounts to be found in standard textbooks on electricity differ somewhat from one another. First, let us consider how the breakdown of an insulating material is initiated. One's first inclination might be to suppose that breakdown starts when the electric field in the material is strong enough to pull electrons out of its "molecules". The following comparatively sophisticated description given by Scott (1959) seems to support this view: "At relatively high fields, the electrons in the filled [energy] bands... as well as the few available conduction electrons or ions, can gain enough energy by successive small quantum jumps or single large ones to be able to knock other charged particles loose and make them available for conduction." Corresponding statements will be found in more elementary textbooks. Cf. Rogers's (1960) description of "point-discharge": "A strong electric field near the point tears electrons off some atoms, leaving positively-charged atoms." However, as Goodlet (1955) remarks, "Since breakdown occurs at much lower field strengths than are necessary to tear electrons directly out of molecules, some other mechanism must be postulated." In ordinary air, tne few free electrons and ions that are needed to start off the "avalanche" discharge which constitutes breakdown, are always present as a result of cosmic rays, radioactivity of some kind or some other ionizing agent (cf Goodlet (op. cit.) and Chalmers (1957)).

/It.....

It is also possible that some of the "initiating" electrons are supplied by field emission from one of the "terminals" between which the discharge passes (Gledhill 1968). This can happen if the electric field is strong enough and if the terminal is negatively charged. "Appreciable" field emission can be obtained if the field strength at the terminal is of the order of 10^8 V/m (Bleaney and Bleaney 1957), and some field emission will take place at field strengths lower than this. Field strengths of this order of magnitude can be obtained if the terminal has sharp-pointed projections on it of radius 0.1 mm or less and potential differences of the order of 10kV are applied between the terminals.

As for the formation of the spark itself, there seems to be general agreement that the theory originally put forward by Townsend (1915) gives the correct account of the phenomenon. According to this theory, if at a given pressure of the gas the applied electric field is made sufficiently intense, the free electrons present will be able to gain enough energy from the field between collisions to be able to eject electrons from the molecules with which they collide, thereby releasing additional free electrons and producing additional ions. The newly-released free electrons can also be accelerated by the field and so produce ionization by collisions with other molecules, so that the number of electrons and ions present in the gas increases rapidly and the gas is readily able to conduct.

In the vicinity of a sharp point on a charged conductor, the intensity of the electric field will be very much greater than at greater distances away or near other parts of the conductor, so that it is possible for ionization by collision

/to.....

to take place near the point even though this does not occur further away. In this case one obtains what is known as a "point discharge", which is a steady leakage of charge away from the point, with ions of the same sign as the charge on the point being repelled vigorously away from it. If the charge on the point is positive, the free electrons and any negative ions that may be produced are drawn towards the point and tend to neutralize the charge there, while the positive ions are repelled away, the net effect being that of positive charge being carried away from the point itself. Should the charge on the point be negative, the description of what happens is much the same, with appropriate reversal of the signs of the ions mentioned, with one or two additional points of interest. The positive ions formed by collision will be accelerated strongly towards the point and those which strike the surface of the point can liberate further electrons. The phenomenon of the "electric wind" blowing away from the point is just as apparent when the point is negatively-charged as when it is positively-charged. This means that negative ions must play a prominent role in point discharge, since the negative electric wind can scarcely be a swiftly-blowing breeze of electrons. The formation of negative ions is seldom, if ever, mentioned in elementary discussions of gaseous discharges, so that it is worth elaborating on it a little here. The description given by Page and Adams (1931) of point discharge suggests one possible mechanism: "[The field] is seen to be very intense close to [the point]. Therefore air molecules in the vicinity of the point are strongly attracted to it. On reaching [it] they acquire some of the charge on the point and are then repelled". Neutral air molecules near the sharp point would be attracted towards it, because the molecules would acquire induced dipole moments under the action of the field if they were not already polar, and the field in the vicinity of the point is

/markedly.....

markedly non-uniform. If the point were negatively charged, it would be possible in principle for those molecules capable of forming negative ions by electron attachment to "acquire some of the charge on the point". (This would not happen to nitrogen molecules but would be possible for oxygen molecules.) In addition, there are several mechanisms by which negative ions can be formed in the surrounding gas by interaction between free electrons and the molecules of the gas (cf Massey 1950), such as capture of a free electron by an atom in the presence of a third body which takes away the excess energy, capture of a free electron by a molecule accompanied by vibrational excitation of the molecule and subsequent stabilization, and capture of a free electron with dissociation of the molecule.

There are sundry complexities connected with gaseous discharges to which we have not alluded in the above account, such as the role of the "space charge" in point discharge, but these need not concern the teacher interested in a "first-order" understanding of the above phenomena.

Turning to the problem of describing the phenomena of spark and point discharge in simple terms, we find that many textbooks are content with the following style of description: "The field in the immediate vicinity of the point... may be sufficient to ionize the air there. Those ions which have the same sign of charge as the conductor are violently repelled away from it, giving rise to an appreciable 'electric wind' " (Bennet 1965, 1968). This is correct as far as it goes, but it leaves out all mention of how the field near the point ionizes the air. We think that the mechanism of ionization should be discussed with pupils, so that they can be prompted to consider how it might be explained in terms of their present knowledge of electricity and the particle model of matter. A "first-order" idea of what happens can

/be.....

be conveyed to pupils in simple terms; there is nothing intrinsically difficult about the idea of electrons speeded up by strong electric forces being able to knock other electrons out of atoms or molecules, and pupils ought to be told about this important aspect of the formation of a discharge, as well as of the possible attachment of electrons to molecules to form negative ions. In these terms pupils should also be able to understand the ionization of air by particles emitted from radioactive sources.

Point discharge should be demonstrated to pupils, because of its intrinsic interest and because it plays a vital part in the action of the Van de Graaff generator. There are a number of ways of doing this: holding a sharp-pointed object in the hand and pointing it at the main dome of the generator when the latter is producing sparks - the sparks stop; using the phenomenon to produce the "electric wind"; taping a pin to the dome and aiming towards an uncharged gold-leaf electroscope placed about one or two feet away (Lewis 1964). An ion-electron "model" of the phenomenon can be described to pupils, who are now in a position to understand how charge can get from the belt of the generator to the inside of its dome. It will not be obvious to pupils that most of this charge must make its way to the outside of the dome. As reasons why some, at least, of the charge must go to the outside it can be pointed out that the dome is a conductor and that there are repulsive forces between like charges.

Many other useful points arise for discussion in connection with the Van de Graaff generator: the need to supply energy to the charges that are being transported to the dome, where this energy comes from and what becomes of it; the fact that the rate of sparking depends on the separation of the domes, and the less frequent the sparks, the fatter they

/are; .....

are; the fact that when charge is going from the main dome to a nearby conducting "point", it is leaking off in a steady stream; the fact that a neon discharge tube (assuming one is available) held along a radial line from the centre of the dome will glow continuously - these are only some that come to mind. Perhaps it would be a mistake to spend too much time in discussion of these points at this stage; but pupils should make their acquaintance with these ideas and phenomena and be given the opportunity of renewing it in due course.

(vii) The electric field

The Van de Graaff generator can also be used to introduce the idea of the electric field. One way of using it for this purpose is to fit an inverted paper "hula-skirt" over its dome, and switch on. The "hula-skirt" can be very simply made by attaching a number of narrow strips of light paper to a cardboard ring of suitable diameter. When the generator is running, the paper strips stretch themselves out along lines of the electric field of the dome. (The "hula-skirt" can assume the role of a paper crown which, when placed on a pupil's head, can be used to demonstrate the electric field in that vicinity when the pupil makes electrical contact with a Van de Graaff generator and is insulated from the floor.) This is the pupil's first encounter with an extremely important physical concept. For the moment, the idea of an electric field need mean no more to him than "the region around a charged object in which other objects can experience electrical forces." The naive demonstration with the Van de Graaff generator shows that there is a "pattern of directions" associated with this field. To show up such patterns much more vividly and beautifully, one could use the now-familiar technique, which may be said to have been popularized by the PSSC, of sprinkling the surface

/of.....

of an insulating liquid with elongated insulating objects such as grass or lettuce seeds, hair clippings or powdered semolina, mounting an electrode or electrodes of suitable shape in the liquid and applying a D.C. or A.C. voltage of the order of several kilovolts. It would be as well not to use this demonstration, beautiful and effective as it is, to introduce the idea of an electric field. If one does this, there will be no obvious connection for the pupil between what he has been doing in "electrostatics" and what he now sees happening. By using the Van de Graaff generator first, this important link can be established; and this demonstration then has the additional advantage of being able to show the pupil that a device like a 5-kv power supply or an H.T. Transformer can also "make" electric charge. At this stage, only the simplest forms of electric field (e.g. those due to a single point charge, equal and opposite point charges, equal and like point charges and equal and oppositely charged parallel plates) need be shown.

3.15 Current electricity: elementary treatment

3.15.1 Introduction

In the junior secondary school years, there is no need to carry the above treatment of elementary electrostatics any further. The pupil knows of the existence of two kinds of charge; he knows something about the forces between like and unlike charges and a little about electric fields. He has also been given a model with the aid of which he can interpret electrical phenomena. Later he will be able to elaborate this knowledge in a quantitative development of the subject, which he is not yet mature enough to be able to appreciate. There is, however, another aspect of electricity whose quantitative development he is mature enough to appreciate - up to a point. This is the behaviour of simple D.C. circuits. Children are certainly capable of

/acnieving.....

achieving a superficial but useful understanding of D.C. circuit theory at quite an early age. Thus the elder of the present writer's two sons once expressed an interest in some electrical experiments he came across in a book of the "Science is Fun" type. Some torch bulbs and 1½ volt cells were purchased for him, and the writer brought home some pieces of copper wire and a "multimeter". Very soon the young experimenter had, with only slight assistance, made certain fundamental discoveries about closed circuits containing single torch bulbs and cells. He then went on to make predictions about what would happen if two cells in series were used to light one bulb by itself or two bulbs in series, and so on, and in due course verified these predictions experimentally. He did all this some months before his seventh birthday. There is clearly a case for letting children twice this age carry out experiments of this type in the classroom.

3.15.2 <u>The introduction of the concepts of current and potential difference: some preliminary considerations</u>

While it is easy for children to carry out these experiments and to perceive the patterns into which the results fall, this should not be the limit of their acquaintance with D.C. circuits in the junior secondary school. They should see how these patterns fit into some scheme of explanation - a theory, in fact. This necessitates the introduction of the concepts of current, potential difference and resistance - and this is where the difficulties arise. In the past, these difficulties seem not always to have been properly appreciated by teachers and authors of textbooks. To illustrate the defects and dangers of an approach to current electricity in which proper attention is not paid to the logical

/development.....

development of the subject, we consider that presented by Gordon and Neser [no date]. In this approach, the sequence of presentation is as follows:-

(i) Survey the magnetic effects of a current, including the principle of the moving-coil galvanometer, without considering the question of how the reading of a galvanometer is to be related to the current in it.

(ii) Describe primary and secondary cells, introducing the concept of EMF as the "maximum electrical pressure" between the poles of a cell.

(iii) Introduce the concept of resistance by means of experiments in which wires of different lengths and materials are introduced in a circuit and the current (in amps!) is measured in each case. Define the unit of resistance in terms of a column of mercury of chosen length and cross-section.

(iv) Introduce the concept of potential difference as "electrical pressure difference" by means of the water analogy.

(v) Introduce Ohm's law by way of an experiment with a circuit containing a number of accumulators and a calibrated ammeter, stating the law in the form "The current in any conductor, at uniform temperature, is proportional to the potential difference between its ends."

(vi) Introduce the units of current and potential difference, defining the former in terms of the rate of silver deposition in a silver nitrate electrolytic cell, and the latter in terms of the ampere and the ohm.

(vii) Discuss measurement of current and potential difference, with special reference to the principles of converting galvanometers to ammeters or voltmeters.

(viii) Introduce the watt as the unit of power, pointing out that "a watt is the rate of working when a steady current of one ampere flows under a pressure of one volt." (The relevant section of the textbook contains the following remarkable

/statement.....

statement: "Since electrical energy can be converted into mechanical energy there is a definite relation between watts and horsepower." This gives the impression that the authors believe that the watt is essentially a unit of electrical power, and the horsepower essentially one of mechanical power).

The first defect of this approach to which we would draw attention is its introduction of the concept of electric current without any attempt to justify the use of the term "current" (unless we may regard the following sentence as a justification: "Everyone is familiar with the fact that an electric current may be generated by a battery or a dynamo."). Before the treatment has gone very far, the pupil is required to accept as natural the measurement of current in amperes, although no means has been suggested for calibrating the ammeter used in terms of amperes. The second, very serious defect is the manner in which the concept of potential difference is introduced: potential difference is linked vaguely to a discussion of differences in water level, but is not as yet given any quantitative significance in terms of power or current. The unit of potential difference is eventually introduced via resistance and Ohm's law, but the connection between potential difference, current and power is only made plain at a later stage. Finally, these concepts are linked together in the most casual and careless manner possible. There are other defects in the above approach, but the three which we have singled out are more than enough to condemn it. The first two go counter to what one would assume to be a fundamental principle of sound physics teaching, viz. to back up the introduction of a new concept with adequate discussion and careful attention to the problem of its measurement; and the third makes nonsense of any pretence that this is an attempt to teach a branch of the subject as though it were a coherent part of a science.

/The.....

The pupil will gain a very poor view of science if he is led to believe that this distorted sample of it is at all typical.

From this object lesson in how not to teach elementary current electricity, we may draw the moral that there are two main requirements to be satisfied in an adequate approach to the subject, viz.

(i) the new concepts, especially current and potential difference, must be introduced "with adequate discussion and careful attention to the problem of their measurement";

(ii) the relationship between these concepts must be as carefully treated as one would treat such relationships in the discussion of any physical theory. If, for example, the theory makes some use of certain concepts and ideas derived from areas of physics already familiar to the pupil, then this aspect of the matter should be given due prominence. The elementary theory of current electricity is one of the more elaborate theories the pupil meets in his junior secondary school years; he should find the experience stimulating and instructive, rather than bewildering and confusing.

To ensure that requirement (i) is met, it is necessary first of all to give an adequate introduction to <u>current</u>. This means that the pupil must be shown that there is a measurable quantity associated with a closed circuit which is the same everywhere in a simple series circuit, and that if the circuit is not a simple series one, it is still possible to state a simple law regarding this quantity, viz. that if at some point P a branch A of the circuit "divides into" a number of branches B, C and D, the value of this quantity in A is the same as the sum of its values in B, C and D. In other words, the pupil must be introduced to the concept of current by being shown that there is a quantity associated with a circuit which conforms to what would be called

/"Kirchhoff's.....

"Kirchhoff's First Rule" in more sophisticated circles. For this "measurable quantity" to conform to Kirchhoff's First Rule it must be measured in terms of either the magnetic or the chemical effects of the current. Nowadays it is generally agreed that it is sensible to work from the start in terms of the magnetic effect. From the purely practical point of view this is advantageous, because although one has to use a balance in a fundamental measurement of current whether one bases the measurement on magnetic or chemical effects, use of the chemical effect also requires a time measurement and the taking of a number of precautions in setting up the experiment properly, so that what should be essentially a simple experiment, quickly executed, would instead become a time-consuming affair with all sorts of complications attached to it. Use of the magnetic effect in this connection also has the advantage that it links up with the approved method of establishing current standards in standards laboratories and with the principle of operation of conventional ammeters. The fact that it would be impractical in an elementary course to base measurement of current on the internationally-agreed definition of the ampere is of no consequence; as Tricker (1955) puts it, "In spite of the fact that international committees have specified [the unit of current] by means of the mechanical force between two conductors there is fundamentally not the slightest necessity to introduce mechanical units at all."

To introduce potential difference adequately, it is necessary to establish as early as possible that the rate at which energy is transferred between a given pair of points in a circuit is proportional to the current and to another factor which is characteristic of the part of the circuit in question and which also depends on the number of cells used to supply energy to the circuit. If this is success-

/fully.....

fully done, then however hazy the pupil may feel about the underlying significance of potential difference, at least he can give it a precise quantitative meaning. The concept of resistance is less troublesome to a beginner than current or potential difference and there is thus no need to make any preliminary observations on its presentation here.

Requirement (ii) above will be met in part by introducing current and potential difference properly, since the introduction of these concepts naturally forms part of the development of the elementary theory of D.C. circuits. The explicit use of "certain concepts derived from areas of physics already familiar to the pupil" is necessary for the proper treatment of potential difference since, as we have seen, it is necessary here to make use of the concepts of energy and power. In addition, one can try to suggest a model for the observed facts concerning currents in circuits, a model based on the pupil's experience of electrostatics and on certain demonstrations that suggest that there is a connection between "electrostatics" and what happens in a current-carrying circuit. This can be done without dogmatic assertions about electron flow or the number of electrons in a coulomb; the point of this model-building is that it is a preliminary exercise in trying to "explain" certain electrical phenomena in electrical terms. The pupil is likely to benefit much more from this kind of discussion of the subject than by the more traditional one which confined itself strictly and frustratingly to the water analogy.

3.15.3 Current electricity in the classroom: where to start?

The above survey of the essential points to be made in the elementary teaching of current electricity might have given the impression that this material is too austere

/to.....

to be taught in the junior secondary school. This is not the view of the Nuffield Physics Project, whose plan for teaching electricity in Years II and III of their course follows essentially the lines suggested above. The Teachers' Guide to the work to be covered in these years contains a number of suggestions which, if followed, should help to make the subject palatable to pupils. We consider some of these ideas below, together with some others that have been put forward in recent years.

The first decision to be made concerning the teaching of current electricity is where to start. If the subject is taught at first without reference to electric charge, then it has to be developed on the basis of the magnetic and heating effects exhibited by a wire when connected to certain kinds of "black box". The idea that some kind of current is present in the wire can be built up with the aid of the analogies with the flow of water and traffic. These analogies, particularly that of the "water circuit", should not be overstressed, otherwise the pupil might find difficulty later in putting his idea of "potential difference" as "electrical pressure" in the right perspective. On the other hand, to attempt to lead straight into a study of electric currents from elementary electrostatics raises the difficulty of convincing pupils at the outset that charge is actually flowing round a circuit containing, say, a cell, a torch bulb and some pieces of connecting wire. It is possible, we think, to present evidence that at least points in this direction, but the accompanying arguments would be better understood at a later stage. There is always the alternative choice of simply telling pupils that current is a flow of charge (cf Brommert 1968), but this kind of dogmatism is scarcely to be encouraged in the teaching of science. On the whole it seems better to accept the kind of compromise approach employed in the Nuffield and new South African junior secondary school syllabuses, in which the pupils first study

/electrostatics.....

electrostatics and current electricity as essentially independent subjects, and are later shown that an electric current is actually a flow of charge.

If the approach to current outlined above were strictly followed, pupils would begin their study of currents by carrying out some experiments showing them that magnetic effects are associated with a closed D.C. circuit. Then it could be shown that by connecting more cells in series into the circuit, "bigger" magnetic effects are produced; and the teacher could argue as follows:- "When we close the switch, the circuit shows magnetic effects which are not there when the switch is open. When we increase the number of cells in the circuit, these magnetic effects are more obvious. With the switch closed, there seems to be something in the circuit whose presence we can tell by the magnetic effects it produces. Putting more cells into the circuit makes this 'something' bigger and stronger. We can measure some of these magnetic effects quite easily" - here the teacher could give examples - "so perhaps we could take one or other of these measurements as an indicator of the size of this 'something'". From this point on, experiments with a current balance and with ammeters calibrated with a current balance could be carried out to show that this "something" conformed to Kirchhoff's First Rule. All this is adequately logical, down to its studied avoidance of the

/word.....

word "current" (cf Tricker (1955): "... in the beginning this term is best avoided because of its prejudicial associations"). Sooner or later this word will have to be introduced, although initially it can simply mean "that which the circuit has which gives rise to magnetic effects"; there would be no more special significance attached to its use at this stage than there is to the use of the word "charge" for "that property which a plastic ruler acquires when it is rubbed with a handkerchief".

To introduce the subject in this way would have the merit of moving quickly to the desired goal. However, it is at present fashionable to require pupils first to play with cells and torch bulbs and so to discover the results of series and parallel connection of cells and resistors. This is how the subject is introduced in the Nuffield Physics Project, and also now the new junior secondary school syllabuses in South Africa require it to be introduced. The Nuffield Guide (II) quotes glowing testimonials from teachers to the enjoyment and excitement experienced by pupils in conducting these experiments and, as noted above, the present author's own son has likewise enjoyed himself. Perhaps, then, it should be agreed that while a satisfactory treatment must ultimately follow the lines suggested above, it can be "humanized" by being introduced in the way laid down by the Nuffield Project. But if pupils are to begin their study of electricity by direct experimentation with torch cells and lamps, the introduction of the concept of "current" at an early stage cannot really be avoided. This means that pupils will meet "current" first as the "something" that is "in" the torch bulb that makes it glow; i.e. they meet current through its heating rather than through its magnetic effects. There is, we feel, no real objection

/to.....

to this provided that when the measurement of current is taken up, it is done properly through the magnetic effect. Pupils should not be confused by the fact that when a circuit is closed, there is something in it which can cause both heating and magnetic effects; and the fact that one rather than the other of these effects is taken as the means of measuring this "something" need not cause them any alarm.

One can agree with Tricker that the word "current" has "prejudicial associations" which it might be best to avoid. On the other hand, the junior secondary school pupil's understanding of the word "current" outside the context of electricity might not yet be so fully developed as to beg too many questions from him when the word is first used in connection with electricity. It would certainly make matters easier for the teacher if he could use "current" from the start, and perhaps he should, provided that he does not make the mistake of using the term too literally at first. A possible approach for him is to say that "'current' is the name we give to whatever is 'in' the torch bulb that makes it glow", and that "the brighter the lamp glows, the bigger the current." He need make no attempt for the moment to suggest that something might be flowing in the circuit, and indeed there would as yet be no evidence for such a flow available for him to use. An alternative approach which might be held to "spoil the fun" somewhat but which does try to make the idea a little more meaningful, is to say "Let us try to invent some kind of model of what is going on in a closed circuit. Suppose we make the guess that the lamp glows because something is flowing through it and through the rest of the circuit of which it forms a part. How could we test a guess like this?" As Gledhill (1968) has suggested, perhaps the most obvious response here is to suggest that a wire in the circuit should be cut to see if anything drips out; the experimental fact that nothing drips out of

/the.....

the wire whether it is cut before or after insertion into a current-carrying circuit would suggest that the current in the wire is not to be identified with anything that is macroscopic and obvious. So pupils and teacher could continue the pursuit of a model for current, possibly by a discussion of what means by a current in a river and of how one could measure it, of what would be meant by a current in a closed water circuit and of how the "current" might vary round such a closed circuit. Here one could let pupils do what they might already have thought of doing, namely to try connecting up a simple series circuit with the same components arranged in different ways. In a circuit containing a pair of cells and a pair of similar lamps all connected in series, the lamps glow with the same brightness no matter in what order the components are arranged round the circuit. This can be taken as an indication that the "current", whatever it is, is the same everywhere in the circuit; and this provides support, though not conclusive evidence, for the view that there is indeed something flowing round the circuit. This idea can be pursued further by connecting up a circuit containing at least three lamps, two of which are connected in parallel; the latter two glow with equal brightness, but not as brightly as the other lamp. The brightness of the lamp is much too crude a measure of "current" to permit the conclusion being drawn that the current in the third lamp is the sum of the currents in the other two; but the evidence is at any rate not inconsistent with that interpretation, which could be tentatively suggested by the teacher, if members of the class have not already offered it.

3.15.4 The measurement of current

At this point it is clearly appropriate for pupils to embark on an investigation of the magnetic effects of

/currents.....

currents, leading on to the measurement of current by this means. Unfortunately the new South African syllabus for junior secondary schools requires the ammeter to be introduced as a current-measuring instrument before the pupils have studied the magnetic effects of an electric current. In an experiment suggested in the Orange Free State Teachers' Guide for this syllabus, the ammeter is to be successively connected in series with a single lamp and one, two and three cells respectively; the pupil is expected to note the correspondence between the reading of the ammeter and the brightness of the lamp. The logic of the experiment goes something like this: "We have previously taken the brightness of the lamp as a rough measure of the strength of the current; now we see that when the lamp is brighter, the reading on the ammeter is bigger; so let us take the reading on the ammeter as a measure of the current. This is a sensible thing to do, because using the ammeter in this way means that we can express the strength of the current by means of a number on a scale, and this is the sort of thing we try to do when we measure something." The weak point of this procedure is not just the use of a black box - in this case the ammeter - to measure current; as we shall see, this can be justified. What does seem unsatisfactory is that this black box is not calibrated in terms of some measurable effect of the current. This kind of approach is bound to have pupils asking themselves or the teacher, "How do I know that this instrument [the ammeter] measures 'current', when I don't really know what 'current' is and I have no idea how the instrument works?" Certainly pupils can learn how to use ammeters and go on to work out "Ohm's Law" or "energy dissipation" problems in which ammeter readings have to be substituted into formulae; but if their state of knowledge of

/currents.....

currents and ammeters is as indicated by the above imaginary question, have they really understood anything about the physics of current electricity?

Obviously these difficulties can be removed if the pupil does not meet the ammeter until after he has observed some of the magnetic effects of electric currents. Now that he knows about both heating and magnetic effects of currents, he can be led to think more deeply about how a current can be measured. This raises the very important principle of measuring a current (whose true nature is still unknown) in terms of the effect it produces. For the pupil, there can be no obvious reason apart from practical convenience for the choice of the magnetic instead of the heating effect of a current as the basis for the measurement of current. Only later will he learn that "current" as measured in this way is truly rate of flow of charge. But this does not matter at this stage. What <u>is</u> important is that the magnetic effect does provide a means of "putting the numbers in" which is simple and direct.

The fundamental method of measuring current in terms of its magnetic effect requires the use of a current balance. For preliminary work, it is desirable to use the kind of current balance in which the force measured with the balance is directly proportional to the current. The pupil is already familiar through his experience of temperature measurement with the principle of taking the change in the value of a certain property to be proportional to the change in the quantity for which a scale is being set up, so that he should be able to accept the application of this principle to the measurement of current with a current balance. Operationally, that is to say, "current" is going to be defined for him as being proportional or equal to the reading on the balance. But if the balance is to give readings of current

/which.....

which will lead to Kirchhoff's First Rule and to calibration of moving-coil ammeters which will give linear scales to the latter, it must give readings which are proportional to the "current" as ordinarily defined. This requirement would not be met by a parallel-wire type of current balance (Williamson 1958) or by that used in the PSSC course (PSSC Lab Guide and Teachers' Resource Book and Guide, Book Four), in both of which the measured force is proportional to the square of the current. Current balances in the latter category are more "fundamental", in the sense that they measure the forces of interaction between current-carrying conductors and so come closer to the approved method of establishing standards of current, whereas "linear" current balances measure the force of interaction between a current-carrying conductor and a permanent magnet (or, less satisfactorily, another current-carrying conductor in which an independent current is maintained at a set value). But considerations of this kind need not concern the pupil. From his point of view it would be perfectly legitimate to set up a scale of current based on a "linear" current balance.

In the Nuffield course, the type of balance used is of the "moving-magnet" type; a small magnet hangs from one end of the balance arm inside a coil, so that when the current is turned on in the latter, the magnet is pulled into the coil and a counter-force must be applied to the other end of the beam to balance it. For a given position of the magnet inside the coil, the force on it is directly proportional to the current; and as in the operation of the balance, the magnet would always be brought to the same position for equilibrium, we see that this type of balance meets the requirement of linearity. Other balances which would also do this are of the force-on-the-conductor type, in which a current-carrying conductor on which the force is measured is

/suspended.....

suspended in the field of a permanent magnet. Current balances have unfortunately not been specified for the corresponding South African syllabus; if the course is one day modified to include them, one would hope that the syllabus planners would follow the example of the Nuffield course and specify that the pupils themselves are to assemble the balances. Then there would be nothing of the black-box about these fundamental current-measuring instruments, at least as far as their construction was concerned.

In early work with the balance, pupils could take readings corresponding to currents previously recognized by them to be equivalent to "full" lighting of one torch bulb or lamp, or greater or less than this. They could also measure the current in two or three "fully lit" lamps connected in parallel and compare it with that in a single "fully-lit" lamp; and they could use the balance to show that the current is the same all round the circuit. These latter measurements yield results which are consistent with the idea that electric current is something flowing round the circuit, and if this idea has not yet been introduced, it might well be introduced here.

The pupil is now well-prepared to meet the familiar form of ammeter. In accordance with the principles of the satisfactory use of black boxes, the ammeter should first be connected in series with the current balance in a complete circuit and shown to give the same readings - or, if the current balance is not calibrated in amperes but in arbitrary units instead, the ammeter and current balance readings can be shown to be proportional. Use of the ampere as the unit in which the ammeter is calibrated is not something about which a great fuss need be made at this stage. Talk of the coulomb and of a rate of flow of 6.25×10^{18} electrons per

/second.....

second would serve no good purpose. We know very well that the definition of the ampere is a matter of arbitrary choice, and there is no reason why we should not "explain" the ampere by saying that physicists decided long ago to call the current that produced a certain size of magnetic effect an ampere; and that in our current balances, this definition of the ampere amounts to saying that a force on the conductor of [x] newtons means that the current in the conductor is 1 ampere.

The principle of operation of the ammeter is not something that is vitally necessary for pupils to know about as yet, though teachers wanting to make the black box a little greyer in appearance could easily do so by opening up the instrument and revealing the mechanism. It will be evident then that the (moving-coil) ammeter works on the magnetic effect of the current. If pupils know enough about magnetic effects of a current, they will be able to understand that when there is a current in the coil, forces on its sides will tend to turn it round, while its turning motion will be restrained by the hairsprings.

3.15.5 Summary of requirements for a satisfactory introduction to current electricity

Before we consider how the subject of current electricity might be further developed, we shall pause to sum up what would seem to be the requirements for a satisfactory introduction to this subject.

(i) Background. It is implied in the foregoing discussion that if the measurement of current is to be properly introduced, it will have to be done on the basis of magnetic effects - in particular, on the use of a suitable kind of current balance. This in turn implies that pupils must be familiar with the rudiments of magnetism, viz. what is meant

/by.....

by a magnet, magnetic poles and the idea of a magnetic field. We take up these points in a separate section.

(ii) Treatment. (a) Pupils can first become familiar with the idea of an electric current through experimentation with simple circuits containing torch cells, pieces of wire and torch bulbs or lamps. To begin with, the current is "what makes the lamp light up". As a first, crude measure of this current, the brightness of a lamp is used.

(b) As a preliminary to proper measurement of current, the pupil studies the magnetic effects of an electric current. These should include the making of a simple electromagnet (e.g. by wrapping a piece of insulated copper wire around a nail), and the force on a current-carrying conductor in a magnetic field. It would not be necessary at this point to show the magnetic field patterns of current-carrying wires or solenoids; the main theme for the moment is the measurement of current, and it would be well to stick fairly close to it. Then the principle of the measurement of current by its effects can be discussed, leading to measurement of current by current balance and "calibration" of an ammeter by comparison with the current balance.

The variety of experimentation in all this should be enough to maintain interest on the part of the pupils, who are at the same time given a treatment of the subject which has no awkward gaps of logic in it.

(Note: It will be clear from the foregoing that the cells used in these experiments are treated as black boxes throughout. At the moment, they are simply "the things we seem to have to use in circuits in order to obtain electric currents". On this point, see also our earlier remarks on cells in Section 2.4).

/3.15.6 .....

3.15.6 Elementary treatment of resistance

It is easy and convenient to introduce "resistance" qualitatively after the techniques of current measurement have been studied. (It can also be introduced before the use of the ammeter has been treated, the brightness of a lamp being treated as a "rough ammeter", but it would be neater to use genuine ammeters for this work.) An intuitive pre-quantitative grasp of "resistance" can be taught with the aid of experiments in which conductors of different kinds are inserted in a simple series circuit which includes an ammeter. These experiments would show that there is a property associated with any conductor which helps to determine what the current will be in the circuit in which the conductor is connected. Of course there is nothing to stop one saying, "I am going to define this property in such a way that its value will be bigger when the current is bigger", in which case one would be talking about conductance. But tradition has accustomed us to dealing with resistance instead, at least in elementary treatments, and so we think of our conductors having bigger resistance when the current is smaller. Use of the word "resistance" points strongly towards the idea of "something flowing" in the circuit, but as this should already have been emphasized in the discussion of the experiments with the ammeter, it should cause no trouble. In the course of qualitative comparison of the resistance of different things connected in the circuit, the teacher could remind pupils of the difference between "conductors" and "insulators" as noted in electrostatics. The experiments with circuits will show that metals have comparatively small resistance, while strips of plastic material reduce the current to zero. This is not conclusive evidence that one is dealing with the same kind of thing in the two situations - an experiment with an ordinary ammeter cannot

/disclose.....

disclose the huge difference in order of magnitude between the resistance of a metal conductor and that of an insulator of similar dimensions - but it is suggestive, all the same. And the pupil knows that the material with which copper wire used in the laboratory and in the home is covered is called "insulation"; he is bound to wonder, "insulation against what?"

The use of a variable resistance or rheostat in a circuit can perhaps be introduced to pupils by way of a discussion of how one might adjust the current in a circuit to some desired value. Pupils could make their own crude rheostats using lengths of bare wire with crocodile clips as sliding contacts, and then progress to the real thing. This is also a means of starting the treatment of the effective resistance of a pair of conductors connected in series: the resistance of a piece of wire of given material depends on how long it is, so if we connect together two pieces of wire in series, we would expect that the resistance of the two together is bigger than the resistance of either.

Experiments on resistances in series and in parallel can be carried out with lengths of "resistance"-wire and with lamps and so provide a basis for the quantitative conclusions to be used later. The results of these experiments should also be interpreted in terms of the "flow" model of electric current. It will also be instructive for pupils to compare the resistances of conductors of similar length and material but having different diameters. It would be as well in the course of this study of resistance to repeat the experiment showing that the current is the same everywhere in a simple series circuit. Although the point has been made earlier, it should perhaps be reinforced in case the idea of "resistance" has prompted some pupils to think that there is less current going out of than is coming into a "resistance."

/3.15.7.....

3.15.7 Potential difference

We come now to that part of the subject which is likely to present most difficulty to pupils and teachers alike, viz. the introduction of "potential difference". It is a comparatively simple matter to familiarize pupils with the way in which a voltmeter is used in a circuit; it is another matter altogether to get them to understand what the voltmeter really measures. The teacher should take the latter as his main aim, even if he feels it may not be completely realized in a junior secondary school course. This means that he will try to convince his pupils that a voltmeter measures energy transfer per unit charge, and that as far as circuits are concerned, "energy transfer per unit charge" may be taken to be the definition of potential difference. If he does this, he should be able to safeguard himself against unthinking use of the term "electrical pressure" in teaching potential difference (see our earlier reference to this matter in Section 2.2).

It is obvious that if potential difference is to be taught in this way, the pupil will not only have to be conversant with the "flow" model of an electric current, but will also have to accept that the quantity measured as current is actually rate of flow of charge. This does not mean that he has to be convinced that the "charge" flowing round in a circuit is the same kind of thing as the "charge" whose effects he observed in electrostatics experiments. What he will really have to accept is that "something" flows in a circuit (he might as well call the "something" charge) and that 1 ampere is actually the rate of flow of 1 unit of this something (he might as well call the unit the coulomb) per second.

/3.15.8.....

3.15.8 Digression: a sophisticated approach to potential difference via the water analogy

Any method of treating the topic within the present context is going to have to make use of the water analogy to some or other extent. Although it will probably be thought too difficult for the average pupil, we think it of interest to give here an outline of the most sophisticated treatment of potential difference by means of this analogy, in a form set out in part by Rogers (1960). The treatment builds on the (by this time) well-known fact that in elements of electric circuits, energy is transferred. Accordingly, one uses the concept of energy in discussing what happens in a "water circuit". Suppose that part of such a circuit consists of a long, thin, uniform pipe. The pressure decreases uniformly along the pipe in the direction of flow. (It is easily demonstrated that we have to establish a difference in pressure between the ends of the pipe in order to make water flow in it.) Application of elementary mechanics shows that a difference in pressure between the ends of such a pipe implies that energy conversion must be taking place in the pipe, as work is being done on the steadily flowing liquid. No extra kinetic energy is gained by the liquid, so that the energy transferred to the liquid as a result of work done all appears as heat. (This presupposes conservation of energy. See our later comments in Section 3.15.9 on the subject of voltmeter calibration.) Here we have an interesting correspondence between a pipe in a water circuit and a conductor in an electrical circuit; in the pipe water flows steadily and energy is dissipated as heat; in the conductor, we think that there is something called "charge" flowing steadily, and energy is dissipated in the conductor as heat. Let us take the analogy a step further; we have a pressure difference between the ends of the pipe,

/so.....

so perhaps we can speak of a difference of "electrical pressure" between the ends of the conductor. This helps us very little unless we can specify a means of measuring "electrical pressure". Indeed it can be argued that we could have got to the idea of electrical pressure much more quickly if we had wanted to; many faulty textbooks (e.g Gordon and Neser) show us how! But the water analogy does provide a useful clue as to how to proceed: experiment shows that the rate at which heat is developed in a pipe is proportional to the pressure difference between its ends and to the volume rate of flow, so that the pressure difference is proportional to

$$\frac{\text{energy liberated as heat per unit time}}{\text{volume flowing past a given cross-section per unit time}}$$

(In fact it is equal to this ratio, as can easily be shown.) Perhaps "electrical pressure difference" between the ends of a conductor can be defined as the ratio

$$\frac{\text{heat developed per unit time}}{\text{current}},$$

where the current in amperes is now to be interpreted as the rate of flow of charge in "coulombs per second" past a given cross-section of the conductor.

We still have to show that the actual behaviour of electrical circuits makes this definition meaningful. One way of doing this is to perform an experiment described by Tricker (1955). As specified in the original description, the apparatus required consists of nine six-volt car lamps, three six-volt car batteries and an ammeter capable of measuring currents of up to 12 amps; it is used in the following way:-

(a) Connect one lamp in series with the ammeter and <u>one</u> battery, thereby fully lighting the lamp, and note the corresponding current.

/(b).....

(b) Connect two lamps in series and show that they are only fully lit if the current is supplied by two batteries connected in series, and that the current is the same as in (a).

(c) Show that three lamps in series are fully lit by three batteries in series, the current being the same as in (a).

(d) Connect two lamps in parallel - one battery is enough to light them and the total current is twice the former value.

(e) Proceed as above with series-parallel combinations of the lamps, showing that in all cases the number of lamps fully lit is given by the product of the number of batteries and the number of "units" of current flowing, a "unit" being the current registered in (a).

If the units of power and "electrical pressure difference" are temporarily taken to be the power dissipated in one fully-lit lamp and the "electrical pressure difference" set up by one battery respectively, the result of the experiment can be summed up in the statement

Total power dissipated = "electrical pressure difference" times "current".

The term "electrical pressure difference" can and should now be dropped, the correct term "potential difference" substituted and officially defined as energy dissipated per unit charge or power dissipated per unit current. The volt can then be introduced and defined as a joule per coulomb.

As we have implied, the reasoning in the first part of this sophisticated treatment of "electrical pressure" is probably too advanced for a junior secondary school pupil. However it would be helpful for the teacher to keep it in mind, because it does make clear the origin of "electrical pressure" as a quantitative concept. The development of

/the.....

the idea in this way should make clear that the word "pressure" used in this context carries with it none of its familiar connotations; in particular, it should on no account be thought of as "electrical force per unit area" which, in so far as we can give a meaning to it, would be the same anywhere along a uniform current-carrying conductor. On this subject the Nuffield Guide (III) comments that "there would be no harm in thinking of [p.d.] as shorthand for 'pressure difference'", and goes on: "We may say that voltage or potential difference is the equivalent, for an electrical circuit, of pressure difference for a water circuit. Some of us enjoy thinking of p.d. as a real pressure difference, of something pushing quite hard; of something which, pushing hard enough, hurts us by giving us a bad shock." Those of us who think that we do understand what "p.d." is, may indeed find it amusing to think about it in the manner just described. But it may be questioned whether a pupil, trying desperately to grasp something entirely new to him, would find this way of thinking equally enjoyable. It is quite possible that he may be seriously troubled by it, that he might grope vainly for some explanation of this supposed electrical "push". Did he but know it, there can be no such explanation, because there is no electrical "push" - at any rate, not in the sense implied by careless use of the term "electrical pressure". The present writer vividly remembers being troubled by considerations of this sort at school; in consequence, he never felt happy about potential difference until the subject was taught him at university. When much later in life he looked again at his school textbook, he realized that neither he nor any other pupil could have been expected to understand potential difference when the topic was presented in this manner. Perhaps he cannot now be blamed for doubting whether "there

/would.....

would be no harm in thinking of p.d. as pressure difference."

3.15.9 Potential difference as energy transferred per unit charge; the "voltmeter" approach

We return to the task of trying to describe a simple approach to potential difference which aims at revealing its significance as energy transferred per unit charge. This approach should also show that a voltmeter measures "potential difference" as defined in this way. One way of beginning is to focus attention on the voltmeter, which is initially a black box whose function is to be revealed by experiment. Its correct mode of connection into an already established circuit could be discovered by the pupil by trial and error. Suppose that the "already established circuit" consists of a battery, three lamps and an ammeter, all connected in series; the voltage of the battery and the wattages of the lamps should be chosen so that the lamps will glow with their normal brightness in the circuit. Connection of the (moving coil) voltmeter in series into this circuit would produce a reading on the voltmeter and a reduction of the current in the circuit to a value too small to be read on the ammeter. The pupil must now solve the problem of how to connect the voltmeter into the circuit in such a way that it does not destroy the conditions existing before its connection. Eventually he should discover (or be guided towards discovering) that when the voltmeter is connected in parallel across, say, one or more of the lamps, it does give a reading without seeming to upset existing conditions in the circuit. Further investigation would show that the voltmeter gives a bigger reading when connected across two lamps than when connected across one only; that it gives the same reading when connected in turn across each of the cells composing the battery, and that if

/it.....

it is connected across different numbers of cells, its reading is proportional to the number of cells across which it is connected.

The pupil now has accumulated quite a lot of evidence about voltmeter readings for the circuit under study. In interpreting it in the correct way, he will have to be firmly guided by the teacher, who must now ask the pupil to think about energy transfer in electric circuits. In the circuit just described, attention could be directed to the energy transfer going on in each of the lamps. It would seem reasonable to take as a fundamental assumption that the total rate of energy dissipation in n lamps is n times the rate of energy dissipation in one lamp. Experiment has shown that the voltmeter reading across n lamps is n times the voltmeter reading across one lamp. Evidently the voltmeter reading is a measure of how much energy is being dissipated in the portion of the circuit across which the voltmeter is connected. That it is not a measure of this alone is easily shown by a modification of the above experiment with the lamps. Let the lamps now be connected in parallel with one another and in series with the ammeter and the battery, and the voltmeter be connected across the lamps. Now the reading of the voltmeter is the same as it was across one of the lamps in the previous experiment, while the reading of the ammeter is three times its previous value. Since we still have three lamps glowing with their normal brightness, the total rate of energy dissipation by the lamps is the same as before. By extending the experiment still further to include other series-parallel combinations of lamps, with the number of cells composing the battery adjusted to suit, it can be established as a general result that the rate of energy dissipation in a given portion of a circuit is proportional to the voltmeter reading across that

/part.....

part of the circuit multiplied by the total current. This proportionality relation can then be turned round to read "voltmeter reading $\propto \frac{\text{rate of energy dissipation}}{\text{total current}}$". Using the idea that current is "rate of flow of 'charge'" and that "rate of energy dissipation" is what it says it is, we can also write this relation as "voltmeter reading $\propto \frac{\text{energy given up}}{\text{charge}}$" or "voltmeter reading $\propto$ energy given up per unit charge." The next step is to state that the voltmeter is in fact calibrated in such a way that its reading is equal to the number of joules given up per coulomb, and that the volt is the name given to the unit "joule per coulomb"- hence the name "voltmeter" for the instrument and "voltage" as the common name for the quantity measured by the voltmeter. The official name for "voltage" must be introduced, together with some justification for its form. Here one can say that the energy that charge gives up as it flows through the circuit is a form of potential energy, and the term "potential difference" implies that the charge would have more of this kind of potential energy at one end of a circuit/than at the other. element

It will be recognized that the experiments just described are essentially those described by Tricker (1955) approached from a slightly different viewpoint. The major assumption of such experiments is that energy is being "given up" by charge flowing round the circuit; what is observed is that the lamps glow, and it is taken that the energy which the lamps are getting rid of is the energy that was supplied to them from the flowing charge. This is an assumption of a kind of energy conservation; yet it is not full-scale energy conservation, since the interpretation of the experiment does not depend on the assumption that all of the "electrical" energy lost by the charge appears as energy given out by the lamps. On the other hand, any proper method of

/calibrating.....

calibrating a voltmeter directly in terms of energy transfer per unit charge does depend on the validity of the Principle of Conservation of Energy; if we want to know the number of joules transferred by a heating coil, for example, we can only do this by measuring the "heat" output from the coil in joules, and that measurement must depend on an assumption of the value of the relation between the joule and the calorie. The establishment of that relation is itself a proof that conservation of energy can be extended to embrace "heat" as well as mechanical energy. Naturally we cannot for the purposes of a "fundamental" calibration of a voltmeter rely on a value for "J" that has been obtained from electrical measurements; only a mechanically determined value for "J" is permissible here. (On this point, cf. the Nuffield Guide III and IV and Tricker (1955).)

The fact that we have to rely on conservation of energy in calibrating our voltmeters may be held to be a serious objection to the above approach to voltage in the context of the South African junior secondary school syllabuses. In those syllabuses, conservation of energy is not to be treated as fully established. However, it will certainly not be possible in this introductory treatment to perform a calibration of the voltmeter. In the absence of the crucial experiment itself, there is no pressing need to explain its logic! Instead, refuge can be taken in the following account of the matter: "We see from experiment that it is reasonable to claim that the reading on a voltmeter is a measure of energy transferred per unit charge. But to prove that a reading of, say, 2 volts on the voltmeter does really mean that 2 joules of energy is transferred for every coulomb of charge, needs an experiment that we shall only really be able to understand in the senior years. You may take it that the voltmeters we use have been or could be calibrated like

/this,.....

this, so that their readings tell you directly how much energy is transferred per coulomb." It seems a pity to have to fall back on "teaching by assertion" at this point; yet the Nuffield Guide (IV) adopts much the same attitude in connection with this matter, even though Nuffield pupils are expected to know about energy conservation before they deal with volts and voltmeters. In the Nuffield course, calibration of a voltmeter is designated as an "Advanced buffer option", to be done only by fast groups of pupils who have teachers who are "skilful at running this as a class experiment". For the rest, the Nuffield Guide (IV) says "This is a place where we should ask many pupils to take our word for the instrument's action - as we do with stopwatches, balances and (quite often) ammeters." We cannot resist quoting this Guide's comments on "teaching by assertion"; according to it, "Teaching by assertion is a weak method; and if we indulge in it much we shall give science a poor name. Yet occasionally we may save pupils from confusing arguments for which they are not ready; and then we may be wise to make assertions. These are not so damaging if we ourselves remember clearly that we have made them." Nor, we might add, if our assertions are based on facts and principles which we can make a reasonable claim to have understood. In sum, we would suggest that if potential difference and the function of a voltmeter are not to be taught in terms of energy transfer per unit charge (with the minor drawback of having to indulge in "teaching by assertion" at the end), they had best not be taught at all, as the disadvantages of any other approaches far outweigh any alleged advantages.

3.15.10 Potential difference as energy transferred per unit charge: direct approach

As an alternative to the line of attack on "voltage" just described, in which one begins with the voltmeter,

/one.....

one may try going straight to the root of the matter by talking about energy transfer in circuits right from the start. (This is the approach described in the Nuffield Guide (IV)). Two lamps of different wattage connected in series in the same circuit show obviously different rates of energy transfer for the same current; and from this it is concluded that if we want to know how much energy is transferred in a given time in a given circuit element, we will have to know both the current and the energy transferred per unit charge in the element. A convenient instrument is needed to measure the latter; by means of experiments of the kind already described, it is shown that the voltmeter meets this requirement.

It can be argued that a defect of both these approaches to "voltage" is that at some point in the treatment, a black box is produced which happens to have just the right characteristics in order to measure energy transfer per unit charge. No doubt it <u>is</u> a defect: in an ideal physics course in an ideal worldwe would have our pupils carefully consider tne problem of how to measure the desired quantity and eventually, with a little assistance from us, they would arrive at the principle of a practical voltmeter. But this is not an ideal world, and it seems that no courses in science can ever be ideal; so perhaps we can rest content with the fact that, after all, our black box is not just taken for granted - its mode of use and its function are carefully established by experiment.

3.15.11 The "voltage" of cells

An elementary treatment of potential difference would scarcely be complete unless it made some reference to the function of cells or batteries in circuits. The experiments carried out by the pupils show that the voltmeter

/gives.....

gives a reading proportional to the number of cells across which it is connected; in other words, to quote Tricker (1955) and the Nuffield Guide (III), a voltmeter can be used as a "cell counter". This is interesting, but not specially valuable unless one can explain some of the underlying physics. To do this requires some more "assertion" from the teacher, who would have to present the case somewhat along these lines: "Take a circuit consisting of a lamp and a piece of wire joining its terminals together; nothing happens in it. Alter the circuit to include the kind of black box known as a 'cell', and the lamp glows. We think that in the lamp, electrical energy is being changed into radiant energy; it is a reasonable guess that the charge that gives up electrical energy when it flows through the lamp, gains this kind of energy when it flows through the cell. We have seen simple forms of cell in which we have been aware of chemical reactions going on; again, it seems a reasonable guess that in the cell, the charge gains electrical energy at the expense of chemical energy. Perhaps the voltmeter, when it is connected across a cell, tells us the electrical energy gained for each coulomb of charge passing through the cell." Behind this argument, too, lies the assumption if not of "full-scale" conservation of energy, then at least of energy "disappearing" in one form and "appearing" to some extent in another form. We feel that it is unnecessary at this stage to mention explicitly the concept of EMF. To mention the term "electromotive force" here would do much more harm than good. If pupils can accept the idea that a cell is something which makes electrical energy available from chemical energy, they are off to a good start in understanding cells; and as beginners in the theoretical study of circuits, it would be much better for them to get used to remembering that the label "$1\frac{1}{2}$ volts" on a cell means

/that.....

that the cell makes available $1\frac{1}{2}$ joules of electrical energy to every coulomb of charge flowing through it, than for them to make the simpler correspondence "That means the cell has an EMF . of $1\frac{1}{2}$ volts". The term EMF . can be saved up for the senior years, where it will be most useful as a general term for the rate of energy conversion from non-electrical to electrical forms per unit charge flowing through a "seat" or "source" of EMF. Here it is neither an aid to understanding nor a particularly convenient abbreviation.

3.15.12 The use of analogy in teaching potential difference

It will be noticed that we have scrupulously avoided any reference to "electrical pressure" in outlining the above teaching approaches to potential difference. We did this in order to demonstrate that there is no absolute necessity to use this term in this context; nor do we have to invoke the water analogy, once the idea of "flow of charge" is accepted. Many teachers will nevertheless want to use this analogy as a teaching aid in presenting the concept of potential difference. They will especially want to make use of the model water circuit (cf. Nuffield Guide (III)) with its flowmeter and pressure gauge as analogues for the ammeter and voltmeter respectively; the flowmeter is connected "in series" with the circuit and the pressure gauge "in parallel" with that part across which the pressure difference is to be measured. There can be no objection to this, provided that the teacher clearly recognizes the limitations of the analogy and makes them plain to the pupils. Then the analogy can be a useful means of helping to drive home some points in the argument and as an aid to remembering some of those points. The limitations of the water analogy particularly concern the correspondence between

/"pressure.....

"pressure difference" and "potential difference"; we have already drawn attention to tne dangers inherent in careless use of the term "electrical pressure", and we feel it is the duty of the teacher to insist that this term is a mere synonym for potential difference, in which the word "pressure" has altogether lost its usual meaning. In the senior years, when the pupil knows more about charge and about electric fields, he can be taught another and much better analogy for the behaviour of charges in circuits. (See Section 7.13.4) In the meantime, the teacher who wants to use an analogy will have to do the best he can with the water-circuit, and remind himself constantly that the use of an analogy is intended to promote understanding and not to obscure it.

3.15.13 The quantitative treatment of resistance

In the presentation we are recommending here, the pupil will already have been introduced qualitatively to the concept of resistance and have gained some feeling for it. Now that he knows something about potential difference, he is able to study resistance quantitatively. This study should begin with a discussion of how resistance could be measured. The discussion could be illustrated by means of a very simple circuit consisting of a battery, a coil of "resistance"-wire and an ammeter, all connected in series. The current in the circuit could be read and noted. Then a coil of different resistance is substituted for the first coil, and the current is read again. Suppose it is smaller: according to earlier idea of resistance, this means the second coil has a bigger resistance than the first. So far, nothing new has emerged. But one can now point out that the potential difference across the coil was the same in the second case as in the first (a voltmeter can be used to check this), and stress that for a <u>given</u> potential difference, the current in the coil said to have a <u>bigger</u> resistance is <u>smaller</u>.

/This.....

This can be re-worded in the form "To produce the same current in the coil we say has a bigger resistance, we have to apply a bigger potential difference across it" - a statement which is also subject to experimental check. At this point the plunge can be taken and the suggestion made that we could measure the resistance of a circuit element, in fact define it, by measuring the potential difference across the element and the current in it and dividing the latter into the former. (Some doubts might be raised in certain quarters about the logical status of this definition of resistance. Is it purely operational, and therefore suspect? Can resistance only be defined constitutively in terms of resistivity? The answers to these questions are not as clear-cut as might at first be supposed. We consider them in detail in Section 7.19. In the meantime, we suggest that in the junior secondary school it will be satisfactory to take $\frac{\text{potential difference}}{\text{current}}$ as the definition of resistance. Cf. also our remarks below concerning Boulind's definition.) From this one proceeds without difficulty to the unit of resistance, the volt per amp, which is re-named the ohm. It may be felt by some that the introduction of the quantitative definition of resistance is a shade abrupt in this presentation. Certainly it is more usual to spend time first on Ohm's Law and so demonstrate that the quantity $\frac{\text{potential difference}}{\text{current}}$ is a constant for certain classes of conductor, and then to say, "Let's call this quantity the resistance of the conductor." But this latter method has the obvious disadvantage of seeming to suggest that the concept of resistance is somehow bound up with Ohm's Law. Such a faulty association of ideas will naturally be reinforced if Ohm's Law is mis-stated in the form "The current in any conductor, at uniform temperature,

/is.....

is proportional to the potential difference between its ends"! (Gordon and Neser, no date.) Perhaps nowadays there are fewer teachers and authors of textbooks who believe that Ohm's Law holds for any conductor. Indeed for any treatment of resistance which approaches the concept via Ohm's Law to have any real claim to respectability, it would have to go on to show that there are certain classes of conductor for which Ohm's Law is not true. (It is remarkable, by the way, that in the PSSC textbook, after an exhaustive treatment of voltage-current characteristics, resistance is introduced as the proportionality constant in the statement of Ohm's Law.) But even if this kind of care is exercised, there is still a danger that the initial entry of "resistance" in the context of Ohm's Law will lead some pupils to believe that there is an inseparable connection with the two. Therefore we would recommend that resistance should first be defined as $\frac{\text{voltage}}{\text{current}}$, and that if time permits, class experiments should be done to show that for some kinds of conductor, the resistance is independent of the current at constant temperature, whereas for others it is not. We say "if time permits", because in our opinion there is no urgent necessity for a detailed treatment of resistance in the junior secondary school. It is sufficient for the moment that resistance should be defined as $\frac{\text{voltage}}{\text{current}}$, and that the impression should not be left with pupils that this quantity has to be constant. Perhaps the teacher could raise the question in this way: "Here is a coil of resistance wire. I connect it into this circuit, and measure the potential difference across it and the current in it. I divide the second result into the first, and I get its resistance. Will I always get the same answer for the resistance? If I changed the circuit by putting another

/cell.....

cell in series with everything else, this would probably give me a new current. Would the new current divided into the new voltage have to be the same?" (Let it not be too hastily objected that this kind of question is, strictly speaking, tautological if the coil is an ohmic conductor; we can certainly use an ordinary ammeter and voltmeter to tell an ohmic conductor from a non-ohmic one. We return to this question later.) To which the teacher or his pupils must supply the answer that one cannot possibly know until one has tried it; that the safest guess to make is that the resistance will not be the same. As we have said, if there is time the pupils could investigate this question experimentally, using examples of ohmic and non-ohmic conductors carefully chosen by the teacher, but we do not think this has to be "in the syllabus." In the senior years pupils will be better equipped to look into the significance of ohmic and non-ohmic behaviour of conductors, and so we feel this topic is better studied there. Nor need any attempt be made to introduce or to prove the formulae for the effective resistance of a number of conductors connected in series or in parallel.

We wish to say a final word here about the definition of resistance. If any definition of this concept has to be learnt, it should be the simple quantitative statement that "the resistance of a conductor is the potential difference across it divided by the current in it." This is brief and precise and it tells one all one needs to know about resistance for practical purposes. Boulind (1948) gives the following definition (at least we presume that it is intended as a definition, since he gives it twice in three pages, the first time in italics): "Resistance is the property of a conductor wnereby potential difference is necessary to drive current through it." There is no serious objection

/to.....

to this as a statement telling us about resistance, although one might take exception to the idea of potential difference driving current through a conductor. But the briefer definition given above tells us the same thing and more besides. It is clearly to be preferred both as a definition and as something to be remembered about resistance.

3.16 Elementary electromagnetism

3.16.1 Elementary magnetism

We have referred to the importance of a knowledge of magnetism as a prerequisite for this introductory study of electricity. This knowledge need be qualitative only, since the relationship between magnetic forces and the magnetic field vector is only needed in the senior secondary school course. The basic material is so simple and so familiar that nothing need be said here about techniques of demonstration or experimental procedures. On the other hand the interpretation of some of these demonstrations and experiments does perhaps need some comment. The first point concerns the nature of a magnet. It is important for pupils to learn to describe a magnet by its properties (it will set itself roughly along a north-south line if free to rotate horizontally, it attracts "magnetic" objects to its ends), since the principle involved is one of very general application in physics. The second point concerns the description and nature of a magnetic field. In this first acquaintance with the idea of a magnetic field, the pupil learns that he can associate a definite direction with any point in a magnetic field. In Section 2.2, we referred to the pointlessness of defining this direction in terms of a "free north pole", and recommended that it should be defined instead as the direction in which a tiny compass-needle would point.

/Pupils.....

Pupils should also learn to realize that "lines of force" have no physical existence, so that when they deal later with electromagnetic phenomena, they should not place too great a reliance on "lines of force." Finally, the three-dimensional nature of a magnetic field should be stressed.

It is usual in textbooks on magnetism to offer some simple explanation of the behaviour of magnetic materials. In its most naive form, this "explanation" is a description of the "molecular magnet" theory in which no attempt is made to suggest to the reader that the "molecular" magnets are anything other than molecules which happen to be magnetic. This form of the "molecular magnet" theory has long since been discredited and it should not now be taught to pupils. It is not necessary to tell pupils about magnetic domains, though there might be no harm in mentioning the term. What they should know is that a theory of magnetic behaviour based on the hypothesis of "elementary magnets" does seem to give us a correct picture of the behaviour of ferro-magnetic materials, but that this does not entitle us to make any rash assumptions about the nature of the "elementary magnets." In Professor Rogers's book (1960) and in the Nuffield Physics Course (Nuffield Guide III), the "elementary magnet" theory is used "to teach the use of a theory." The basic assumptions of the theory are arrived at from observation of what happens to magnets when they are cut in two, and is used to predict the magnetic behaviour of a magnetized ring. To clinch the argument, the theory is put to experimental test. This is an exercise well worth carrying out if time permits. At the same time it would also be worth telling pupils (or at any rate the more advanced ones) that <u>in this simple form</u>, the theory cannot take us very far; we need to try to make it quantitative, and we need to try to fill in the details of what the "elementary magnets" consist of. These matters can be mentioned

/as.....

as an indication of the deeper problems facing physicists concerned with studying the properties of matter.

3.16.2 The action of "keepers"

It is also traditional to make some mention of the action of "keepers" when talking about magnets. Some accounts of this action are misleading, if not downright inaccurate. Consider the following quotation: "When individual magnets are allowed to lie about, they slowly lose their magnetism...Free poles at the end of each magnet act inductively on the steel to reverse the magnetism existing in the bar, so that the magnet tends to reverse itself in the absence of a keeper." (Gordon and Neser, no date). As set down, these statements leave the impression that a magnet, left entirely to itself, will gradually lose its magnetism under the influence of its own demagnetizing field. Of course this is not true; but perhaps the most unfortunate aspect of these statements is their implication that a magnetizing or demagnetizing field does not immediately have its full effect on the material to which it is applied. For pupils to be left in no doubt as to the true state of affairs, they should do an experiment in which a specimen of ferromagnetic material is magnetized by momentary application of the magnetic field of a current in a coil. They should convince themselves that merely flicking the switch in the circuit on and off will result in as effective magnetization as leaving the current on for several seconds or minutes. If they are then led to wonder why it is necessary to stroke a specimen repeatedly when magnetizing it by stroking with a magnet, so much the better. As to the action of keepers, there is no reason why pupils should not be given an accurate account of their function, viz. that magnets and keepers form closed magnetic rings which cannot easily be influenced by the presence of other magnets in the vicinity,

/and.....

and that the effects of temperature changes and of mechanical shocks are less severe with keepers in position.

It is possible to go further and discuss why the presence of free poles makes it easier for a magnet to lose its magnetization under the action of thermal and mechanical shocks. That is, the teacher can talk about the effect of the "demagnetizing field" set up by the "free poles" at the ends of the magnet. The concept of a "demagnetizing field" may well be more easily understood by a pupil whose knowledge of magnetism is as yet rudimentary, than by a more advanced student, who knows that the $\vec{B}$-vector is the fundamental vector of the magnetic field and that inside the magnet, the "lines of force" of $\vec{B}$ go from the South to the North pole. Such a student (and possibly some physics teachers who have thought about the problem - see below) may wonder why it is that the magnetic domains are supposed to be influenced by the $\vec{H}$-field of the "free poles" rather than by the $\vec{B}$-field. As Pugh and Pugh (1960) point out, if the domains <u>were</u> influenced by the $\vec{B}$-field inside the magnet, the effect of thermal and mechanical shocks would be to enhance rather than destroy the existing magnetization. One of the points at issue here is the meaning of the $\vec{B}$-vector inside the magnet. The $\vec{B}$-vector whose lines go from South to North pole inside the magnet is related to the "magnetizing force" ($\vec{H}$) and "magnetization" ($\vec{M}$) vectors by the equation $\vec{B} = \mu_o(\vec{H} + \vec{M})$, if SI units are used in general and, in particular, "Sommerfeld" units are used for the magnitude of $\vec{M}$ (cf Coulomb's Law Committee 1950). In principle, the value of this $\vec{B}$ could be obtained by cutting a slot in the material perpendicular to the direction of $\vec{M}$ and measuring the value of $\vec{B}$ in this slot. In practice, averaged-out values of this $\vec{B}$ are obtained by measurement of the

/flux.....

flux through a given cross-section of the specimen or, in more sophisticated experiments, by measuring the deflection undergone by very fast charged particles or slow neutrons that have passed through the specimen (cf. Section 7.15.2). It should be clear that neither the "slot" value of $\vec{B}$ nor the average $\vec{B}$ obtained in conventional measurements correspond to the $\vec{B}$-vector "felt" by the atoms of the material. The atoms are essentially in free space, and the $\vec{B}$-vector acting on them will be (in SI units) equal to μ_o times the effective value of $\vec{H}$ in their vicinity and not to $\mu_o(\vec{H} + \vec{M})$. This $\vec{H}$ will not be equal to that normally specified, which is determined by contributions from macroscopic currents and from "free poles" on the surfaces of the material: to those contributions will have to be added another from the material surrounding the atom in question. For instance, just inside the ends of a long cylindrical bar magnet, the value of the "conventional" $\vec{H}$ is easily shown to be equal to $-\frac{\vec{M}}{2}$ (cf. Scott 1959). However, the effect of this "negative" field is partially cancelled by the contributions from the surrounding material which, if one is justified in assuming spherical symmetry, would give a resultant of $+\frac{\vec{M}}{3}$. This still leaves an effective $\vec{H}$ equal to $-\frac{\vec{M}}{6}$ which is oppositely directed to $\vec{B}$ and is thus a demagnetizing field. In general, it is this "effective" $\vec{H}$ which must be thought of as determining the "demagnetization" in the material.

That these ideas still cause difficulty and confusion is illustrated by some recent correspondence in the journal "Physics Education" (Birch 1968, 1969; McCaig 1969). Birch correctly observes that it is questionable to imply that what is usually called the "self-demagnetizing field" is <u>directly</u> responsible for a demagnetizing action. He considers that "taken in conjunction with the amperian current model there is no place in the <u>macroscopic</u> MKS-Sommerfeld approach for

/an.....

an <u>active</u> self-demagnetizing $\vec{H}$ field; certainly the approach does not give place to a genuine self-demagnetizing $\vec{B}$ field. Perhaps a satisfactory atomic model for the latter can be devised, but one is still faced with the problem of reconciling this with a macroscopic picture which seems to preclude it." This is unduly pessimistic. From our discussion above, it will be apparent that Birch is right in suggesting that the use of macroscopically defined $\vec{B}$ and $\vec{H}$ vectors (i.e. those related by the equation $\vec{B} = \mu_0(\vec{H}+\vec{M})$) cannot clarify the demagnetizing process at the microscopic level. However, our approach shows that within the "MKS-Sommerfeld" context and with the aid of the "free-pole-on-surfaces" concept, it is possible to account for a "genuine self-demagnetizing $\vec{B}$ field", which is not to be identified with the macroscopically defined $\vec{B}$ or related to the macroscopic $\vec{H}$ by the proportionality factor μ_0. As for the problem of devising a "satisfactory atomic model" for self-demagnetization, based on an "Amperian current" approach, this would seem to have been solved by Warburton (1934), whose method is simply to picture a specimen of magnetized matter as containing vast numbers of similarly-aligned atomic current loops. From the well-known properties of the field of a magnetic dipole, he shows that at a particular point in the specimen, the resultant magnetic field will be made up of a "demagnetizing field", resulting from the contributions of all the "atomic dipoles" lying within a certain region of the specimen, and a "self-magnetizing field", resulting from the contributions from all the atomic dipoles lying outside this region. The resultant field will depend on the relative magnitudes of these two fields, which in turn will depend on the position of the point considered and on the ratio of length to diameter of the specimen..

There is clearly no problem here of reconciling this picture with a "macroscopic picture which seems to preclude

/it".

it". The so-called "macroscopic picture" simply tells us that where there is a certain macroscopic $\vec{H}$, there must be a corresponding macroscopic $\vec{B}$, or vice versa. How self-demagnetization "works" must be explained at least partly in terms of a microscopic picture, i.e. in terms of the field that must act on a microscopic particle within the material. Our discussion above has outlined two equivalent ways of arriving at an expression for this field.

In view of the interest and importance of these points, it is a pity that they are often omitted from discussions of permanent magnets to be found even in advanced textbooks. Even a recent treatment of the topic intended for the guidance of physics teachers (Thumm 1968) sums the matter up with the somewhat misleading statement that "The torque due to $\vec{H}$ [i.e. the $\vec{H}$ arising from the "free poles" alone] exerted on the aligned domains is responsible for the domain alignment in the direction of ($\vec{B}$-iron) when domain rotation is facilitated by thermal or mechanical shocks."

To the pupil who knows nothing of $\vec{B}$, $\vec{H}$ and $\vec{M}$, the matters just discussed can be of no concern. As we have implied, he need only appreciate that the presence of poles on a magnet must give rise to a "demagnetizing field" inside the magnet, so that if the magnet were heated or knocked about this field would help it to lose its existing magnetization. He should certainly <u>not</u> be told that the magnet "tends to reverse itself in the absence of a keeper"; for the magnet to "reverse itself", there must somehow be a demagnetizing or reverse field in the magnet even if the magnet has lost all its magnetization! Just how this is supposed to arise the authors of the above quotation do not explain.

Finally, the teacher might like to note that keepers are not as necessary with the best magnets as they once used to be, since the best of modern materials are much more stable than those that were once in common use. (cf. Lee 1963).

/3.16.3.....

3.16.3 The magnetic effects of electric currents

We have said earlier that some work on the magnetic effects of an electric current should precede any treatment of the measurement of current. We now consider just how far a study of these magnetic effects should proceed in the junior secondary school. A typical syllabus (Orange Free State 1967) specifies the magnetic fields of a current-carrying straight wire and solenoid, the electromagnet, the electric bell, the force on a current-carrying conductor in a magnetic field, the principle of the ammeter and the principle of the electric motor. With this list of topics one can certainly rest content. Regarding the magnetic fields of the straight wire and the solenoid, the vital points to make are that the field lines appear to be closed and that there is marked similarity between the external field of a solenoid and that of a bar magnet. Of course it is desirable that pupils should see real "iron-filings" demonstrations of the magnetic fields of these current-carrying conductors, or that they should plot these fields with plotting compasses. To show the similarity in magnetic behaviour of the current-carrying solenoid and the bar magnet, the forces of interaction between a bar magnet and such a solenoid should be demonstrated. Demonstrations of the force on a current-carrying conductor should make plain that this force is a maximum when the conductor is at right angles to the field, though no formula should be given in this connection; naturally one cannot in this context demonstrate that the force depends on the current in any quantitative way, since we assume a relationship in making our fundamental measurement of current depend on the magnetic effect.

A question of some importance arises in connection with the elementary teaching of the force on a conductor in a

/magnetic.....

magnetic field. Pupils must be convinced of the fact that if the conductor is held at right angles to the field (which it always will be in applications of any interest to them), the direction of the force on it will be at right angles both to the conductor and to the field. (They should also know that the force is <u>always</u> at right angles to the conductor.) Is it also necessary for them to know a rule relating the directions of the field, current and force? In the Nuffield Guide (III) it is strongly argued that in a first encounter with the subject, it is not necessary for pupils to know such rules. The Guide points out that it is a good deal more important for a pupil to know (i) that the three directions are at right angles to one another and (ii) that if the direction of the current is reversed, so is the direction of the force, than for him to memorize a rule which might in the end bulk larger in the pupil's mind than the physics on which it is based. In the words of the Guide: "... the rule and the mnemonics used by pupils, take charge and appear to some pupils to be the main substance of electromagnetism... Physicists often say that the rules are also essential for pupils who are to become electrical engineers; but we are told that intelligent electrical engineers just turn the current on and see which way the wire moves." On this view, when discussing the principle of operation of the D.C. motor or the moving-coil galvanometer, one should stress that if the force on one side of the rectangular coil is one way, the force on the opposite side must be the opposite way - the actual directions do not matter. If there is indeed a danger that emphasis on a left-hand or right-hand rule might obscure the physics behind the rule, we would feel it advisable to play for safety and not teach the rule in the early stages. It can always be produced when needed.

/Should.....

Should any attempt be made to "explain" the force on a current-carrying conductor in a magnetic field? In our view, the answer must be firmly in the negative. Relativity apart, we cannot explain this force (except in terms of the forces on moving charges, and this is not a significant step forward here) and we should not pretend that we can. Thus while we may demonstrate the "catapult" pattern of the combined uniform magnetic field and a current-carrying wire, the appearance of the pattern should not be used in any pseudo-explanation of the force on the wire.

Much has been said elsewhere on the question of whether or not we should continue to use the well-established convention that "current is in the direction from positive to negative" in the teaching of electricity. To say that "in ordinary wires it is the electrons which carry the charge and they go from negative to positive" is not, in our opinion, sufficient reason for reversing the familiar convention. For we do need a convention, which has either to be the old one or its opposite; and whichever we decide on, we shall have to contend with the complex situation that arises in a current-carrying electrolyte, semiconductor or plasma, in which we have charge carriers of opposite sign moving in both directions. Perhaps the use of the "direction-of-the-electron flow" convention would be found marginally easier by the beginner, but for this minor advantage we do not feel an impelling urge to start the kind of large-scale revolution which would be needed to revise existing textbooks and teaching aids in this respect. There are too many other matters in physics teaching about which we feel far more concern.

3.16.4 Electromagnetic induction

It seems now to be customary for junior secondary science syllabuses to include an introductory treatment of electromagnetic induction. This is perfectly satisfactory,

/as.....

as long as the treatment remains "phenomenological"; ; that is, it confines itself to demonstrations of the basic phenomena and to discussion of some of their applications. It is well-nigh impossible to give a satisfactory elementary explanation of the facts of electromagnetic induction; the attempts that have been made in this direction only serve for the most part as proofs of the correctness of this statement. In our earlier section on oversimplification we drew attention to the absurdity of a typical "explanation" of the action of the transformer in terms of lines of force, according to which the expanding or collapsing lines of force around the primary coil cut the secondary coil on their way out or in. That this kind of statement is still encountered is illustrated by the following extract from a recently-published textbook: "'n Transformator bestaan uit twee klosse wat gerangskik is sodat die magnetiese kraglyne van die een klos gesny word deur die ander klos." (A transformer consists of two coils so arranged that the magnetic lines of force of the one are cut by the other coil.) (Prinsloo et al. 1965) In the Orange Free State Teachers' Guide, the following alternative explanation of transformer action appears: "Electric wave inside conductor causes a magnetic wave outside it. Magnetic wave motion outside second conductor induces a current inside it". Lest there be any doubt as to the supposed meaning of the term "electric wave", we quote the remarks in this guide concerning alternating current: "Introduce alternating current; Draw attention to alternating polarity; describe as an electric wave passing through the conductor." Once one appreciates in what sense the term "wave" is being used here, or rather misused, one recognizes that there is a glimmering of sense in this explanation. But the form in which it appears illustrates rather well the desperate straits to which people are appar-

/-ently.....

ently driven in trying to make electromagnetic induction palatable to beginners.

One might think that a way out of these difficulties might be provided by Faraday's Law of Electromagnetic Induction itself, which does after all sum up the whole subject. Evidently this is the view adopted by Boulind (1948), since near the beginning of his chapter on electromagnetic induction he quotes Faraday's Law in the following form:- "The induced E.M.F. is proportional to the rate of change of the lines of force passing through the circuit." Properly interpreted, this statement of the law is perfectly correct; but what is it to convey to the uninitiated pupil? Boulind's own suggestion to his readers to "Think of the magnet as having attached to it an imaginary web of lines of force" would help to make the statement intelligible in situations in which there is relative movement between a circuit and a magnetic field. But it will not help to explain the action of a transformer, in which there is no motion relative to a magnetic field. Only if the pupil knows the convention by which the strength of a magnetic field is related to the number of lines of force per unit area of a surface drawn in the field, will he able to understand what Faraday's Law is saying in this case. Since the concept of "magnetic field strength" is not one with which the pupil is acquainted as yet, it is not possible for him to interpret "number of lines of force" in this way, and this kind of statement of Faraday's Law can only confuse him as to the meaning of lines of force.

We suggest that if an "explanation" of electromagnetic induction is to be given at all to junior secondary pupils, it should be formulated in the following terms: "When a magnet moves towards a solenoid, the part of the magnetic field enclosed by the turns of the solenoid <u>changes</u>.

/When.....

When a dynamo coil is rotated in a magnetic field, the magnetic field inside the coil changes (compare the situation when the coil is edge on to the field with that in which it is at right angles to the field); and when a current is switched on in one coil which is near another coil, the magnetic field in the vicinity of the second coil changes (there was no magnetic field to begin with - if we neglect the earth's field - and then there is some field, so clearly there has been a change). In all these cases a current is induced in the circuit through which the magnetic field changed, so we conclude that if the magnetic field surrounded by a circuit changes, there will be a current induced in the circuit". The bigger and the more rapidly we make the changes in the field, the bigger the current seems to be. The idea of a change in a magnetic field is used here in a way which we hope will be felt to be unexceptionable on the one hand and not entirely devoid of meaning on the other. To those who would complain that the above account does not "explain" electromagnetic induction, it just tells one what is happening in terms of the language of fields, we would retort that most so-called "explanations" of the phenomenon amount to no more than this anyway, and that the account just given at least provides a reasonable start for a later discussion of Faraday's Law. Of course one can give a more direct explanation of motional induced EMF's (i.e. EMF's induced by "flux-cutting") by the use of the force on a current-carrying conductor in a magnetic field; when the conductor moves in the magnetic field, it is carrying its free charges along with it, these are moving charges and so must experience "magnetic" forces, etc., etc. Although it is tempting to "wrap up" the phenomenon in this way by linking it to something the pupil already knows about, we think it best to resist the temptation

/in.....

in this case. The author's experience in talking to first-year university students about this topic leads him to state that to explain this aspect of electromagnetic induction properly needs time and care on the part of the teacher and a thorough quantitative familiarity with "magnetic" forces on conductors on the part of the pupil. Accordingly, it is best left to the senior secondary school.

"A simple introduction to alternating current" is also a typical entry in junior secondary syllabuses, and indeed it follows naturally enough upon a discussion of electro-magnetic induction. "A.C." can be introduced with reference to a "transformer"-type pair of circuits, in which a regular switching on and off of the current in the primary circuit will produce a regular alternation of current in the secon-dary circuit, as could easily be shown by a short-period galvanometer. Or it could be done with the aid of a simple dynamo (rectangular coil rotating in a uniform magnetic field) but only, we suggest, if the voltage output of the dynamo can be displayed on an oscilloscope screen. It is out of the question in an elementary treatment of the subject to try to deduce the sinusoidal shape of the voltage-time curve from the motion of the coil in the field - even to try to bring in Lenz's Law and so deduce the alternation of current di-rection every half-cycle is, we feel, an unnecessary compli-cation at this point. So we suggest that the pupil should see the sine-curve (preferably <u>not</u> to be called a sine-<u>wave</u>) on the oscilloscope screen, as produced by a real dynamo. This would require (i) a dynamo consisting of a rectangular coil of at least 20 turns, (say) 8 cm long by 4 cm wide; (ii) an electric motor capable of driving the dynamo at a rotational speed of at least 1000 revolutions per minute ; (iii) a strong permanent magnet or electromagnet capable of supplying a reasonably uniform field of at least 200 gauss

/over.....

over the space in which the coil rotates ; (iv) an oscilloscope with a vertical deflection sensitivity of 0.1V /cm or better and capable of synchronizing or triggering successfully at frequencies as low as 20 Hz. Increase in the number of turns on the coil, the width of the coil, the effective cross-sectional area of the field, the strength of the field and the speed of rotation of the dynamo would all increase the output of the dynamo. On the assumption that the magnet has pole pieces just over 4 cm apart and 4 cm square and supplies a field of 200 gauss, that the coil is 4 cm wide and has 20 turns and that its speed of rotation is 1200 revolutions per minute, this dynamo would have a maximum output of the order of 75 millivolts. The magnet is the main limiting factor here in designing a dynamo capable of producing a detectable pattern on the screen of the oscilloscope. We would urge that schools be supplied with such magnets, which would be useful not only in making a dynamo but in experiments on the force on a conductor in a magnetic field as well.

Use of the oscilloscope to demonstrate the alternating nature of the output from a simple dynamo must naturally have been preceded by an introduction to the oscilloscope along the lines suggested in our earlier section on black boxes. Once pupils have seen the typical dynamo "wave-form", they can be showr that the output from a mains socket (suitably stepped down) has the same "wave-form". It is not necessary to lay stress on the frequency of the mains voltage in this treatment; the background idea can be expressed by saying that "the current changes direction 100 times a second."

Since alternating current is the principal source of household electricity, and since much of its use in the home is concerned with the heating effect of a current, some attempt should be made to justify to pupils that the heating

/effect.....

effect does not depend on the direction of the current. (By this time pupils should be thoroughly familiar with the fact that electricity has a heating effect and with the standard application of this effect.) It is not necessary to produce a "mathematical" justification of this fact, e.g. by deriving Joule's Law and so showing the dependence on the square of the current; this can be left to the senior years. Instead the matter can be approached experimentally by way of the study of a simple circuit consisting of a cell and a lamp. Reversal of the connection of the leads from the cell to the lamp produces no change in the brightness of the lamp; if an ammeter is included in the circuit and the connections to it are reversed also, this will show that the size of the current in the lamp is unaltered by the change in connection. Next, pupils can be asked to imagine (or to try) the operation of rapid to-and-fro reversal of the connections from the cell to the lamp, and to consider how this would affect the energy delivered to the lamp. They should not take much convincing that this "square-wave" current in the lamp would produce the same "heating effect" in it as a direct current of the same value, if the time spent in "switching over" were made negligibly small. From here it is a small step to convincing pupils that a sinusoidally varying current must be capable of producing a heating effect.

The above introductory syllabus for electricity and magnetism should prove more than adequate as an introduction to the subject and as a basis for a more advanced treatment in the senior secondary school. Provided that descriptions and explanations of phenomena are kept strictly within the confines indicated above, the teaching will have at least the negative virtue that the pupil is not likely to be too seriously misled by what he sees, does and hears in class. This negative virtue is by no means to be despised: if the

/pupil.....

pupil is to have any chance at all of gaining something worthwhile from his matriculation studies of electricity, it is essential for him to come to those studies unhampered by serious misconceptions concerning fundamental concepts. And because of the central importance of electricity to many other aspects of his matriculation physical science, one must hope that his understanding of this topic will not be in any way obscured. If it is thought that such an introductory syllabus as the one we have assumed above is too ambitious in its scope, then let its scope be reduced; for example, the work on electromagnetic induction can be omitted, without any real loss.

3.17 Wave motion

3.17.1 Waves in the school curriculum

One of the novel features of the new generation of syllabuses in school physics was their inclusion of substantial sections on wave motion. So successful has the inclusion of this "classical" material proved to be that it has now become standard practice to incorporate it in any present-day school syllabus, "modern" or "conservative." The advantages of its inclusion are plain enough. As a means of energy transfer, wave motion is of fundamental importance to mankind; and the physical theory which has proved to be most successful in accounting for the facts of the microscopic as well as the macroscopic world can be expressed in a "language" which has a great deal in common with that used to describe wave motion. If a course in school physics is to give any kind of introduction to our means of communicating with one another and to our understanding of the physical world, it will have to lay an adequate foundation for the understanding of wave motion. Fortunately the work of the P.S.S.C. has helped a great deal to enable the teacher to make wave motion both palatable and exciting to his pupils.

/The.....

The design of cheap ripple tanks for individual use by pupils and the decision to use "Slinky" springs to demonstrate the propagation of both longitudinal and transverse pulses along springs are two of the most noteworthy contributions of the PSSC team. These developments make it feasible for pupils to begin their study of wave motion in the junior secondary school.

3.17.2 Waves on springs

An easy way of using a "Slinky" spring to demonstrate longitudinal and transverse waves is to stretch the spring out along the floor or the lecture bench to a length of three or four metres, secure one end and then send a transverse or longitudinal pulse down the spring from the other end. Although there is some damping of the motion, resulting in a marked reduction of amplitude as the pulse travels down the spring, the effect is by no means large enough to obscure what is happening. If it is desired to do these experiments with less damping, the PSSC system of suspension of the spring may be adopted (see PSSC Teachers' Resource Book and Guide, Part Two); that is, the spring is supported by long pieces of string spaced at regular intervals along the spring and having their upper ends attached to curtain rings which can slide along a rod fixed to the ceiling. One way of arousing interest in these experiments with springs is to ask the pupils to time the passage of a pulse down the spring with a stopwatch and so measure its speed of travel, and then invite them to explain just what it is they have measured the speed of. A transverse pulse travels quite slowly down a "Slinky", which is one of the reasons why this demonstration is so effective. To provide a contrast with this and to show propagation of a pulse in another "medium", a pulse can be sent down a few-metres'-

/length.....

length of rubber tubing. (Pupils could be put on their mettle by being asked to time the travel of this pulse!)

With the aid of these simple experiments, a number of the fundamental characteristics can be made plain to pupils: a disturbance can travel through the medium without the medium as a whole undergoing any movement; successive "bits" of the medium carry out their motions in succession, not in step; the motions of the "bits" can be at right angles to or along the line in which the disturbance travels or both; and energy is transferred by this means from one point in the medium to the other. This last and most important point can perhaps be driven home by showing the contrast between giving energy to something by throwing an object at it, and giving energy to it by jerking it by means of a pulse sent down a spring. A point worth experimental investigation by enterprising pupils is the uniformity of the speed of propagation along the medium; an approximate check of this can be made for transverse pulses along a "Slinky". This point is not as trivial as it may seem. Those of us who know something of the theory of wave propagation will know that the speed of wave travel depends on the characteristics of the medium, and will therefore conclude that if the medium is uniform, the speed of propagation must be uniform as well; the pupil may or may not have some intuition about this, but it would be a good thing for him to check for himself.

There is only one more thing that can usefully be done with these "one-dimensional" waves at this stage, and that is to demonstrate the reflection of waves. This is easiest to do with a length of rubber tubing or round-wire spring; the pulse is reflected with little diminution of amplitude at the ends of the "medium". This is not the moment to draw special attention to the inversion of the pulse on reflection; it is the phenomenon of reflection and not the associated phase

/change.....

change which is important here. Reflection of a longitudinal pulse can be shown with a suspended "Slinky".

3.17.3 Waves on surfaces: the ripple tank

The pupil's attention should now be turned to the ripple tank. This is necessary because he has to become familiar with waves in two dimensions as a preparation for the study of sound and light, and because the ripple tank is perhaps the simplest means of illustrating the concepts of wavelength and frequency. To some extent one can demonstrate the latter concepts with the aid of a "Slinky" or a piece of stretched tubing by shaking one end and so show the generation of a periodic wave; but this sort of demonstration is complicated unavoidably by the reflections from the far end of the "medium". In the ripple tank on the other hand it is easy to generate trains of circular or straight waves without hindrance from reflections. A valuable preliminary demonstration for work with a ripple tank is that which shows surface waves in liquids in section. The Nuffield Guide (III) suggests generating these on a paraffin-water interface in a tank one-third of full of water and one-third full of paraffin on top. The slower motion waves thus produced enable the motion of the liquid near the interface to be more easily observed. On the ripple tank, pupils should first generate circular and straight pulses and observe their reflection from barriers of various shapes. Just how far this kind of work should be carried depends on the requirements of the syllabus. In the Nuffield course a considerable amount of time is allotted to individual work with the ripple tank, and in the course of this pupils are led to the idea of image formation by waves and make preliminary investigations of interference, diffraction and scattering. The new South African syllabuses for science in the junior secondary

school would not seem to allow room for this kind of extended introductory treatment; nor indeed do they specify it. In accordance with these syllabuses, we shall assume that reflection of waves will be no more than noted at this stage and that interference and diffraction will be omitted altogether.

The study of periodic circular and straight waves on the ripple tank should be carried out with the aid of the "rotating disk" type of stroboscope made popular by the P.S.S.C., so that pupils can "stop" the waves and make measurements on them. In the Nuffield course, teachers are urged to wait for pupils to make the suggestion that they should measure the "wavelength" (or whatever term they initially use for this concept). In the present context, we do not feel that this recommendation has to be followed; but whether teachers prefer to approach the topic in this way or not, we <u>would</u> recommend that the treatment of periodic waves be accompanied by a simple description of their generation in a "one-dimensional" medium, so that pupils can grasp that one up-and-down-and-back-again motion of the source generates one complete "wave". This should make it easy for pupils to understand the concepts of "wavelength" and "frequency" and to see that the frequency of the waves must be the same as the frequency of the source. A very simple definition of "wavelength" will be perfectly adequate here; pupils do not yet know about "phase", and even to define wavelength as "the distance, in the direction of propagation, between any point along the wave and the nearest point which is in the same state of vibration" (Prinsloo et al 1968) is too sophisticated a procedure for the moment. Pupils should appreciate that they are measuring the wavelength of the projected pattern on the sheet of paper beneath the ripple tank,

/not.....

not the actual wavelength of the surface waves in the ripple tank. This is not a point which need cause pupils or teacher any concern; in particular we would not regard it as essential for pupils to work back from these measurements to the true wavelength of the waves. Combination of frequency and wavelength measurements to yield a velocity of the wave should also be done by pupils; explanation of the relation "wavelength times frequency equals wave velocity" should give the teacher no trouble. It would be instructive for pupils to carry out these measurements for various frequencies of the source; since water is a dispersive medium, they should find that the wave velocity is different for different frequencies. (One wonders what the unsuspecting teacher would make of this, if he accepts the following piece of advice from the O.F.S. Teachers' Guide: "Point out that the velocity of the wave remains the same, even though the crests pass a fixed point in more rapid succession".) Early acquaintance with the phenomenon of dispersion would eliminate the danger of the pupil coming to accept non-dispersive media as the norm. We do not suggest that special attention should be drawn to this phenomenon or that it should be given its official name; we would merely wish pupils to take it as natural that the speed of waves can depend on their frequency.

At the moment it is accepted practice in the design of South African syllabuses that refraction should first be studied in a ripple tank as showing another aspect of the behaviour of waves. In the P.S.S.C. and Nuffield courses and also in the South African Joint Matriculation Board syllabus of the mid-nineteen-sixties, ripple tank studies of refraction follow a first encounter with the phenomenon in the context of optics. Either teaching sequence has points in its favour; we shall not debate the matter here,

/but.....

but continue on the assumption that refraction will be treated as part of the general introduction to wave motion. The points to be brought out in connection with the refraction of water waves are (i) that when the waves move into a region where their velocity is different, their frequency is unaltered; (ii) that if the waves are other than normally incident on the boundary between two media, their direction will be more nearly normal to the boundary in the medium in which their velocity is slower. Optimum conditions for securing clearly observable refraction in a ripple tank are set out in the P.S.S.C. Teachers' Resource Book and Guide, Part Two. In principle, it would now be possible to make measurements on the angles of incidence and refraction and "discover" that the ratio of the sines of these angles was equal to the ratio of the measured wavelengths and hence of the wave velocities. We do not recommend this, however, as the limitations on the accuracy of these measurements are such that the pupil could not hope to "discover" this law; the best that could be done would be for the teacher to tell him the law so that he could confirm that his measurements were consistent with it within the limits of experimental error. (Certainly no attempt should be made at this stage to derive this result theoretically.) At some stage the pupil should carry out these measurements; but it would be better for him to do this by way of testing that the behaviour of light is in this respect like the behaviour of waves on water. The focusing of waves by means of refraction can be shown with the aid of a plate of glass shaped like the cross-section of a lens; if this is placed in the ripple tank, the water above it will act like a lens on the waves incident upon it.

/3.17.4 .....

3.17.4 Graphical representation of transverse and longitudinal waves

Rather surprisingly, the graphical representation of transverse and longitudinal waves is included in the section on waves in the South African junior science syllabuses. We say "rather surprisingly", because we would have thought that the formal graphical treatment of waves implied by the inclusion of this topic was a little premature for junior secondary school pupils. This reservation applies especially to the graphical representation of longitudinal waves, since here pupils have to grasp the fact that longitudinal displacements are plotted vertically on the graph, so that the resulting curve bears no resemblance to the physical appearance of the wave. Still, if pupils are to cope with sound waves to the extent to which the same syllabus requires of them, they had certainly best try to master this material. The O.F.S. syllabus specifies that the cathode-ray oscilloscope must be used here as an aid to understanding. In the spirit of this recommendation, the O.F.S. Teachers' Guide says: "Demonstrate a sine wave on a cathode ray oscilloscope and point out what is meant by wavelength, frequency and amplitude. Then draw a sine curve to illustrate the concepts". All this is very odd. The alleged "sine wave" on the oscilloscope screen does, it is true, resemble the shape the pupil will already have noticed when he saw periodic waves being sent down a "Slinky". However, it is in fact a graph of "something" (voltage) varying sinusoidally with time; as such, it has no advantage in the present context over an arbitrary sine curve the pupil might plot for himself. The recommendation to use the oscilloscope here can only be understood on the assumption that authors of syllabus and Teachers' Guide have unaccountably confused a sine wave with a sine curve. How else can

/one.....

one explain the suggestion to "point out what is meant by wavelength, frequency, and amplitude" in reference to the "sine wave" on the oscilloscope screen? We need no clearer evidence that there is great danger of confusing the pupil about the terminology of the subject, its fundamental concepts and the graphical representation of waves. Obviously the teacher will need to draw a very careful distinction between curves of displacement versus time taken at a given point in the medium and curves of displacement versus distance taken at a given instant. We see no way in which the oscilloscope could help to make this distinction clear at this stage; on the other hand it is very clear to us that use of the oscilloscope could considerably increase any existing confusion. Accordingly, we cannot recommend its use here at all.

3.18 Sound

3.18.1 Sound as wave motion

In drafting the new common basic syllabus for physical science for South African Schools, the responsible committee placed most of what they had under the heading of "Sound" in double square brackets, meaning that this material could be covered in the junior secondary school. The implied recommendation has been duly adopted and we accordingly discuss the treatment of sound here.

A study of this topic begins quite naturally with the sources of sound, and especially the sources of musical sounds. We need not pursue this point here, as the usual means of showing the association between vibrating sources and musical sounds are familiar and uncontroversial. Since the emphasis for the moment is on sounds rather than their sources, there is no need to say more about the latter at this stage. In some syllabuses and treatments of sound, the next point to be taken up is "the necessity for a medium".

/Usually.....

Usually this is illustrated by the experiment in which a bell-jar containing an electric bell is steadily evacuated. When the pressure inside is low enough, the bell can no longer be heard. From this it is concluded that air is necessary to transmit sound from the bell to the jar, and this is taken as a particular case of the necessity for a medium for the propagation of sound. Of course the conclusion is correct but, as Lindsay (1948, 1963, 1968) has pointed out, it does not follow directly from the above experiment. What the experiment really shows is that as the density of the air in the jar is reduced, sound is transmitted less readily from the bell to the air and from the air to the walls of the jar. It is known that sound is still satisfactorily transmitted at pressures of the order of a thousandth of an atmosphere. At this pressure, a negligible fraction of the acoustic energy from the bell will be transmitted to the walls of the jar. If we are to argue from this experiment that air is needed to transmit sound from the bell to the jar, then we should not say that "when there is no air, no sound gets out, therefore we need air to transmit sound", which is the usual argument, but instead that because the sound gets fainter as the air pressure is reduced, it seems that the air inside might have something to do with the transmission of the sound from the bell. If it is still thought worthwhile to do the experiment for the pupils' benefit, well and good; but we should not hoodwink them as to its interpretation.

An alternative approach might be the following (cf. Helm et al, 1969b): "The physicist is interested both in the source of the sound and in the way sound travels. We have all heard the phrase 'sound waves' so often that we are now pretty well conditioned to believe that sound travels in the form of wave motion. And so it does, but we should not be

/properly.....

properly satisfied in this matter until we know why we are justified in talking of sound waves. Let us think of some of the things we know about sound and try to relate them to some of the things we know about waves in general". This approach has the merit that it does concentrate attention on a fundamental problem, viz. the nature of sound. In a number of the existing treatments of the subject, the wave nature of sound is either taken for granted from the beginning (cf Jardine 1965) or asserted early in the discussion. It seems strange that when discussing light, we should take care not to make assertions about its nature without some supporting evidence, whereas in the case of sound this cautionary attitude is no longer thought necessary. The pupil now knows enough about waves to be able to grasp that the behaviour of sound has features in common with the behaviour of the waves he has seen and studied. Thus he can note that sound takes a finite time to travel from one place to another, that it can be reflected and that it also can be refracted. This last point can be demonstrated by focussing sound with a balloon filled with carbon dioxide (cf. Jardine 1965, for example). One could also show that sound exhibits interference and diffraction, but since the pupil has not yet encountered these phenomena in his study of waves, such a demonstration would not avail much here. While it is true that without showing interference and diffraction, we cannot provide conclusive evidence for the wave nature of sound, some evidence is undoubtedly better than none; it should at least convince pupils that we do not simply take the wave nature of sound for granted. Nor need we leave diffraction entirely out of the picture; the pupil will have observed waves "going round corners" in the ripple tank, and it can be pointed out to him that sound does the same sort of thing. The apparently contrasting behaviour of light in this respect can be considered when the nature of light is dealt with in the Matriculation years.

/The.....

The speed of sound should be made the subject of experimentation by the pupils themselves. There are a number of straightforward methods of measuring this quantity, all of which can be both interesting and enjoyable to pupils (cf. Jardine 1965). The simplest of these is to measure the time elapsed between the firing of a gun and the receipt of its sound some distance away. If it is desired to carry out the experiment using a much shorter distance, an "echo" method can be adopted. A pupil can stand a hundred yards or so from a high wall and clap two pieces of wood together with such a frequency that he hears the echo exactly midway between the claps. A friend can measure the time taken for a known number of claps. By way of creating additional interest, a teacher with a suitably equipped laboratory could invite his pupils to co-operate with him in the experiment with a demonstration scaler described in Section 2.4 in our section on "black boxes". The results obtained in this way could be compared with those obtained in the experiments conducted over very large distances.

After consideration of this preliminary evidence, which points towards the idea that "sound" might be a kind of wave motion, the teacher could argue: "As far as we have tested it, sound seems to do the kind of thing that waves do. Let us accept that it is a kind of wave motion and see how far we can get. Since we normally think of waves as travelling <u>in</u> or <u>on</u> something, we must ask what sound waves travel <u>in</u>. It is obvious that sound waves travel <u>through</u> the air, and we can do some simple experiments to show us that sound waves travel through water and also through all sorts of solid materials. But this does not quite answer our question; we wanted to know what sound waves travel <u>in</u>." (This distinction between what waves travel <u>through</u> and what they travel

/<u>in</u>.....

in or on is far from trivial; we have only to think of the example of light to see why.) How is this discussion to be continued? We have already seen that we cannot legitimately use the bell-in-a-"vacuum" experiment to decide the question for us. It seems to us that the best that can be done in the circumstances is to say that if we develop a theory of sound propagation on the assumption that sound is a kind of wave motion in a material medium, we can use this theory to predict the speed with which sound will travel in different gases, liquids and solids. The predictions of the theory are always confirmed by experiment; in the case of air, for example, the theory gives us a formula relating the speed of sound to the pressure and density of the air, and if we plug the right values for pressure and density into the formula, we get the right value for the speed of sound in air. Unfortunately, the details of this theory go well beyond a school course in physics, and for this reason the above argument must leave some feeling of dissatisfaction with both teacher and pupil. However we shall see below that a naive "theory" of the propagation of sound waves in gases can be developed within the confines of the syllabus, and that the predictions of this theory can be linked up with work to be done in the Matriculation years.

3.18.2 The nature of sound waves

Having got to the point that sound waves are waves of some sort in material media, the next point to be established is the nature of these waves. The fact that they are longitudinal is often taken to be self-evident; it is hardly likely to appear thus to the uninitiated, and so we suggest that some attempt should be made to justify this idea. A possible way to do this is to discuss with pupils how a sound could travel down a metal rod as a wave.

/To.....

To send a pulse of sound down the rod, one end of the rod is struck; the end-most layer of rod is compressed in this action; in returning to its normal volume it overshoots and compresses the next layer, and in this way a compression travels down the rod. This picture can be compared with a longitudinal pulse travelling down a "Slinky" and with other, more homely illustrations of longitudinal waves - a jerk travelling down a tightly-packed queue of would-be rugby spectators. This kind of reasoning should help to convince pupils that sources of sound produce compressions or rarefactions in the surrounding medium which are propagated as longitudinal waves.

Particle models for the propagation of sound waves should certainly be discussed with pupils, especially with reference to sound waves in solids and gases. For solids the model is a straightforward one: a row of little balls joined to one another by springs. For gases the model is also easy to describe - it is just the molecular model of a gas the pupil will already have encountered - but the mechanism by which sound waves travel through it is not so easy to explain in terms of the pupil's vocabulary. To give a proper account of what is happening, one should refer to the momentum transferred from the source to the adjacent molecules and from the latter to other molecules with which they collide. In terms which the pupil might understand, this might become something like the following:- "When the prong of a tuning-fork is moving outward, it gives the molecules next to it an extra push, so that they will be travelling a little faster in the 'forward' direction. When they knock into other molecules, which they always have a good chance of doing, they will be able to pass on this forward push to them." One important idea that should emerge from this account is that the speed with which the sound travels in a

/gas.....

gas should be something like the average speed of a gas molecule. When the pupil makes a quantitative study of the kinetic theory of gases in the Matriculation course, he should be reminded of this prediction and shown that the speeds of gas molecules calculated from the formulae of the kinetic theory are indeed comparable with the measured speeds of sound in gases. He should also appreciate at that stage that the observed increase of the speed of sound with temperature is also to be expected in terms of this "kinetic" theory of sound propagation.

Were the syllabuses so arranged that the pupil would study sound after having met quantitative kinetic theory, the idea that

speed of sound in a gas $\sim$ average speed of gas molecule (or root-mean-square)

could be used as the basis of number of predictions concerning the speed of sound. Since the pupil will have had the formula for the root-mean-square speed derived for him in an elementary way, this formula can be used to show that at constant temperature, the r.-m.-s. speed is independent of the pressure; that the r.-m.-s. speed must be proportional to the square root of the absolute temperature (this follows by taking into account the definition of the ideal gas scale temperature); and that at the same temperature and pressure, the r-m-s. speeds of sound in different gases will be inversely proportional to the square roots of their densities. On the assumption that the speed of sound will conform to the same rules, one at once obtains a number of verifiable predictions concerning it without having to drag in $c = \sqrt{\frac{\gamma P_0}{\rho_0}}$ as an unsupported formula.

Some teachers may wish to quote this formula and to account for the presence of the "γ" in it. There are two

/objections.....

objections to be raised against this. First, the meaning of "γ" will have to be explained to the pupil, who will have to accept without justification from the teacher that the presence of "γ" in the above formula is somehow associated with <u>adiabatic</u> compressions and rarefactions of the gas. Second, in many textbooks, including a number of advanced ones, an incorrect or, at best, incomplete explanation is given of why the compressions and rarefactions are adiabatic. A typical statement is "... the pressure-volume changes in a sound wave occur much too rapidly to be isothermal and must be regarded as adiabatic changes" (McCormick 1965). This is essentially a modern rendering of Laplace's explanation, first published in 1816, of why the observed velocity of sound was appreciably higher than the value calculated on the assumption that the compressions and rarefactions were isothermal (cf. Feather 1961, Lindsay 1968). As seems first to have been pointed out by Herzfeld and Rice (1928), this explanation ignores the fact that as the frequency of the waves increases, the distance between adjacent compressions and rarefactions decreases, which would tend to make it easier for heat to flow from compression to rarefaction and so make the process isothermal. Other writers (e.g. Condon 1933, Wood 1940 and Zemansky 1957) have drawn attention to this point; Zemansky gives a simple analysis of the situation which shows that "for ordinary frequencies, say from 20 to 20,000 cycles per second, and for even the best-known heat conductors, the wavelength is too large and the thermal conductivity too small for an appreciable amount of heat to flow" (Sears and Zemansky 1964). From this one may conclude that the compressions and rarefactions at "ordinary frequencies" are adiabatic because their frequences are so low. In a reply to Whelan (1968), Blitz (1968) has suggested that this is also too simple a

/view.....

view of the matter, since it fails to take into account the fact that "a medium of propagation will in practice have boundaries across which heat may be conducted. When the frequency is sufficiently low the dimensions of the medium become comparable with the wavelength and thus heat exchanges will occur. Hence isothermal conditions should be approached as the frequency tends to zero." Certainly it is true that both in Laplace's original discussion and in the other above-quoted treatments of the problem, it is tacitly assumed that the dimensions of the medium are large compared with the wavelength of the sound waves. At any rate it is plain that Laplace's original discussion of the problem, while it has the ring of plausibility about it and leads to the right result, leaves one important factor out of account. In view of the complexity of the matter, it will be seen that it is best not to discuss the formula $c = \sqrt{\frac{\gamma P_o}{\rho_o}}$ in class at all; brighter pupils who show an interest in this as a result of their reading can be given additional references to consult.

3.18.3 The waveforms of sound waves

Another important idea arising from a study of sound waves in gases is that at any point in the gas through which the waves are travelling, there must be a variation in pressure caused by the passage of the waves. The cathode-ray oscilloscope can be used to advantage in stressing this idea, since if a microphone is connected to the terminals of its vertical amplifier, the oscilloscope will, in effect, plot a graph of pressure variation versus time. (It is likely that the microphones used for this demonstration in any school laboratory will be pressure-sensitive). In line with our earlier remarks about the use of the oscilloscope to demonstrate waves, we should like to insist here that the teacher should be sure to describe the pattern on the screen

/correctly.....

correctly - that it is a graph of variation of pressure against time and only in this sense a "picture of the sound wave". The oscilloscope should be used to illustrate the aperiodic nature of "noise" (e.g. hissing) and the periodic nature of musical notes. If it is desired to illustrate pure tones and so produce "sine waves" on the oscilloscope screen, there are a number of ways of doing this:-

(i) The sound of the human whistle is a good approximation to a pure tone; unfortunately it is difficult for the average whistler to keep his whistle steady in frequency and amplitude.

(ii) Notes in the upper register of the descant recorder (e.g. G or A in the second octave of the instrument's range) are very nearly pure tones, but a player with good breath control is needed.

(iii) If a tuning fork is mounted on a resonator of the "open-box" type, the note issuing from the box will be both pure and intense enough to elicit a satisfactory response from the oscilloscope.

(iv) As an alternative version of (iii), a bottle can be "tuned" to a given tuning fork by filling it with water to the right extent. When the vibrating fork is held over the mouth of the bottle, the note produced by the latter will produce an adequately pure tone.

(v) If the school laboratory possesses an audio-frequency oscillator and a loudspeaker, which it should, this combination can be used to produce a substantially pure tone over a wide range of frequencies. (That the sound comes from "black boxes" is of no consequence here; it is the sound and its "wave-form" that are of interest).

The human voice singing sustained vowel sounds on notes of definite pitch is the cheapest source of "complex" musical

/sounds.....

sounds if it is desired to show the "waveforms" of these on the screen of the oscilloscope. If pupils or teachers can make string or wind instruments available, so much the better. For the success of this demonstration, it is important that the oscilloscope should be able to give a stable display witnout too much knob-twiddling on the part of the demonstrator. An oscilloscope with good synchronization or, better still, good "triggering" is thus needed. Fortunately oscilloscopes specially designed for school use are now available with automatic "triggering" in the frequency range of interest (i.e. below 30 kHz).

3.18.4 Sound, the ear and hearing

The correlation between pitch and loudness with frequency and amplitude respectively should be demonstrated with the aid of the oscilloscope and also, in the case of the pitch-frequency correspondence, by mechanical means; Savart's wheel and König's siren are the familiar items of apparatus for the latter demonstration. (Note that since "intensity" is a quantity as yet undefined for pupils, it is not used here as the aspect of the stimulus which correlates with loudness. "Amplitude" is easy to define with reference to a "displacement"-time or pressure-time curve.) Care should be taken not to use interchangeably terms describing measurable properties of the stimulus and those describing aspects of the sensation. What, incidentally, is one to make of the following suggestion (from the Orange Free State Teachers' Guide)? "Use oscilloscope to show that pitch depends on frequency and intensity (loudness) on amplitude."

There is very little that a teacher can say to a pupil about the relationship between pitch and frequency and between loudness and amplitude beyond that as the frequency

/increases.....

increases, the pitch rises, and that as the amplitude increases, the sound gets louder. To the musically-inclined pupil a musically-inclined teacher could (and no doubt would) explain the relationship between musical intervals and frequency ratios, but this should not be regarded as being for general consumption. The range of frequencies over which the ear can hear can be quoted in the usual round figures of "20 to 20,000 Hz", although the present writer has yet to hear of or come across anybody capable of hearing sounds of frequency 20,000 Hz. (With a loudspeaker and an audio-oscillator it is possible for pupils to conduct crude "audibility range" tests for themselves). The teacher who favours the "believe-it-or-not" approach can tell his pupils about the enormous range of amplitudes to which the ear is capable of responding (something like 10^7 : 1 between the threshold of pain and the threshold of audibility), though he will be able to offer no evidence for his astounding claims!

Pupils will inevitably ask how the ear works, and certainly this is something that every physics teacher ought to know something about. The chief point of interest here as far as the physics of the ear is concerned is the action of the middle ear; if one thinks about this at all, one might at first be inclined to wonder why sound has to be conducted to the cochlea via the eardrum and the ossicular chain instead of going straight from the air to the oval window. In technical terms, the middle ear acts as a transformer which matches the acoustic impedance of the air to that of the cochlear fluid; in terms which a pupil might understand, the middle ear is a lever system which ensures that as much as possible of the energy entering the ear is transferred to the cochlear fluid. As to what happens inside the cochlea, the less said about this at this stage the better. The description given in the new Joint

/Matriculation.....

Matriculation Board Teachers' Guide for Physical Science (Prinsloo et al 1968), should suffice: "The cochlea contains a membrane to which the nerve endings of the auditory nerve are attached...The vibrations stimulate the nerve endings attached to it, and signals travel along from these to the brain via the auditory nerve." (We have taken the liberty of improving the wording of the second sentence.) Incidentally the statement that "Different tones affect different sections of the membrane," which appears in the same section of the Guide, is an oversimplification; while localization of tones on the membrane is quite specific for higher frequencies, this is not the case for lower frequencies.

3.18.5 Resonance

The phenomenon of resonance, which was the basis of the two demonstrations of pure tones with tuning forks to which we referred above, might receive a little attention in view of its great importance. It has been supposed that vibrations as such would not be studied in this part of the course, so that any explanation of resonance must needs be very simple. It would have to take as its starting-point the fact that something that is set vibrating and left to itself would do so with a definite frequency (the possibility of more than one mode of vibration need not be stressed for the moment). If one tries to force it to vibrate by pushing and pulling on it with some frequency, it will vibrate most easily if the frequency at which it is "driven" is the same as the frequency at which it would vibrate if left to itself. This can very easily be demonstrated with a weight hanger dangling from a rubber band. If one moves the upper end of the band up and down at the right frequency, resonance is so pronounced that the hanger flies off the band.

/Although.....

Although superposition of sound waves, standing waves, beats and resonance of air columns and determination of wavelength are also included within double square brackets in the official Common Basic Syllabus, we feel along with those responsible for the junior syllabuses that these topics are better dealt with in the senior years. Not all of them are essential to a school course, but pupils should at least have some experience and understanding of standing waves, in view of their relevance to the problems of modern physics.

3.19 Light

3.19.1 Why teach light?

One of the aims of teaching optics to pupils is to try to give them an insight into the nature of light. Such an insight has proved valuable to physicists not just for its own sake but also because it has yielded important clues to the way in which they could solve the puzzles of the microscopic world. The quantization of radiant energy led physicists to formulate theories according to which the energies of microscopic systems would likewise be quantized; and the most satisfactory of those theories embodies a description of the behaviour of microscopic "particles" which endows them with properties which are in part wavelike. Thus the "duality" of the nature of light finds a parallel in a corresponding "duality" in the nature of matter. Ideally, the teaching of light should point the way to these exciting developments in a manner which is itself interesting and exciting.

3.19.2 Introduction to geometrical optics: P.S.S.C. and Nuffield approaches

Unavoidably, this teaching must begin with geometrical optics. Very likely the teacher will tackle this aspect of his task handicapped by the distaste for geometrical

/optics.....

optics which his own teacher and more especially his university lecturers will have succeeded, intentionally or otherwise, in communicating to him. Yet he knows the job must be done; and so he and his pupils plod their way through the traditional sequence of events, beginning with experimental proof of the approximately rectilinear propagation of light and some of its consequences (the formation of shadows and eclipses), going on to the laws of reflection and refraction and thence to the behaviour of mirrors and lenses and to the principles of optical instruments. This approach is unassailable on purely logical grounds, but does not usually inspire much in the way of enthusiasm for the subject. There are perhaps two main reasons for this: the topic is covered in too much detail and the customary approach is too formal. It is to be hoped that the incorporation of most of the traditional material on geometrical optics in the square-bracketed sections of the new Common Basic Syllabus for Physical Science will result in an improvement in both these respects. In its traditional form, geometrical optics was first-rate examination material for examinations of the traditional type with its laws and many definitions to be learnt, its ray diagrams to be drawn, its experiments to be described and its "plug-in" problems to be worked out. Relegation to double-square-bracket status means that the emphasis on this work in examinations will be much less than before; consequently there will be less temptation for the teacher to cover the subject formally and in great detail. It should also be easier for him to remember that his main concern is with laying a foundation for the understanding of the nature of light.

These improvements do not alone ensure that geometrical optics will be found more attractive by the pupil. For advanced pupils, the kind of approach adopted in the P.S.S.C.

/course.....

course might prove successful in this respect; in outline, the topic is treated in much the usual manner but not as extensively, but the underlying physics is constantly kept in view and the pupil should receive the impression that he is studying what light does in order to try to find out what it is. In particular the Laws of Reflection and Refraction are carefully considered and a particle model for light is advanced to explain them. Finally it is shown that while a simple particle model can be used to account for reflection, it does not account satisfactorily for the refraction of light, since it predicts that light should travel faster instead of slower in the optically denser medium.

All this is good training in experimentation and the deduction of scientific laws from it and in the making and testing of physical theory. Given the right kind of stage management by the teacher it could, one hopes, achieve its aims. But for this to happen the right mental attitude on the part of the pupils is also needed. The P.S.S.C. expect this attitude from pupils in their seventeenth year; can it reasonably be expected from pupils some three or four years younger? The Nuffield Guide (III) says emphatically that it cannot, and that for the pupils with whom its course is concerned (in their fourteenth year) the P.S.S.C. approach is still much too formal. According to the Guide, pupils of this age will not feel any urge to discover any laws about the behaviour of light until their experience of the uses to which light can be put leads them to believe that such laws might be helpful. Therefore the Nuffield course has its pupils in Year III begin their acquaintance with optics by playing with pinhole cameras, magnifying glasses and simple telescopes and microscopes, and so receive an introduction to rays and images. Only after the pupils

/have.....

have spent a good deal of time in experimentation with lenses and mirrors are the phenomena of reflection and refraction isolated for special study. Then pupils find out about the laws of reflection and "very fast" pupils hear about Snell's Law of Refraction. For those who do not come into the latter category, the facts of refraction are kept qualitative; in the words of the Guide: "The algebraic form of the law is not essential to the discussion of the 'crucial experiment' to distinguish between waves and bullets." (For a discussion of whether this experiment really is "crucial", see Section 8.2). In this course, pupils consider both the "wave and the "bullet" theories of light; as in the P.S.S.C. course, they learn that if light were "bullets", it would have to travel faster in the denser medium, whereas if it were waves, it would have to travel slower. The suggestion is made that if facilities are available, and if pupils like some drill, refraction of waves can be demonstrated with a small "army". This demonstration can be used to show total internal reflection and focussing by a converging lens.

It is easy to imagine pupils enjoying the Nuffield approach, and one wishes it well. One would be mistaken in criticizing it on logical grounds, since the preliminary work on lenses and the formation of images is well supported by observations on the behaviour of "ray streaks". Whether the approach as such will gain wide acceptance remains to be seen; the new South African syllabuses, for example, still favour the conventional outline of presentation. In the meantime, one can benefit by taking note of any lessons that the Nuffield approach may teach. Chief among these is perhaps the thoroughness with which the subject of optical

/images.....

images is treated. It is easy for pupils to gain a superficial understanding of virtual and real images (and, of course, easy for them to learn their standard definitions) while still retaining a feeling of insecurity about them. This is particularly true of their understanding of virtual images. Whether the teaching of geometrical optics is to follow conventional lines or not, there is much to be said for introducing early in the course a demonstration of the formation of real images and of virtual images with the aid of lenses and a ray-box, accompanied by a careful discussion of how we see things. This could help to remove some of the mystery and misunderstanding that needlessly surrounds the subject of optical images. It is astonishing how some misconceptions of this sort can persist, even in the most unlikely places. Recently the present writer had to spend several minutes convincing an able physics teacher of many years' experience that one really could <u>see</u> a real image even when the latter was not formed on a screen; apparently he had not seriously considered this possibility until a question in a Joint Matriculation Board examination brought it to the fore. If pupils are properly instructed on this point, there is no reason why they should ever gain this curious impression. As for virtual images, the problem here is really to convince the pupil that the eye must see an image at the point (or region) from which the rays appear to diverge. He would understand this most clearly if he knew that the eye "sees" by virtue of real images formed on the retina. Thus a pair of diverging rays arriving at the eye and focused on the retina must give rise to an image there whether they diverge from a real or a virtual source; and this image is interpreted as being the image of something situated at the real or apparent point from which the incident rays diverge. It would not be difficult for a teacher to

/demonstrate.....

demonstrate this; all that is required is a luminous object, a pair of converging lenses and a ray-box. In the first stage of the demonstration, lens A is used to form an image of the luminous object on a screen. In the second stage, the distance of lens A from the screen is kept the same as before, the luminous object is moved forward and lens B is interpolated between it and lens A in such a position that an image is once again formed on the screen. Between the lenses, the rays are seen to form a similar pattern to that which they formed when coming straight from the object to lens A. Lens A has clearly treated the rays reaching it as though they had come from where the object originally was, which is now the position in which B forms a virtual image of the object. We do not suggest that this full-scale demonstration of the formation of virtual images be given at the outset. It should be treated as the clinching demonstration in a sequence of hints and arguments all pointing in the same direction.

3.19.3 Errors to be avoided in the teaching of geometrical optics

By its very nature, geometrical optics is not a subject in which many logical pitfalls lurk, so that it is not necessary to issue a list of warnings concerning such pitfalls. The teacher may, however, need to be warned to avoid certain errors of omission or commission. In teaching refraction, for example, he should make it quite clear that some reflection always takes place at the surface of separation of two media and that "total internal reflection" represents the limiting case of such reflection. If he teaches the subject by way of experiment and demonstration, then this point should give no trouble; taught "from the book" it might well do. It would be as well for him to emphasize that when he draws diagrams showing refraction in

/wnich.....

which reflected rays do not appear, the latter are omitted simply for the sake of clarity. Warren (1965) has warned against the uncritical use of the figure of 25 cm as "the least distance of distinct vision", pointing out that this is clearly unrealistic for an average pupil of normal sight. It is best to treat this figure as the Nuffield Guide (III) suggests, i.e. as a kind or "official" least distance of distinct vision that is applicable to adults of normal sight. In drawing diagrams illustrating the formation of images with microscopes or telescopes, care should be taken not to show the rays performing peculiar unmotivated changes of direction as they pass through a real image (cf. Prinsloo et al 1965, in which the action of a microscope is illustrated in this way without a word being said in its defence). The explanation usually given of the way in which a telescope produces a real image in the focal plane of its objective, leaves much to be desired. Typically, the diagram shows rays incident on the objective parallel to one another and at an angle the axis of the telescope. The rays converge on a point in the focal plane, which is shown as the head of the image of the distant heavenly object; the foot of the image is shown as resting on the axis. It is not usually made clear by the textbooks in which such diagrams appear, that the rays shown incident at an angle originate from the "top" of the object, and that rays originating from the "bottom" of the object (which is assumed to be "standing" on the axis) and which are not shown will reach the objective as a set of rays parallel to the axis. If this is not explained to pupils, they will be needlessly puzzled about the working of the telescope. Incidentally, in some textbooks (e.g. Prinsloo et al 1965) the distant object which gives rise to a finite-sized image in the focal plane of the objective is cheerfully described as a "star". One presumes the authors had the sun in mind!

/3.19.4.....

3.19.4 The laws of reflection and refraction

In South African syllabuses, it seems to have been accepted that pupils will study reflection and refraction as such in tne junior secondary school. Since, in fact, there are not many laws included in South African junior physical science syllabuses, it is not unreasonable that pupils in this category should study the laws of reflection and refraction. In order that they should find this interesting, it is desirable that pupils should come to know them by a process akin to self-discovery. The observation that the angles of incidence and reflection are equal ought not to present any difficulty to pupils; it would be obvious even in a demonstration experiment. The fact that incident and reflected rays and the normal all lie in the same plane is, in its way, obvious enough too, although the terminology in which it is conventionally expressed might not be. In drawing attention to this law, teachers should emphasize that while it is simple and obvious, it is by no means trivial; some hint might be given of what might happen if incident and reflected rays and the normal are not coplanar. In regard to Snell's Law, the matter is less straightforward. It is safe to say that no standard Eight pupil, however bright, is likely to hit on this Law if left to find unaided the relationship between the angles of incidence and refraction. As has been pointed out (e.g. in the P.S.S.C. textbook), data on refraction were known and used for a thousand years before "Snell's Law" was discovered independently by Snell and Descartes. Thus, while the teacher should see that this law is "discovered" by the pupil from his own data, he will have to give strong hints converning the way in which the process of "discovery" should continue: "Is the ratio of i to r constant?... No? Try the ratios of the tangents, or the cosines or the sines." No doubt this sort of

/"discovery".....

"discovery" is highly artificial. But it does give the pupil a little insight into the nature of the dilemma which physicists have often faced in the past, and which they are likely to have to face again, viz. the reduction of a set of data to a simple, summarizing statement - a "Law", in fact. In this case the "Law" is far from obvious, and the pupil is thus enabled to gain a little appreciation of the need in such cases to try out various possibilities in the hope that one may fit. We have quoted the Nuffield Guide (III) to the effect that the precise form of Snell's Law is not essential to the discussion of the experiment in which the velocity of light in a "dense" medium is determined in order supposedly to decide between "wave" and "particle" theories of light. We think, however, that the discussion of this experiment would gain in force and significance if it were to take the form of a comparison between the precise predictions of a theory and the precise results of an experiment. This comparison need not be studied until the Matriculation years, but it would be convenient to have the concept of a "refractive index" available for it when the time comes.

Some teachers might wish to deduce Snell's Law from Huygens's construction applied to wavefronts arriving at and leaving the interface between the two "media". This raises the question of whether Huygens's construction should be used at all in elementary teaching and, if it is, to what extent. In considering the first of these points, one needs to be clear about the status of Huygens's Principle on which the construction is based. Contrary to what is sometimes supposed, it is not a fundamental physical principle of general validity. Even when it is applied to the simplest of situations in wave propagation, such as the propagation of plane or spherical wavefronts, it is necessary to add

/"ad hoc".....

"ad hoc" assumptions to it in order to account for the absence of a "back wave". Once this is done, there is no doubt that Huygens's construction or Fresnel's extension of it can be used to predict many of the standard results of the wave theory in a very simple way, and because of this, one has the feeling that there is something right about it. But, as many modern critics have pointed out, Huygens's Principle rests on insecure foundations; at best, "it is an artificial scheme that physicists have come to accept by repeated use" (Nuffield Guide III). We consider the difficulties attached to the Principle in Section 6.2.4-5, where we also show that it is possible to derive Snell's Law for waves on water by a method which seems to resemble Huygens's construction but which is not open to the same objections. We do not think that there is much to be gained by the use of Huygens's Principle in the treatment of wave motion in the junior secondary school. If, as is asserted in the Nuffield Guide (III), it is "far from obvious to pupils," and if it is known to be a "convenient mathematical device" (Mendoza 1965) rather than a genuine physical principle, the teacher is probably better advised not to use it - at any rate in the earlier stages of teaching wave motion and optics.

(The concept of "refractive index" will also be convenient for a partial interpretation of the dispersion of light, which should be qualitatively studied by pupils at this stage as an introduction both to colour and to the spectrum of white light.)

Quantitative work on refraction by <u>lenses</u> should, we suggest, be kept to a minimum. It is easy to show the general rules relating the position, magnification and nature of the image to the distance of the object from a lens; pupils should certainly experience all this and learn the principles applicable to the drawing of the appropriate

/diagrams.....

diagrams. They should learn about the focal length of the lens and learn the "distant object" method of finding it for a converging lens. But it is not necessary for them to learn about " $\frac{1}{u} + \frac{1}{v} = \frac{1}{f}$ " (or whatever form of this equation the reader favours), with its attendant complications of sign conventions and whether or not one should provide a "proof" of the formula and if so, what kind of proof.

3.19.5 Treatment of the nature of light in the junior secondary school

In order that the discussion of the nature of light should make its proper impact in the matriculation years, it is desirable that any treatment of this topic in the junior secondary school should be very carefully considered. The full argument in favour of the view that light might be a form of wave motion is based on observations conducted on a variety of phenomena, only some of which are covered in the junior years. Certainly there is no harm in inviting pupils to consider the "wave" and "bullet" theories of light and to think of possible arguments for and against these theories; but to bring them down firmly on one side of the fence would be premature at this stage. A curious entry in the Orange Free State syllabus for Standards Seven and Eight reads: "Light as transverse waves...Use ripple tank [i.e. "waves on water in a ripple tank"] as analogy." Thus not only is the view put to pupils on insufficient evidence that light is a form of wave motion, but they are required also to recognize the transverse character of light waves. This is going much too far in this context, and one can only wonder whether the planners of this syllabus really intended this.

/One.....

One final point needs to be made. An underlying assumption in any standard treatment of optics is that light is something which travels. It would be just as well to emphasize this early in the discussion since the point sometimes seems to escape pupils. The present writer once had to coach a first-year university student who, until otherwise informed, had imagined that light was just "there"; the thought that energy (or anything, for that matter) might be travelling outwards from the source was completely new to her. A very elementary discussion of the possibility of measuring the speed with which light travels would be a useful interpolation in the course, especially since most pupils will know of an approximate value for the speed of light anyway. The principle of, say, Fizeau's method of measuring this quantity is so simple that it could well be treated in class. As has been pointed out by the P.S.S.C. authors, the disc type of stroboscope could, in principle, be used as the "light chopper" in Fizeau's experimental arrangement.

The speed of light in vacuo is such an important quantity in present-day physics that it cannot safely be neglected in any high-school treatment of the subject. It seems to us that this topic could conveniently be dealt with in the junior secondary school.

CHAPTER 4: CLASSICAL MECHANICS IN THE SENIOR SECONDARY SCHOOL

4.1 Introduction

The importance of careful teaching of the fundamental principles of classical mechanics cannot be over-emphasized. Not only do these principles form the basis of so much that the pupil is to study later in the course; their introduction provides a wonderful opportunity for the teacher to show how physical concepts are invented in order to facilitate the analysis of physical problems. If the pupil is to gain a proper idea of what physics really is all about, this opportunity will have to be seized and properly exploited. For example, it is possible to teach the subject of momentum and its conservation by defining the concept of momentum without preamble and proceeding with the aid of Newton's Third Law of Motion to deduce the conservation of momentum. In this way all the requirements of syllabus and examination can be met. But to satisfy the need of the pupil to develop a sense of the way in which concepts evolve, it is surely better to allow the concept of momentum to emerge naturally from, say, a properly-conducted discussion of collisions.

Like other branches of physics, classical mechanics should be regarded by the teacher as a means of stimulating in the pupil an inclination to think about the world around him in scientific terms as well as in other ways. So much of what the male pupil finds exciting today has its origin in the direct application of classical mechanics, and the subject itself owes its origin to man's desire to give a rational account of the motions of heavenly bodies. If the teacher can keep these points in mind when presenting the principles of mechanics, he can do a good deal towards making his pupils realize that the appreciation of science is as

/natural.....

natural and healthy an activity as the appreciation of the arts.

4.2 Vectors and scalars, with special reference to the vector nature of displacement and velocity

Pupils will have gained some familiarity with elementary ideas about motion and force prior to reaching Standard Nine. Any advance on such elementary and largely qualitative treatment of motion requires explicit use of the vector nature of the concepts listed above. Fortunately it is quite easily possible for a pupil to gain a "first-order" understanding of the nature of a vector quantity, which will be fully adequate for the requirements of a school syllabus. A "higher-order" understanding of vectors, which involves a knowledge of the transformation laws obeyed by vectors and of the difference between polar and axial vectors, can be left to the university to impart; since it is not absolutely essential for a teacher to have it in order to teach the subject properly, we will not elaborate on it here, although we will briefly discuss polar and axial vectors later with reference to the magnetic vector $\vec{B}$ (cf. Section 7.15.5).

Discussion of an imaginary trip undertaken from point A to point B via point C which does not lie on the straight line AB should help (a) to introduce the pupil to the concept of displacement and to show that it has some use, and (b) to introduce the idea of a vector quantity and the method of adding vectors. The vectorial nature of displacement can be contrasted with the nature of other physical quantities previously encountered such as length, time, mass, area, volume and density which are not directional and are known as scalars.

In connection with vectors, the Common Basic Syllabuses specify the topics "addition of vectors by the parallelogram

/law.....

law, graphically; resolution of vectors, graphically into components at any angle and quantitatively into rectangular components." The most important topic here is the quantitative resolution of vectors into rectangular components; it is essential for later work that the very simple sine and cosine formulae for the components of a vector should be absorbed by the pupil. A small terminological point arises here which perhaps ought to be cleared out of the way. In the literature and in teaching, the word "component" is used in two ways; in one, it refers to the magnitude of the component vector in the specified direction (cf Arons 1965); in the other, it refers to the component-vector itself (cf. Spiegel 1959). In our opinion, it would be safer to stick to the latter alternative. It is easy to introduce components by saying that they are vectors which add up to the original vector and are therefore equivalent to it. The teacher can always refer to the magnitudes of components if he wishes to.

Undoubtedly it is highly desirable to use some directly <u>physical</u> approach in formally introducing the concept of <u>velocity</u> (as opposed to speed) and making clear its vector nature. A good way of doing this is to use the P.S.S.C. approach of discussing the problem of how to find the speed of an aeroplane relative to the ground when there is an appreciable wind. Many boys would know that the airspeed indicator of an aeroplane measures its speed relative to the surrounding air and would therefore appreciate that there <u>is</u> a problem. They would see easily enough that if the aeroplane and the wind had the same direction, the speeds would have to be added to give the required answer, and that if the two directions were opposite, the speed of the wind would have to be subtracted from that of the aeroplane. What if the wind were at right angles to the direction of the

/aeroplane?

aeroplane? This case provides the opportunity for the teacher to introduce the idea of "velocity" as a vector quantity.

It will be noted that it has been assumed in the above discussion that the concept of speed is understood to at least some extent by the pupil. Certainly this ought to be true of boy pupils, who are likely to be interested in cars, rockets and aeroplanes and at some stage to have formed the idea that if the speedometer of a car reads "60 miles an hour", this means that "if the car keeps going at this speed, it will travel 60 miles in one hour." On this kind of intuitive understanding of the concepts of speed and, in particular, constant speed, it is possible to build the foregoing discussion of what are essentially constant speeds and velocities. There is no need as yet to introduce instantaneous velocity.

By introducing the idea of "trips" (displacements) made in given times, one can justify the fact that velocities add vectorially. A point worth making even in an elementary discussion of vectors, is that the same answer is obtained whether one adds vectors "simultaneously" (as in the parallelogram method) or "successively" (resultant completes thetriangle), and that in the latter case the order of addition is immaterial. This is easily justified by means of the thought- (or actual) experiment of a man going for a short trip on a truck across which he proceeds to walk.

If velocity is introduced by way of the type of example we have suggested (velocity of an aeroplane relative to the ground), the idea that velocity is specified <u>relative</u> to something is made explicit. This is the first hint of the concept of a <u>frame of reference</u>, which is so important both to classical mechanics and to special relativity. As yet, the term need not be used, but if the pupil is properly aware

/that.....

that the specification of a velocity implies a system relative to which that velocity is measured, he will have grasped a most important point. We discuss frames of reference as such in Section 4.6.

In discussing both displacement and velocity, we have implied that their manner of addition should be justified in each case. We suggest that the teacher should take it as a matter of general principle that whenever he has to introduce a new vector quantity, he should regard it as necessary to show that it satisfies the laws of vector addition. Sometimes this may be obvious from the way in which it is defined (e.g. if it is defined as a scalar x known vector, or $\frac{\text{known vector}}{\text{scalar}}$); in other cases it will not be, and then pupils should not simply be expected to take its vector nature for granted. Pupils should become thoroughly familiar with the rules that a vector multiplied by a scalar yields a vector and that a vector divided by a scalar yields a vector. This is very easy to prove if the "scalar" is in fact a pure number; if it is any scalar quantity, then the result is easy enough to argue in any particular case (e.g. velocity as $\frac{\text{displacement}}{\text{time}}$).

4.3 The concept of acceleration

In an earlier section, (2.6) we referred to the possibility of introducing the concept of acceleration through a class discussion of how one might specify a suitable measure for an as yet undefined "acceleration", and we asserted there that given the chance, pupils might come forward with suggestions which were equivalent to $\frac{dv}{ds}$ and $\frac{dv}{dt}$. If the teacher is not pressed for time, he might well try this approach to acceleration; he could then guide pupils to making the choice of "$\frac{dv}{dt}$" (some of them will be used to the idea of

/this.....

this already from advertisements for motor cars or from reports on road tests), pointing out to them that this choice will turn out to be highly convenient later. Indeed, one aspect of its convenience will soon appear, when it is found that the motion of freely falling bodies is characterized by constant acceleration, where acceleration is defined as "$\frac{dv}{dt}$".

The vector nature of acceleration needs to be emphasized; that is, pupils need to be reminded from time to time that acceleration is defined as increase in velocity per unit time and not increase in speed. Since pupils may tend to miss the point of this if their studies in kinematics are confined to purely rectilinear motion, they should have other types of motion discussed with them also, so that they may appreciate that if a body is travelling in a certain direction now and in a different direction later, it has undergone an acceleration whether its speed has changed or not. Again, the teacher should assure his pupils that it is no mere caprice on the part of physicists which makes them define acceleration in this way, and that the full advantages of this definition will appear in due course.

4.4 Rectilinear, uniformly accelerated motion

4.4.1 Some general considerations

It is a well-established practice to preface a study of the principles of classical mechanics with a fairly detailed study of rectilinear, uniformly accelerated motion. This is an essential first step towards a quantitative understanding of motion and it is no surprise to find it accepted as such in present-day school syllabuses, including South African ones. In such an analysis of motion, the concepts of instantaneous velocity and acceleration are implicit.

/Should.....

Should they be explicitly taught at this stage? They are in the P.S.S.C. course, and in the Teachers' Guide to the New Common Basic Syllabus in Physical Science (Prinsloo et al, 1968), the present writer has gone so far as to say the following: "In this ... study of linear motion it is of course inevitable that average velocity and acceleration should first be defined, followed by the use of the concept of a limit to define instantaneous velocity and acceleration. 'Limits' in this sense are fundamental to any but the most naive analyses of motion, and should be carefully explained." This view has not been challenged in South Africa, but in the Nuffield Physics Course, a contrary attitude seems, by implication, to have been adopted (cf. Nuffield Guide (IV)). That is, while the principle of starting with the definitions of average velocity and acceleration has been adopted, the word "average" is not officially used in connection with these quantities, and the idea of instantaneous velocity as a limiting value for average velocity in an interval is not introduced at all. (The nearest the Guide comes to discussing this sort of idea is in its remarks about approximating the area under a curve by means of a series of vertical rectangular strips; in referring to taking the limit in such circumstances it says that this is "probably very wise mathematical teaching; but it will probably spoil our present teaching of physics entirely if we labour it.") Despite this, the idea of "velocity at an instant" is explicitly used, as indeed it has to be if the well-known equations of rectilinear motion are to be given any meaning. As the Nuffield Year IV is concerned with pupils aged fourteen or fifteen, the reluctance of the planners of the course to deal explicitly with limits and their apparent willingness to treat instantaneous velocity as intuitively understandable can be readily understood. With pupils in South African senior

/secondary.....

secondary schools the situation is different; they will be at least sixteen years of age and should have some experience of graph drawing. For several years now, pupils writing the Joint Matriculation Board examinations in physics have been required to approach velocity and acceleration via the use of averages and limits, and their teachers have not yet complained. We shall proceed on the assumption that use of these ideas will not "spoil our teaching" but help in achieving its aims. The Common Basic Syllabus for Physical Science does not specify the use of the equations of uniformly accelerated, rectilinear motion, but instead asks for treatment of such motion by means of graphs of displacement, velocity and acceleration against time. The introduction of these graphs is certainly excellent training for the pupil, for from them he can learn that

(i) the slope of a displacement-time graph at any point gives the velocity at that point, so that it is possible to construct a velocity-time graph from a displacement-time graph;

(ii) the area under a velocity-time graph corresponding to a given time interval is the displacement undergone in that time interval;

(iii) the slope of a velocity-time graph at any point gives the acceleration at that point, so that an acceleration-time graph can be constructed from a velocity-time graph; and

(iv) that the area under an acceleration-time graph corresponding to a given time interval is the increase in velocity in that interval.

The pupil should also make the acquaintance of velocity- and acceleration-time graphs in which negative velocities and accelerations are plotted. The interpretation of the areas under the "negative" portions of the curves as <u>decreases</u> in displacement or velocity should also be made clear to him.

/Since.....

Since the study of these graphs brings the pupil so close to "$v = u+at$" and "$s = ut + \frac{1}{2} at^2$", he might as well see these equations derived from the graphs and go on to derive "$v^2 = u^2 + 2as$" from the equations. This is not for the purpose of giving the pupil a few extra formulae to memorize for the examination (we fully support Rogers's suggestion (1962) that these and other formulae should be printed on the examination paper), but simply to make matters easier for pupil and teacher when certain important topics are discussed. For example, if the formula $s = ut+\frac{1}{2} at^2$ is not available for use by the pupil (and it should not be used by him unless some attempt can be made to derive it for him), then it would not be possible for him to do an experiment to find the value of "g" by timing a falling object. (See below). And "$v^2 = u^2 + 2as$" is an indispensable relation when kinetic and potential energy and the conservation of their sum are under discussion. Therefore, even if a syllabus does not specify these formulae (the Common Basic Syllabus for Physical Science does not), the teacher should still regard it as his duty to see that they are derived in class, although his pupils will not be expected to make use of them in examinations.

4.4.2 Experimental work

It is highly important that this theoretical work on rectilinear motion should receive experimental support. The function of this experimental support must be carefully considered: we cannot prove the formulae for rectilinear motion experimentally, because they are logical consequences of the definitions of displacement, velocity and (uniform) acceleration. But we can conduct experiments to determine whether the motion of an object under given conditions is uniformly accelerated or not. The pupil must be able to

/recognize.....

recognize constantly accelerated motion when he comes across it, since motion of this type will play a central role in his experimental approach to the laws of mechanics. There are various ways in which he can be shown how to detect constant acceleration. If he has initially been given an entirely theoretical treatment of rectilinear motion, then there is no problem at all: it is very easy to show, either graphically or algebraically, that displacements in successive equal time intervals increase by a constant amount, which is the criterion by which he can recognize uniformly accelerated motion from measurements on a ticker tape. The teacher might wish to make the same point directly from experiment, and follow this up with the theoretical treatment. In this case he could arrange for a trolley to be pulled along by a constant force (not to be identified as such at this stage) and for marks to be made at regular time intervals on a length of ticker-tape attached to the trolley. Perhaps the easiest arrangement for introductory purposes would be to have the trolley run down an inclined plane, so that "diluted gravity" (the term comes from the Nuffield Guide (III)) can provide the constant force. The ticker-tape provides a record of the displacements undergone by the trolley in successive time intervals. The quickest way for the pupil to infer constancy of acceleration from his data would be for him to draw up a time-displacement table, consisting of five columns with the following headings:-

TIME	OVERALL DISPLACEMENT	DISPLACEMENT DURING THE INTERVAL	AVERAGE VELOCITY DURING THE INTERVAL	AVERAGE ACCELERATION DURING THE INTERVAL

The data obtained directly from the ticker-tape would be recorded in the first two columns. (For standard Nine pupils we would recommend that the vibrator used for marking the tape should have the interval between its "ticks" calibrated

/in.....

in terms of the second; this interval should preferably be adjusted to be equal to some convenient fraction of a second.) From them, the data to be entered in the third and fourth columns of the table could easily be deduced. If one assumes that the average velocity in each successive time interval is equal to the actual velocity in the middle of the interval - a justifiable assumption for this type of motion - one could plot a velocity-time graph from the data in the fourth column (or just compare successive differences between pairs of numbers in this column). All being well, the graph would be a straight line (and the successive differences constant), showing the constancy of the acceleration. In the fifth column, the calculated values of the average acceleration in the appropriate time intervals could be recorded (they would simply be equal to the differences of successive pairs of numbers in the previous column divided by the time interval). It is worth pointing out here that if the values of the average acceleration in the last column were added up and divided by the appropriate number to yield a mean, this would not represent the best possible use of the data. Suppose that five values of the displacement are recorded in column (2), and that these are labelled S_1 to S_5. Then the displacements in four successive time intervals are respectively S_2-S_1, S_3-S_2, S_4-S_3, and S_5-S_4, or ΔS_1, ΔS_2, ΔS_3 and ΔS_4. To find the average acceleration in three successive time intervals means using the differences $\Delta S_2-\Delta S_1$, $\Delta S_3-\Delta S_2$ and $\Delta S_4-\Delta S_3$. If the values of average acceleration obtained from these differences are added up in order to find a mean, then we are left in effect with only ΔS_4 and ΔS_1 or, in terms of the original displacements, S_1, S_2, S_4 and S_5. We do not suggest that this is of earth-shaking consequence in the present context, but if the teacher wants his pupils to use their data to best advantage in some

/determination.....

determination of an acceleration, he should introduce them to the technique of using overlapping pairs of data. A detailed discussion of this point has recently been given by Lanni (1968).

The Nuffield Guide suggests another way of presenting the data obtained from a ticker-tape in such an experiment which, if there were sufficient time available, would undoubtedly prove instructive. This is to make a "tape-chart" consisting of pieces of the tape representing displacements in successive time intervals pasted up alongside one another, all starting from the same base-line. The Nuffield Guide (III) suggests that "the pasted tapes have a strange virtue for beginners, of insisting on the physical nature of the chart." Of course the line passing through the tops of the strips of tape will be in effect a velocity-time graph.

As part of their experimental study of uniformly accelerated motion, pupils should carry out observations on a freely falling body and deduce the value of "g". There are two main reasons for doing this: (i) "free fall" is Nature's own example of uniformly accelerated motion; and (ii) "g" is a quantity of some intrinsic interest. To establish (i), it is clearly insufficient simply to find the time taken for the object to fall from rest through a given distance, plug the values of distance and time into the appropriate formula and hence calculate "g". The pupil must be convinced that the formula really is appropriate to the situation. For this to happen, the experiment should be done for several different heights of fall, so that it can be shown that the formula yields the same value for "g" for each height - which it should do, if the acceleration were in truth constant. This kind of experiment can nowadays be done by timing the fall of the object with a demonstration scaler set to count

/pulses.....

pulses from an oscillator (e.g. a Leybold scaler, whose "clock" ticks 10,000 times a second, or a Panax scaler, for which the oscillator frequency is 10^3 Hz.). For the justification of the use of this "black box" in this experiment, see our earlier section (2.4) on black boxes. It is easy to arrange a circuit and auxiliary apparatus so that the release of the object (usually a large ball-bearing) starts the "clock" and its striking a target stops it. By the use of illuminated photo-diodes the time of flight of the ball between a chosen pair of levels can be measured with the scaler. This will yield results of better consistency than those obtainable from the simpler form of the experiment, but the added precision is probably not worth the added complication of the algebra required to analyse the data.

A more direct method of showing constancy of acceleration in free fall would be to take a "multiflash" photograph of a falling ball and then obtain a "time-displacement" table from measurements on the photograph. (The ball should be allowed to fall next to a vertically-mounted metre rule.) If a value of "g" is to be deduced from the data, the stroboscope will have to be calibrated. Rather than simply accept the manufacturer's calibration, teacher and pupils should carry out their own calibration with a strobodisc rotated at a known number of revolutions per second. As the interval between successive flashes of the stroboscope need not be much less than $\frac{1}{20}$ second in this experiment, this calibration ought not to present any difficulty.

An essential point to make in connection with "g" is that it is the same for all freely falling bodies. As any physicist knows, this is no trivial matter; not until the seventeenth century did Galileo establish it experimentally and back up his experiments with the kind of reasoning we

/would.....

would now find acceptable, and not until our own century did the full significance of this fact become apparent. Therefore time spent in making this clear to the pupil is time well spent, and the opportunity can also be taken to introduce some useful history; Aristotle, Galileo and Newton can make their entry here to some advantage. The tale of Galileo's alleged experiment of dropping two balls of markedly different mass from the top of the Leaning Tower of Pisa before an astonished crowd is apparently no more than a fable. Certainly it is one of those stories that ought to have been true, but there are good physical reasons for thinking that Galileo would not have attempted to demonstrate his conclusion about the free fall of bodies in this way, and in any case there are no reliable accounts of the supposed event. What <u>is</u> true beyond doubt is that Galileo did carry out "private" experiments (e.g. with "diluted gravity" acting on balls rolling down inclined planes) from the results of which he deduced that all freely falling bodies in a vacuum would fall with the same uniform acceleration. There is something to be said for acquainting pupils with the Leaning Tower of Pisa legend - it may help to fix the essential point firmly in pupils' minds - but for the sake of historical accuracy it ought to be made plain that the story is apocryphal. (For modern accounts of Galileo's work, see Cohen (1961) or Arons (1965).) If the apparatus is available, a modern version of Newton's "guinea-and-feather" experiment should certainly be done, preferably by the pupils themselves. The Nuffield Guide (III) gives a detailed account of how to organize this as an experiment to be done by individual members of a class. The teacher wishing to bring his pupils up to date on modern work in this field can refer them to the results obtained by Dicke and his co-workers

/Dicke.....

(Dicke 1961, Roll et al 1964), which have verified Galileo's conclusions to within 3 parts in 10^{11}. He can also tell them that the most accurate determination of "g" yet made employed a highly refined form of the method in which the motion of a freely falling ball is timed (see Cook 1967).

4.5 Projectile motion

The Common Basic Syllabus for Physical Science places "Projectile motion, by means of the resolution of the motion into two components" in the category of optional extensions to the syllabus; the corresponding syllabus for Physics includes the topic without strings attached. Projectile motion is a simple example of motion that is not rectilinear but is in fact uniformly accelerated; it would thus help to broaden the pupil's understanding of acceleration and bring home to him the principle of independence of motions at right angles to each other. This principle was first enunciated and experimentally checked by Galileo; today it is possible to demonstrate it by means much more direct than those that were available to him. "Multiflash" photographs of different stages of projectile motion illustrate the principle most vividly; fine examples are to be found in the P.S.S.C. textbook and elsewhere, but if the equipment is available such photographs can and should be taken in the school laboratory also. Whether or not this can be done, pupils should certainly see the demonstration known as the "Monkey and Hunter", in which a "bullet" aimed at a "monkey" and fired at the instant the latter is released will infallibly hit the "monkey"- provided the latter is not too small in relation to the bullet! (See Sutton 1938, Nuffield Guide (III)). The teacher may care to note that the principle of independence of velocities breaks down in the "relativistic" region, i.e. for velocities comparable with that of light: if the falling particle has a

/large.....

large "horizontal" velocity relative to the chosen reference frame, its vertical acceleration will be found to depend on its horizontal velocity (cf Katz 1964, Resnick 1968).

There would be no particular virtue in requiring Matriculation pupils to do numerical problems on projectile motion. Pupils should nevertheless appreciate that in the absence of air resistance, the horizontal component of the motion of a projectile is characterized by constant velocity, while the vertical component is subject to a constant downwards acceleration, viz. the acceleration due to gravity. On the basis of these facts, they should be able to gain a semi-quantitative understanding of the parabolic shape of projectile paths illustrated by multiflash photographs.

4.6 Frames of reference

Lurking behind any quantitative description of motion is the concept of a frame of reference relative to which the displacement, velocity and acceleration are specified; and fundamental to any modern account of Newtonian mechanics is the concept of an inertial frame of reference, or "inertial frame". Yet until comparatively recently these terms did not appear in school science textbooks, at any rate not in South African ones; they are not even mentioned in two textbooks published in South Africa during the last four years (Gordon et al 1964, Prinsloo et al 1965). Elsewhere the situation has altered somewhat, perhaps as a result of the work of the P.S.S.C., who give the subject prominence in their textbook and have produced an excellent film called "Frames of Reference" (P.S.S.C. - MLA 0307, reviewed by Brown (1963)) to reinforce the presentation of it. We do now find the term "frame of reference" turning up occasionally in elementary textbooks (e.g. in Jardine 1966), even though it may not yet be mentioned in syllabuses. The points to be considered

/here.....

here are (a) Should the teacher discuss the concepts of "frame of reference" and "inertial frames" with his pupils and, if so, to what extent?; and (b) whether or not the teacher does raise these points in class, how much should be known about them as background to effective teaching of mechanics? It may be that planners of future syllabuses will decide to answer the first part of (a) in the affirmative and officially prescribe the extent to which these matters should be covered. We shall work on the assumption that this has not yet entered their minds. As things stand at the moment, (b) is of more pressing concern to the teacher than (a) and we accordingly deal with it first. We do so by giving an account of the background knowledge of the subject that any competent university teacher of elementary mechanics ought to have; in our view, the schoolteacher ought to know at least much of the background as what follows if he is to teach mechanics adequately and also be able to stimulate the brighter pupils in his classes.

The concept of a "frame of reference" needs little elaboration here. Explicitly or implicitly, we use it whenever we describe the motion of a body in quantitative terms. Often we make it concrete by identifying it in our speech or thought with some concrete object or physical system; thus we speak of the velocity of an aeroplane relative to the surrounding air, the velocity of the wind relative to the ground, or the velocity of a space-probe relative to the earth. But in the back of the physicist's mind when he speaks like this will be the idea of a co-ordinate system "fixed in" whatever object or system is being spoken of as the frame of reference. By this he means no real set of very long rods somehow embedded in or attached to the object or system in question, but an imaginary set of axes relative to which the

/displacement,.....

displacement, velocity or acceleration of the body could be specified. In fact, there is nothing to stop the physicist imagining a frame of reference anywhere he likes and moving in any way he pleases relative to some other frame of reference. This is not to say that such an arbitrarily-imagined frame of reference could not, in principle, be "fixed in" some real object or system which has the appropriate motion and from which measurements could be made. In fact, one usually imagines such a frame of reference with the idea in mind that <u>if</u> one were to carry out actual measurements relative to it of certain quantities one would obtain particular values for those quantities. In this sense one can "change the velocity (say) of an object by changing one's mind", to adapt the dangerous phrase that Dingle (1940, 1956) uses in this connection, but the velocity arrived at should remain measurable in principle. One could imagine a frame of reference moving with a velocity of a hundred times that of light relative to the physical system of interest, for example, but this would be a singularly unprofitable exercise in the solution of any physical problem, simply because no physical system could travel at such a velocity relative to any material object in which a frame of reference might be fixed.

To illustrate the use that is sometimes made of "imaginary" frames of reference in the treatment of physical problems, we discuss a collision between two marbles of equal mass, one of them at rest on a table and the other moving towards it at a velocity 4 m/s due North. Regarded from a frame of reference moving at a velocity 2 m/s due North, this problem is particularly simple to describe and to solve: the marbles approach each other with equal and opposite velocities and bounce apart with equal and opposite velocities. If kinetic energy is conserved in the collision, the marbles will have the same magnitudes of velocity in the "moving"frame

/after.....

after the collision as they had before, from which one deduces that relative to the table, the marbles have simply exchanged velocities, the initially rolling one now being at rest and the other moving off with velocity 4 m/s due north. In this case the use of a frame of reference "fixed in" neither the table nor one of the marbles has somewhat simplified the solution of a physical problem.

The foregoing example illustrates an important fact about frames of reference: that for certain purposes some frames of reference may be more helpful than others. It is this fact which, as we shall see, has led to the introduction of the concept of an "inertial frame of reference" or "inertial co-ordinate systems" into modern presentations of Newtonian mechanics. Suppose we fire a ball bearing into the air from a toy cannon, and we wish to describe the subsequent behaviour of the ball in terms of the forces acting upon it, using the laws of Newtonian mechanics for the purpose. If we choose a frame of reference at rest relative to the laboratory (on earth) our task is very simple; our experience and training have led us to believe that only two forces act on the ball while it is in the air, viz. the force of gravity and the frictional resistance of the air, and if we assume this to be true and use the fact that resultant force equals mass times acceleration, we are able to arrive at a completely satisfactory description of the behaviour of the ball from the point of view of our chosen frame of reference. If our laboratory is situated instead in a "Jumbo Jet" flying at constant velocity relative to the earth, choice of a frame of reference "fixed in" the laboratory would similarly enable us to account satisfactorily for the behaviour of the ball. But if the aeroplane were accelerating relative to the earth and we persisted in using a frame of reference at rest relative to our

/laboratory, ...

laboratory, we would find that the resultant force on the ball would not be equal to its mass multiplied by its acceleration, unless we postulated the existence of an additional force on the ball, whose origin we could not attribute to any obvious agency inside or outside the laboratory. So we reach the conclusion that there is something special about frames of reference that are at rest or moving with uniform velocity relative to the earth; if we use them in our descriptions of motion, Newton's Laws seem to "work" without our having to call in unaccountable or "fictitious" forces.

We are led naturally to wonder why it is that these frames of reference are privileged in this way, and whether we can establish an independent criterion with the aid of which we can select such "special" frames. Inevitably we look first to Newton himself, as the founder of modern mechanics, for a solution to this problem. We find that in Newton's writings (cf. Motte's translation of the "Principia" rev. Cajori 1947) there is no mention of anything resembling a "frame of reference" in the sense in which this term is now used, unless we construe the statement "... it may be that there is no body really at rest, to which the places and motions of others may be referred" in this way, or if we choose to identify Newton's "relative spaces" and "absolute space" with co-ordinate systems. However, Dingle (1967a) has argued that "Although he might have done so, Newton did not introduce co-ordinate systems into his scheme. All his geometry is of the pre-Cartesian type, and he never entertained the idea of imaginary reference frames"; and if we may accept this account of Newton's attitude, we must concede that for him, "frames of reference" in the modern sense did not exist. But Newton certainly did make use of the reference "spaces" and "bodies", and so we can still ask what criteria he employed in selecting them. In particular, it

/is.....

is of interest to know what reference space or spaces he had in mind when in his First Law, Definition IV and Corollary V he refers to a "state of rest or uniform motion in a right line." Newton's observations on "absolute space" and "relative spaces" and "absolute (or "true") and "relative motions" appear to leave no doubt that when he spoke of "rest" or "uniform motion in a right line" he meant <u>absolute</u> rest or <u>absolute</u> motion in a "right line" (cf. Jammer 1954). But, as he says, "It is indeed a matter of great difficulty to discover, and effectually to distinguish, the true motions of particular bodies from the apparent; because the parts of that immovable space, in which those motions are performed, do by no means come under the observation of our senses." Nevertheless, he believed that "absolute space can be determined through the existence of centrifugal forces in rotational motion" (Jammer 1954) (cf. his famous "rotating pail" experiment); and he asserted that "the centre of the system of the world is immovable", which centre he identified with the centre of gravity of the system composed of the sun, the earth and the planets. This, then, is his "absolute space" relative to which all "true motions" take place; but, even if we accept Newton's unique "reference space", it may not seem of much help in selecting frames of reference for everyday purposes. In Newton's own treatment of dynamical problems, he made use of his idea that forces were "the causes and effects of true motions". Thus in his view, a pair of bodies moving in some manner relative to each other may each be acted upon by a <u>real</u> force, which determines their <u>true</u> motions in <u>absolute</u> space. If the motion of one relative to the other is uniform in velocity (which may be zero), the real forces are the same on each body; if instead their relative motion is variable, the real forces on each must be different and the difference between these forces determines

/their.....

their relative motion. If one chooses an "apparent space" in which one of the bodies is at rest as one's reference space, then in this space an apparent force must act on the other body which is the difference between the real forces acting on each of the bodies. Evidently any "apparent space" may be chosen as a reference space and Newton does not appear to single out any class of such "apparent spaces" as being better than others for this purpose. It is interesting to note here that, in accordance with what one might expect from the foregoing reasoning, Newton's writings make it plain that he does not place centrifugal forces in a special category of "fictitious" forces, as is now common practice in modern presentations of Newtonian mechanics.

Thus Newton's own formulation of mechanics neither introduces the concept of an inertial frame of reference nor gives us any direct help in selecting frames of reference relative to which it is possible to give simple quantitative descriptions of the motions of bodies. The first explicit references to an "inertial frame" or an "inertial system" in scientific literature occur in the writings of Ludwig Lange (1885, 1886, commented on by Mach (1902) and Jammer (1954)). Lange's "inertial system" is set up as follows:

Let three "mass points", A, B and C be projected from the same point in space and left to themselves (i.e. be not subjected to any forces). Then we can find a co-ordinate system S relative to which these three points describe three different straight lines. Such a co-ordinate system is defined by Lange to be an "inertial system"; relative to it any fourth mass point, left to itself, would move in a straight line. In effect, this last statement may be held to define an inertial system; that is, an inertial system is a co-ordinate system relative to which a particle "left to itself", i.e. not acted on by any forces, would continue in its state of

/rest.....

rest or uniform motion in a straight line. To state this definition in somewhat shorter form: an inertial system" or "inertial frame of reference" is a co-ordinate system relative to which Newton's Law of Inertia (i.e. his First Law of Motion) holds. This is, in fact, the commonly used definition of an inertial frame today; it was adopted by Einstein for his exposition of the special theory of relativity (cf. Einstein 1920. 1922) and is to be found in textbooks on that subject (see e.g. Katz 1964 or Resnick 1968) or on introductory or intermediate (university) physics (see e.g. P.S.S.C. 1960, 1965, Kittel et al 1965, or Sears and Zemansky 1964). (Dingle (1962) alleges that "widespread confusion" exists in regard to the meaning of the term "inertial system" and quotes fifteen examples of definitions of inertial systems in an endeavour to prove his point, but these examples are only apparently contradictory in content.)

At last we seem to have a prescription for choosing "convenient" frames of reference. "Inertial frames" as defined above are, in fact, frames of reference in which the laws of Newtonian mechanics will be found to hold and in respect of which descriptions of motion will be simpler than in non-inertial frames. To identify such a frame, all we need to do is to select a co-ordinate system relative to which a particle acted on by no forces will move with uniform velocity. But immediately we encounter a difficulty: how do we know whether or not any forces act on a particle? Suppose we adopt Newton's definition of a force (Principia, Definition IV): "An impressed force is an action exerted upon a body, in order to change its state, either of rest, or of moving uniformly forward in a right line". Then our test for deciding whether or not an "impressed" force acts on a particle is to see whether or not the particle is being accelerated.

/accelerated.....

Accelerated relative to what? Presumably relative to an inertial frame, since inertial frames are the "special" frames we have in mind when we enunciate the laws of mechanics. What is an inertial frame? But this is exactly the question we set out to answer when embarking on this argument. The customary way out of this impasse is to employ an additional criterion for the presence or absence of forces on a body, viz. that a force on a body must be associated with the presence of an object or objects in its environment (cf. Eisenbud 1958 or Halliday and Resnick 1966 for an explicit statement of this postulate). Using this idea we are able to argue that if we could find an object in the universe that was very greatly distant from its nearest neighbours, we could to a good approximation identify an inertial frame with a coordinate system moving with uniform velocity relative to this object. (Cf. Weinstock's (1961) definition of an inertial frame: "Whenever the motion of any particle is influenced by no external agency - strictly, if the particle is infinitely distant from all other bodies - that particle moves with constant (vector) velocity with respect to an inertial co-ordinate system ".) Objects that appear to meet this condition are the so-called fixed stars. Thus our journey in search of inertial frames has now led us to the fixed stars, a destination at which Mach arrived (conceptually, at any rate!) some time ago, though by a somewhat different route. Commenting on Lange's definition of an inertial frame, he remarks that "we only <u>apparently</u> evade by such expressions references to the fixed stars" - a conclusion with which we might now sympathize. The fixed stars are a rather remote standard to appeal to for everyday use; again, the postulate associating forces with objects in the environment comes to our aid. Consider a body falling freely in vacuo towards the earth; it is by no means moving with constant velocity relative to

/the.....

the earth, yet to a very good approximation we can account for its deviation from motion with constant velocity by saying that it is acted upon by a force arising from the presence of a very large object in its environment; in other words, we say it is being accelerated by the force of gravity which is found to be proportional to the mass of the particle on which it acts. Provided that this gravitational force is 'real', we would feel it reasonable to extrapolate from the observed motion of particles subject only to the gravitational force to the assertion that if this force were removed, a particle "left to itself" would move with nearly constant velocity relative to the earth. We could then conclude that a "frame of reference" fixed in the earth would constitute a good approximation to an inertial frame. We would be strengthened in this view by the knowledge of certain relevant facts concerning the earth, the sun and the stars: the acceleration of a particle on the equator towards the centre of the earth is about 3.4×10^{-2} m/s^2, the acceleration of the earth in its orbit around the sun is about 6×10^{-3} m/s^2, and the acceleration of the sun relative to other stars is likely to be several orders of magnitude smaller than this (Kittel et al 1965). For certain purposes, allowance would have to be made for effects arising from the rotation of the earth on its axis, but for conventional laboratory experiments these effects are negligible. Similar reasoning can be used to establish that a frame of reference "fixed in" the sun would for many purposes serve adequately as an inertial frame; for example, it may be used in this way in treating the motion of the planets around it.

Our quest for inertial frames has now brought us back to the earth and the sun (at any rate as far as our most pressing needs are concerned), and again a quotation from Mach seems

/relevant: .....

relevant: "I have remained to the present day [i.e. the end of the nineteenth century] the only one who insists upon referring the law of inertia to the earth, and in the case of motions of great spatial and temporal extent, to the fixed stars." Having reached this satisfying conclusion, we are now ready to face the fact that we cheated a little in accepting without challenge the view that there really are gravitational forces. What is our justification for asserting that other bodies exert gravitational forces on a given body? Have we a means of checking this that is independent of one's choice of frame of reference? To test for the presence or absence of an electric field, for example, one can carry out observations on the motion of positively- and negatively-charged particles; if these experience the same acceleration when placed in turn in the same small region of space, no electric field is present (cf. Bergmann 1968). With our supposed gravitational forces, the situation is different. The only test available seems to be one based on Newton's Definition IV - and this leads us to inquire once again what an inertial frame is. We are back where we started.

Or are we? Surely we have already solved this problem, by showing that the acceleration of a particle on the surface of the earth is small relative to the fixed stars and implying thereby that a frame fixed in the earth must be a good approximation to an inertial frame? But this "solution" depends on the assumption that space-time is so constituted that inertial frames can have indefinite extent. If we cannot make this assumption, then we cannot assume that an inertial frame validly defined in one region of space has no acceleration relative to an inertial frame validly defined in another region of space; and if we cannot make the latter assumption, how can we know that the forces which we call "gravitational"

/on.....

on earth are not just fictitious, "inertial" forces arising from our choice of an incorrect frame of reference from which to make our observations? It is an essential feature of "inertial forces" (e.g. the "centrifugal force") that they are proportional to the masses of the particles on which they are postulated to act; gravitational forces are known to exhibit this feature, as has been experimentally established to a very high degree of accuracy.

A means of ultimately escaping from these difficulties was shown by Einstein (1920, 1922), who said, in effect: Abandon your prejudiced view that there are gravitational forces in the universe. Stick to your definition of an inertial frame as one in which a free particle will move with uniform velocity, but do not assume in advance that a frame of reference so defined can have indefinite spatial extent.

According to Einstein's prescription, a particle in free fall should not be thought of as being acted upon by a gravitational force. By definition, the fact of its being in free fall implies that no non-gravitational forces act on it; if gravitational forces do not exist, then it is truly a free particle and, in terms of the above definition, a frame of reference fixed in it must be an inertial frame. Space vehicles in free fall are physical realizations of such inertial frames; it is a matter of familiar experience to space travellers that particles "left to themselves" in freely-falling space vehicles will travel with uniform velocity. Imagine that at some future time someone is born and raised in an orbiting space station. Such a person would not ordinarily have any reason to suspect the existence of gravitational forces; to put it another way, no gravitational forces can be felt in the space station. This gives us an alternative way of stating the definition of an inertial frame: such a

/frame.....

frame is one in which no gravitational forces will be felt (cf. Taylor and Wheeler 1966). If our hypothetical space ship is large enough, our hypothetical inhabitant of it might make an interesting discovery if he made careful measurements on the relative motion of a pair of widely-separated "free" particles in the ship. He would find that the particles experienced a definite relative acceleration. From the "prejudiced" point of view of an observer in a terrestrially-based frame of reference, one would say that this relative acceleration arose from the different gravitational accelerations experienced by the particles in a non-uniform gravitational field. From the point of view of the observer in the space ship, who has never heard of gravitational forces, the particles are evidently subject to some kind of "tidal force". If this observer were sufficiently well educated, he would know that a theory existed in which such tidal forces are explained in terms of the curvature of space-time in the vicinity of the space ship. This theory is Einstein's General Theory of Relativity.

If a relative acceleration between two free particles were observable in the space ship, this would show that the space ship <u>as a whole</u> could not strictly be regarded as an inertial frame. To the extent that this acceleration could be neglected in comparison with other accelerations of interest that might be measured in the space ship (e.g. those resulting from the application of "real" forces), one could treat the latter as a satisfactory approximation to an inertial frame for one's purposes. If one used only a limited region of the space-ship for one's experiments, the tidal accelerations one could observe in this region would be smaller than those that could be observed in the space-ship as a whole, and one would thus be working in an even better approximation to an inertial frame. Thus "inertial frames"

/in.....

in the sense with which we now use the term are strictly local in character, and it is now clear why a warning was given above that inertial frames should not be assumed to have indefinite extent.

The above discussion has shown that if the search for inertial frames is pursued far enough, it takes one inevitably to the doorstep of the General Theory of Relativity. Presumably this is not where teachers would want to lead any but the brightest of their pupils. There are signs that the use of Einsteinian inertial frames (or what Bergmann (1968) calls "free-falling frames of reference") is being considered worthy of trial at an elementary level. The introductory text on relativity by Taylor and Wheeler (1966) introduces the reader to Einsteinian inertial frames on page 8, and Kaempffer (1967), in his introductory text on physics, discusses Einsteinian inertial frames before he mentions "Newtonian" ones (i.e. frames unaccelerated with respect to the fixed stars). Also, one of the approaches to mechanics recommended by the Association for Science Education (1963) for their "advanced phase" syllabus does <u>introduce</u> the pupil to the Einsteinian approach to gravitation; but it is noteworthy that the authors of the syllabus remark that this approach should be confined to "the brightest of top-set sixth-formers". As they say, "Pupils below this level have sufficient difficulty in mastering the more elementary principles". In the view of the present writer, neither pupils nor teachers in South African schools would be ready for such an approach to mechanics for some time to come. Nor would the average South African senior school pupil benefit from any explicit discussion of the concept of an <u>inertial</u> frame of reference; the mental effort required to absorb a further new concept and to understand the manner of its definition are more likely to lead to bewilderment and confusion than

/enlightenment.....

enlightenment at this stage. (Of course, if special relativity were ever to be included in the senior school syllabus, explicit discussion of the concept of an inertial frame could not be avoided.) Nevertheless, some attempt must be made to come to terms with the problem of frames of reference in mechanics teaching; as will have been apparent from the foregoing discussion, the matter is too important to be shrugged off. We wish to propose the following:-

(i) Introduction of the term "frame of reference"

We suggest that the term "frame of reference" should be introduced, so that it can become part of the pupil's vocabulary and the associated idea can become part of his thinking about mechanical problems. The term need not be formally defined and there is no need as yet to speak of it as a co-ordinate system. Instead a frame of reference can for practical purposes be identified with a material object or system, though the pupil should also become accustomed to thinking of imaginary objects as reference frames (e.g. "What would be the speed of that [real] aeroplane relative to a car [imaginary] below it travelling in the same direction with a speed of 60 m.p.h., if the speed of the aeroplane is 150 m.p.h.?... In the imaginary frame of reference of that car, the speed of the aeroplane is 90 m.p.h.") By way of elaboration of the concept, it would be worth discussing with pupils one or two simple examples of the kinematical description of the motion of a particle from the points of view of two different frames of reference which have uniform relative velocity; the motion of a vertically-falling raindrop as viewed from the ground and from a moving car, or the motion of a bomb dropped (in vacuo, preferably!) from an aeroplane as viewed from the aeroplane and the ground would be suitable examples. Perhaps, it should be made clear that both frames

/of.....

of reference yield "correct" descriptions of the motion, although in practice we might find it more natural or convenient to use one or the other.

(ii) Inertial frames in disguise It is our singular good fortune as earthly inhabitants to have beneath us an object which for many purposes can serve as an "anchor" for a satisfactory inertial frame (in the pre-Einsteinian sense). For this reason, it has been possible for elementary mechanics to be taught quite successfully without explicit reference to inertial frames; usually in such teaching the frame of reference has been, tacitly or avowedly, "the earth" or "the laboratory" which, as we have seen, must inevitably lead to the emergence of the laws of Newtonian mechanics, other things being equal. It is not our intention to complain of this; indeed, in this context we would take our stand with Mach in insisting in most cases on "referring the law of inertia to the earth" or to frames of reference moving with uniform velocity relative to it or portions of its surface. What we would recommend is that where the earth or any other object is used as the frame of reference in the treatment of dynamical problems, this fact should on occasion be explicitly stated. The suitability of a reference frame "fixed in" the earth can be taken for granted, and the logical difficulties associated with Newton's First Law can be avoided for the moment by the use of argument given earlier, viz. that a "free" particle near the earth does not move with uniform velocity because it has a large body in its environment which exerts a gravitational force on it. There seems to be no escape from the view that if Newtonian mechanics is to be successfully introduced to senior school pupils, it must be done on the basis that gravitational forces are as "real" as other, more tangible "pushes" and "pulls". Although, as we have seen, the logic of such a presentation is open to certain

/doubts.....

doubts, it does have the merit of leading to a description of the world around us in terms that can be understood by a beginner.

Should "non-inertial frames" (in the Pre-Einsteinian sense) be discussed with pupils? Again we would suggest that the term should not be mentioned. As for the idea, the extent to which this might be discussed with pupils must depend on the teacher's assessment of how much his pupils can assimilate. If he feels that they might find the idea confusing, the course for him to follow is to stick rigidly (if not literally!) to the earth or the laboratory as the frame of reference to be used in virtually all problems of interest. For example, in discussing the motion of an artificial satellite round the earth, he would do so entirely from an earthly frame of reference and not raise the question of how things might look, dynamically speaking, to the occupants of the satellite. He would steer entirely clear of such questions as what would have to be done to make the laws of physics hold inside an accelerating railway carriage or a rotating merry-go-round. With luck, his pupils might not raise these issues themselves. With a brighter class, these issues could be raised and the valuable point made that in non-inertial frames, "fictitious" forces (i.e. forces not associated with any objects in the environment) must be postulated in order to make the laws of mechanics hold. The above-mentioned film "Frames of Reference" could be of great assistance here. If "centrifugal force" has to be mentioned at all (see below, Section 4. 9) this would be the place to introduce it.
In this connection, the authors of the "advanced phase" Physics syllabus mentioned above make an interesting observation: "...we have great difficulty in persuading pupils to answer A-level questions making the accepted distinction [between "fictitious" forces and "real" gravitational forces]." Does this very difficulty imply that an intuitive "feel" for

/relativistic.....

relativistic dynamics comes more naturally than is generally supposed? Perhaps it does; the point is well worth detailed investigation. Perhaps the time may come when sixth-form teachers in Great Britain and university teachers in other countries such as South Africa will feel free to teach mechanics from the Einsteinian viewpoint; but if they do, they will surely have to build on the foundation laid by a sound training in the fundamentals of Newtonian mechanics.

We asserted at the outset of this discussion that the teacher ought to know at least as much as the above about frames of reference if he is to teach mechanics adequately and to be able to stimulate his best pupils. It may seem to some that this is going too far - that the difficulties attached to the concept of an inertial frame need be of no concern to the teacher. But the teacher who chooses to ignore these matters surely does so at his peril; if he has not thought his way through some of the above reasoning concerning inertial frames and their definition, he may be thrown into utter confusion by an untimely question from a pupil who has pondered the problem and been troubled by it. And as we have seen, this kind of discussion leads to general relativity; the teacher who knows and appreciates this can give his best pupils excellent food for thought.

4.7 Newton's laws of motion

4.7.1 Newtonian mechanics and the teacher

Here we reach the hard core of classical mechanics, which is at once one of the most important sections of the syllabus and one of the most difficult to grasp. Consequently, more than ordinary care and thought must be devoted to preparation for its successful presentation. The situation in South African science teaching is such that teachers in this country need a great deal of guidance in this

/preparation.....

preparation (cf. Helm 1967). Teachers in government schools have been handicapped by syllabuses which have carefully excluded even a cursory treatment of the fundamentals of mechanics. With the introduction of the new common basic syllabuses this state of affairs will soon end, but many teachers will then have to try to teach mechanics properly without any previous experience. No-one with such experience would suggest that this task is going to be easy; what follows below may be taken as an attempt once again to outline what is regarded as essential background knowledge for the teacher. This will in turn be followed by a discussion in which we shall not try to lay down One True Method of presentation of the subject, but instead suggest basic principles underlying what seems to us to be a satisfactory treatment of the subject at this level. In addition, we shall try to indicate as clearly as possible the various pitfalls that will have to be avoided if pupils are not to become seriously confused about the fundamental issues.

Classical mechanics as it is customarily taught in schools and in introductory university courses originated with Newton's "Principia", the tercentenary of the publication of which will be celebrated in less than twenty years' time. During the time that Newton's Laws have been the property of mankind, controversies have raged about the precise interpretation of the content of these Laws and the manner in which they should be taught. On the latter subject, one notes with a certain amount of cynical amusement the way in which certain issues are raised, hotly debated for a time and then laid to rest, only to be revived some years later in precisely the same terms as before. Confronted with this state of affairs, the would-be teacher of mechanics might well take flight. Yet his panic must be countered, for Newtonian mechanics <u>must</u> be taught; and panic is in any case unnecessary, because out of the vast amount that has been said

/and.....

and written on the teaching of mechanics it is possible to extract some sound, worthwhile advice. Our hypothetical would-be teacher might take comfort from the evidence of textbooks published over the last few decades which reveals that while there was once a good deal of confusion on the fundamentals of mechanics among textbook authors (and teachers, one presumes) there is much less now; in fact, among the more enlightened members of the world-wide teaching community it seems that something of a consensus is being attained on many of the important issues in mechanics teaching.

4.7.2 Newton's formulation of mechanics

According to Thomson and Tait (1886), "We cannot do better, in all events at commencing, than follow Newton somewhat closely. The Introduction to the 'Principia' contains in a most lucid form, the general foundation of dynamics." There is no doubt that the idea that "the Introduction to the Principia contains...the general foundation of dynamics" is still widely believed. Truesdell (1968) has trenchantly argued that this view claims too much for Newton (whose achievement he in no way wishes to belittle) and gives too little credit to the work of his successors, notably Euler. Himself a "specialist in mechanics", Truesdell asserts that "no modern specialist in mechanics shares the opinion that the three laws at the head of [the Principia] really suffice as a basis for mechanics as a whole." He points out that the attempts Newton made in Book II of the Principia to come to grips with problems in fluid mechanics were largely unsatisfactory, because Newton's method of approaching these problems was to employ "special reasoning, special hypotheses and special guesses, related vaguely to mechanical ideas and principles, but not in the least derived from a system of mechanics", instead of "the two essentials: the concept of

/hydrostatic.....

hydrostatic pressure and the general principle of linear momentum." Evidently neither Newton nor his contemporaries "had a grasp of the principle of momentum sufficient to solve a single problem concerning the motion of a deforming body." Neither the Law of Moment of Momentum (Truesdell's name for the "rotational analogue" of Newton's Second Law) nor the Law of the Lever, without which problems in the mechanics of rigid or deformable bodies cannot be solved, are to be found anywhere in the Principia (see Section 4.13 for a discussion of the status of these laws). Furthermore, "... except for certain simple if important problems, Newton gives no evidence of being able to set up differential equations of motion for mechanical systems...the cold fact is, the equations are not in Newton's book." Evidently the first person to propose the equations $F_x = ma_x$, $F_y = ma_y$, $F_z = ma_z$ "as general, explicit equations for mechanical problems of all kinds", applicable to every part of every system, was Euler, who set forth these equations for this purpose in a paper published in 1752. At that time, Euler believed that these equations "[included] all the laws of mechanics"; later he was to conclude that the Law of Moment of Momentum had to be added to them as a separate principle.

Classical mechanics in the South African secondary school syllabus is concerned essentially with particle mechanics, the basis of which, however imperfectly formulated, is to be found in the "Principia". Hence, although Newton may not have provided a completely general formulation of dynamics, what he accomplished is still relevant to the needs of the present-day South African pupil. Truesdell makes clear that the present-day practice of using "F = ma" to solve problems in particle mechanics derives more directly from Euler than from Newton. However, the pattern of presentation of the subject to be

/found.....

found in elementary courses is still strongly influenced by that established by Newton in the "Principia"; indeed, many textbook authors and teachers consider it obligatory to declare their allegiance to Newton, even to the extent of crediting him with the performance of certain experiments which he no doubt ought to have carried out but unfortunately never did. (Cf. Shortley and Williams (1965), who, in order to adapt Newton's treatment of mechanics to their own quasi-Machian presentation of the subject, refer to experiments in which Newton is alleged to have measured the accelerations of a pair of interacting bodies.) We therefore think it appropriate here to attempt a brief sketch of Newton's formulation of the fundamental principles of mechanics, from which it might be possible to assess the extent to which his writings on the subject can be of use to the present-day teacher or pupil. In Book I of the "Principia", Newton sets forth eight Definitions, three "Axioms, or Laws of Motion", and eight Corollaries. The definitions are concerned with the concepts of "quantity of matter" (mass), "motion" (momentum) and force, all of them crucial to Newton's exposition of mechanics and none of them hitherto defined with any clarity. Definition I defines "quantity of matter" as the "measure of the same, arising from its density and bulk conjointly"; we have commented earlier (Section 3.7.2) on the difficulties associated with this definition. Although Burniston Brown (1965) claims that it is a "perfectly satisfactory physical definition of a measure of an amount of matter", the uncertainty about the meaning of the term "density" as used in the "Principia" (cf. the discussion of this point in Jammer (1960)) and in particular as to how it is to be <u>measured</u> leads us to conclude that the meaning of Definition I is not entirely clear. However, in his commentary on this definition, Newton makes a most significant remark: he says that

/"the.....

"the same [mass] is known by the weight of each body, for it is proportional to the weights, as I have found by experiments on pendulums, very accurately made, which shall be shown hereafter." This shows that as Newton conceived the term, mass was not defined in terms of the gravitational properties of matter, for if it were, his "experiments on pendulums" would have been "Zanzibar experiments". In other words, Newton did not regard mass as a priori measurable with an equal-arm balance. In Definition III he introduces another concept, "vis insita" or "innate force of matter" which is essentially identical with what would now be called "inertial mass". He says: "The vis insita, or innate force of matter, is a power of resisting, by which every body, as much as in it lies, continues in its present state, whether it be of rest, or of moving uniformly forwards in a right line." Once again, in commenting on this definition he makes a most significant remark: "This force is ever proportional to the body whose force it is; and differs nothing from the inactivity of the mass, but in our manner of conceiving it." In other words, if by some means we are able to measure the "innate forces" of various samples of matter, we shall be able thereby to compare their "quantities of matter", or masses. If one takes all Newton's various statements concerning "vis insita" into account, it seems fair to say that operationally, though not conceptually, he identified "vis insita" with "quantity of matter". Although he speaks of this concept as a "force" (an "unfortunate" and "unnecessary" usage, as Feather (1965) puts it; Jammer (1957) regards it as a concession to pre-Galilean mechanics), his use of it is essentially consistent with the interpretation that it is to be identified with the modern concept of inertial mass. Newton himself suggests "vis inertiae" ("force of inactivity") as an alternative name for this concept.

/Definition.....

Definition II defines what we would now call momentum: "The quantity of motion is the measure of the same, arising from the velocity and quantity of matter conjointly." This can be unambiguously rendered in modern terms as "momentum is the product of velocity and mass". The remaining five definitions are all concerned with different kinds of "forces." We have already referred to Definition IV in connection with inertial frames; in full, it reads: "An impressed force is an action exerted upon a body, in order to change its state, either of rest, or of uniform motion in a right line." As countless commentators have pointed out and is in any case obvious enough, the content of this definition seems to be identical with that of Law I: "Every body continues in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by a force impressed upon it." Like certain other features of the "Principia", this confronts one with a puzzle: If Definition IV and Law I really had identical content as far as Newton was concerned, why did he choose to say the same thing twice over, once as a "Definition" and once as a "Law"? And if he did not mean them to be interpreted in this way, how exactly are they to be interpreted? Many commentators have decided that Law I and Definition IV do say the same thing, and have preferred to regard the problem of why they were separately enunciated as insoluble. Feather (1959, 1965) takes a slightly broader view; for him Law I states nothing which is not stated in one or other of Definitions III or IV. Jammer (1957), who thinks that for Newton, force was "a concept given a priori, intuitively, and ultimately in analogy to human muscular force", appears to regard Law I as "an empirical statement describing the motion of free bodies"; and he says that "Definition IV... is not to be interpreted as a nominal definition, but as

/summarizing.....

summarizing the characteristic property of forces to determine accelerations."

Definitions V-VIII all deal with "centripetal" forces. Definition V states that "A centripetal force is that by which bodies are drawn or impelled, or any way tend, towards a point as to a centre." Definitions VI and VII, concerned respectively with the "absolute" and "accelerative" quantities of centripetal forces, have no direct relation to our main theme and will not be discussed here. Definition VIII states that "The motive quantity of a centripetal force is the measure of the same, proportional to the motion it generates in a given time." Here again is a definition with a link with one of the Laws of Motion (Law II in this case). Newton seems to have singled out centripetal forces for special attention in this way in view of their central importance for his main purpose, viz. the exposition of the "system of the world". His remark in connection with Definition IV that "Impressed forces are of different origins; as from percussion, from pressure, from centripetal force" makes it plain that he regarded centripetal force as a highly important special case of an impressed force.

Newton follows his set of definitions with a "Scholium" in which he gives his views on the subjects of absolute time, space, place and motion. We have given some attention to these ideas in discussing "time" and "frames of reference" above; there is no need to add anything here. The "Scholium" is followed by the Laws of Motion. Law I has been quoted and briefly commented upon above. Law II runs thus: "The change of motion is proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed." It is instructive to compare this with the usual modern rendering of the Law: "The rate of change of momentum of a body is proportional to the impressed force and

/takes.....

takes place in the direction of the force." In all respects but one, the modern version may be taken as an accurate rendering of the content of the original. The significant difference is that the modern statement substitutes "rate of change of momentum" for Newton's "change of motion". Neither in his statement of Law II nor in his commentary on it does Newton make any reference to "time" or "rate of change". Does this omission mean that Newton's idea of force as a quantitative concept was different from that now accepted? Maxwell (1876) thought that it was: he writes: "...by impressed force [Newton] means what is now called Impulse, in which the time during which the force acts is taken into account as well as the intensity of the force." Several modern scholars have followed Maxwell in this interpretation (cf. Dijksterhuis (1961), Arons and Bork (1964), Bork (1964) and Arons (1965)). Besides pointing to the absence of any mention of time in the statement of Law II, these scholars draw attention to such remarks of Newton's as "When a body is falling, the uniform force of its gravity acting equally impresses, in equal intervals of time, equal forces upon that body, and therefore generates a whole velocity proportional to the time." Such a statement could certainly be taken to mean that a "force" in the Newtonian sense was something that is generated over an interval of time. On the other hand, Seeger (1962) cites a statement from another part of the "Principia" (Proposition XXIV, Section VI, Book II), which can be held to support the conventional reading of the Second Law: "For the velocity which a given force can generate in a given matter and a given time is directly as the force and the time, and inversely as the matter. The greater the force or the time is, or the less the matter, the greater the velocity generated. This follows from the Second Law of Motion". Here Newton certainly seems to be saying what in mathematical

/form.....

form would be rendered as "$v = \frac{Ft}{m}$", where F would correspond to the modern concept of force. Also, as Feather (1965) points out, Definition VIII does make explicit reference to "the motion which it generates in a given time" in defining the "motive quantity" of a centripetal force, and he feels justified in "[revising Law II] without incurring the charge that we are in any way changing the content which Newton intended it to have". (His "revision" is the usual modern version). It is noteworthy that two scholars who have obviously studied Newton's writings at least as carefully as anyone else, have both in recent works accepted the "usual modern version" without question (cf. Jammer 1957, Hesse 1961).

Yet even if we accept that Law II really does mean that force is proportional to rate of change of momentum, our difficulties with the interpretation of this law are by no means at an end. The position is thus summed up by Hesse (1961): "This...may be interpreted empirically or logically: either as the natural generalization of the fact that when what appear to be two equal forces act upon a body, they produce twice the acceleration produced by one force acting alone; or, if we assume with Newton (Definition I) that a statical definition of mass is already given, the law may be taken as the definition of the measure and direction of the force acting, the constant of proportionality between force and acceleration being the inertial mass of the body." It is impossible now to be certain of which (if either) of these views of Law II was held by Newton. Some commentators (e.g. Jammer 1957) have argued in favour of the first of these interpretations, on the grounds that Newton does, after all, distinguish clearly between definitions and laws; as we have noted, Jammer is of the opinion that force, for Newton, was a concept given a priori. Others (e.g. Feather 1959, 1965 and - rather more tentatively - Hesse 1961) have taken the contrary

/view;.....

view; for them, Law II is indeed a quantitative definition of force in terms of mass and acceleration, so that, according to them, Newton did not intend force to be an independent concept.

In full, Law III runs as follows: "To every action there is an equal and opposite reaction; or, the mutual actions of two bodies upon each other are always equal, and directed to contrary parts." It is to be regretted that in modern renderings of this Law, the second part of Newton's statement, which amplifies the meaning of the first, should so often be omitted. If the full statement were given (with the term "force" used in place of "action"), then at least it would be clear from the first what the Law actually says, even though its full implications might not initially be realized. There is much less room for doubt about the status of Law III (which Mach called "the most important achievement of Newton with respect to the principles" of mechanics) than there was about Laws I and II. Here seems to be an unequivocal statement about the way Nature behaves, with no possibility of its being confused with a mere definition. The only point of doubt is whether Law III is to be regarded as <u>wholly</u> empirical or whether, in the case of forces acting between different parts of the <u>same</u> body, it can be regarded as a deduction from Law.I. As Hesse points out, if Law I is a mere definition, Law III must be treated as wholly empirical; but if a force can be recognized apart from the acceleration it produces so that Law I is truly a law, then Law III is a deduction from it as far as forces acting within a single body are concerned. Newton does appear to be using Law I in this way in his demonstrations in the Scholium following the Corollaries that Law III must hold for a pair of mutually attracting bodies and for mutually attracting parts of the earth. He was of course

/particularly.....

particularly interested in the application of Law III to mutually attracting bodies, as its validity in such situations was essential to his theory of gravitation. He supposes an obstacle to be placed between two such bodies, A and B, to prevent their coming together. Then, he says, if B attracts A more than A attracts B, the obstacle will be pushed harder by A than by B, and in consequence the whole system of three bodies will start to accelerate through space in the direction from A to B "in infinitum". (We might note that if the system is not in equilibrium, the force exerted by A on the obstacle is not necessarily equal to the force exerted by B on A, a point which Newton does not mention in his brief discussion.) But since no external forces are acting on the system, by Law I it cannot proceed to accelerate away in this manner: "and therefore the bodies must equally press the obstacle, and be equally attracted one by the other." Of course this proof does not establish the equality of action and reaction when the attracting bodies do not have an obstacle between them; in the latter case the validity of Law III must be regarded as proved by the correctness of the predictions based upon it. As far as its application to collisions is concerned, Newton believed that he had established this by experiments he describes in the above-mentioned "Scholium": "And thus the third Law, so far as it regards percussions and reflexions, is proved by a theory exactly agreeing with experience." (Newton does not describe how he measured the masses of the colliding bodies used in his experiments, so that there is a possibility that the correctness of Law III was unwittingly assumed in the determination of these masses. See below (Section 4.12) for a discussion of whether or not it is possible to prove the Law of Conservation of Momentum experimentally.)

/Finally.....

Finally in this brief survey of Newton's statement of the fundamental principles of mechanics, we deal with the content of the six Corollaries that follow the three Laws. Corollary I is Newton's proof of the theorem of the parallelogram of forces: "A body by two forces conjoined will describe the diagonal of a parallelogram, in the same time that it would describe the sides, by those forces apart." It is a fundamental principle of mechanics that forces add vectorially, as we would now say. Newton's proof which, interestingly enough, treats of forces in given directions producing uniform motions in those directions (are the forces impulsive in this case?), depends on the assumption that the effect of one force applied to a body is independent of that resulting from the application of any other force. He does not state this assumption, which is, however, validated by the correctness of the results predicted from it. Corollary II points out that the preceding Corollary is the basis for the law of composition of one force from two forces "and, on the contrary, the resolution of any one direct force...into two oblique forces... which composition and resolution are abundantly confirmed from mechanics". Corollary III states, in Newtonian terminology, the Principle of Conservation of Momentum: "The quantity of motion, which is collected by taking the sum of the motions directed towards the same parts, and the difference of those that are directed to contrary parts, suffers no change from the action of bodies among themselves". Newton proves this with the aid of Laws II and III in a manner which is essentially the same as that of any modern textbook deduction of the Principle from Law III. Corollary IV states a simple consequence of the preceding one: that the common centre of gravity of an (isolated) system of interacting bodies remains at rest or moves uniformly in a straight line. Corollary V is a statement of what has been called

/the.....

the "Newtonian Principle of Relativity": "The motions of bodies included in a given space are the same among themselves, whether that space is at rest, or moves uniformly forwards in a right line without circular motion". Corollary VI is also concerned with relative motion: "If bodies, any how moved among themselves, are urged in the direction of parallel lines by equal accelerative forces, they will all continue to move among themselves, after the same manner as if they had been urged by no such forces." By "equal accelerative forces" we may understand "forces that would produce the same acceleration in each body." Therefore this Corollary tells us that if a uniform gravitational field were suddenly to be imposed upon a system of bodies, their relative motions would be the same as in the absence of the field.

This survey of Newton's exposition of the fundamental principles of his system, though brief and superficial, has at least the following advantages:-

(i) it does make known what Newton actually said;

(ii) it makes it clear that it is impossible to base an introductory treatment of Newtonian mechanics on an uncritical use of Newton's own statements as they stand; any attempt in this direction will have to depend on a number of decisions on the interpretation of Newton's Laws and Definitions. This is no criticism of Newton, who was not, after all, trying to write an introductory textbook for school or university use; his urgent concern was with clearing the ground for his account of the "System of the World". If in the course of this he made statements which are not wholly clear to succeeding generations, or the logic of which is sometimes open to doubt, this cannot be held to detract from the greatness of his achievement. Mach (1902) has expressed the matter well: "Like the commander of an army, a great discoverer cannot stop to institute petty inquiries regarding

/the.....

the right by which he holds each post of vantage he has won. The magnitude of the problem to be solved leaves no time for this."

4.7.3 Mach's formulation of mechanics

Nowadays, then, we feel unable to accept the view of Thomson and Tait (1886) quoted at the beginning of the previous section. The difficulties associated with Newton's formulation of the fundamentals of dynamics have led to a number of attempts to reformulate them, of which the most famous and influential has probably been that of Mach (1883, trans. 1902). In the present century, various authorities (e.g. Gardner 1921, Roller 1936, O'Leary 1947b,c) have urged the adoption of the Machian approach in the introductory teaching of mechanics, and a number of introductory textbooks in mechanics or general physics now either adopt this approach more or less explicitly or follow a treatment which is in accord with that of Mach in most respects. (Examples are Lindsay 1933, first edition(also subsequent editions); Ference et al 1956, Feather 1959, Shortley and Williams 1965, Orear 1961, 1967 and Alonso and Finn 1967, to name only some among many. Also, in the treatment of mechanics recommended for "outstanding pupils" taking the "advance phase" syllabus of the Association for Science Education (1963), the "experimental propositions" set out as the basis of Newtonian dynamics are substantially identical with Mach's own "experimental propositions.") Thus it behoves us to examine Mach's ideas in some detail, and also to consider some of the criticism levelled at them in recent years. This should help us to achieve a clearer focus both on the fundamental principles of dynamics and on what is needed in their introductory treatment.

Mach's attitude to physics in general and Newtonian mechanics in particular in many respects resembles that to

/be.....

be found in the writings of Bishop Berkeley (cf. Popper 1953, rep. 1963), who took the view that "Force, gravity, attraction and similar terms are convenient for purposes of reasoning and for computations of motion and of moving bodies, but not for the understanding of motion itself." (Berkeley, "De motu", quoted in Jammer 1957.) If Berkeley is correct, then a proper formulation of mechanics would not contain "force" as an independent concept, but seek to define it in terms of already-existing concepts such as space, time and mass. To this task Mach addressed himself, but carried it a stage further in that he also sought a definition of mass that would be dependent only on concepts of space and time and would thus on the one hand be independent of force and on the other be free of the difficulties attached to the definition of mass as "quantity of matter". For Newton's sets of Definitions, Laws and Corollaries Mach (op. cit.) proposed the substitution of three "Experimental Propositions" and two definitions, which we quote below:-

"a. Experimental Proposition. Bodies set opposite each other induce in each other, under certain circumstances to be specified by experimental physics, contrary accelerations in the direction of their line of junction.

b. Definition. The mass-ratio of any two bodies is the negative inverse ratio of the mutually induced accelerations of those bodies.

c. Experimental Proposition. The mass-ratios of bodies are independent of the character of the physical states (of the bodies) that condition the mutual accelerations produced, be those states electrical, magnetic, or what not; and they remain, moreover, the same, whether they are mediately or immediately arrived at.

d. Experimental Proposition. The accelerations which any number of bodies, A, B and C ... induce in a body K, are

/independent.....

independent of each other.

e. Definition. Moving force is the product of the mass-value of a body into the acceleration induced in that body."

As Mach set out these ideas less than a century ago, we are not faced with any difficulties of interpretation; he uses modern terms in a modern way and without any room for misunderstanding. His definition of mass-ratio amounts to the following equation:

$$\frac{M_A}{M_B} = \frac{|a_B|}{|a_A|},$$

where M_A and M_B are respectively the "mass-values" (masses) of the bodies A and B, and $|a_A|$ and $|a_B|$ are the magnitudes of their accelerations. His second experimental proposition tells us that this ratio is the same for a given pair of bodies whatever the means whereby they induce accelerations in each other. It also makes the following point:-

Let there be three bodies, A, B and C which are allowed to interact with one another in pairs. Then if $\frac{M_A}{M_B}$, $\frac{M_B}{M_C}$ and $\frac{M_C}{M_A}$ are the mass-ratios of these particles that are found by measurements on the mutual accelerations of A and B, B and C and C and A respectively, then it is an experimental fact that $\frac{M_A}{M_B} = \left(\frac{M_B}{M_C} \times \frac{M_C}{M_A}\right)^{-1}$.

In other words, the mass-ratio of A and B may be found "mediately" by measurements on the interactions of B and C and A and C, as well as "immediately" by measurements on the direct interaction between A and B. The remaining experimental proposition is the assumption tacitly made by Newton in proving the theorem of the parallelogram of forces. Mach's other definition is, of course, Newton's Second Law in its conventionally-stated form, but here explicitly stated to be a definition and thus not an experimental law.. One notes

/with.....

with interest that in the first experimental proposition, it is stated in effect that a force on a body arises from interaction with another body (cf. our discussion of this point in connection with inertial frames).

At first sight, Mach's reformulation of the principles of Newtonian mechanics has a number of attractive features. The experimental propositions and definitions are free from ambiguity; the attempt to define force and mass independently of each other and in terms of kinematical concepts only, seems to have been successful; and the overall impression one receives is of formal elegance. Moreover, it is not too difficult to present the essential features of the Machian approach in a manner suitable for an elementary treatment. The following is a description of a typical elementary treatment of mechanics from the Machian standpoint:

Take two trolleys equipped with light, frictionless wheels (or two gliders on a linear air track) and attached to each other by a string so that they compress a spring between them, and place them on a horizontal, frictionless surface. Burn the string and measure the (average) accelerations imparted to the trolleys by the expanding spring. Repeat as many times as desired, compressing the spring to a different extent in each case. It will be found that the ratio of the accelerations is always the same. Now place a block on one of the trolleys and repeat the above experiment. Again it will be found that the ratio of the accelerations of the trolleys, though different from what it was in the first set of observations, is constant. Evidently the ratio of the accelerations depends on some inherent property of each of the cars; if we call this property the <u>inertial mass</u>, we may sum up the results of the above experiment by means of the equation $\frac{M_A}{M_B} = \frac{|a_B|}{|a_A|}$, where the M's and the a's refer to the respective

/accelerations.....

accelerations (strictly, their magnitudes) and inertial masses of the trolleys. (Note that we define the ratio of masses to be the inverse ratio of the accelerations, in accordance with our intuition that the less easily accelerated body has the bigger inertia.) Thus the experiment provides, in principle, a means of comparing inertial masses; if some arbitrarily-chosen body (preferably in the shape of a trolley!) is selected as having unit inertial mass, other bodies can have their inertial masses determined by the method just described. We can now introduce force by defining it as the product of the inertial mass (as measured above) and acceleration. We can justify this procedure for the moment on the grounds that it is a fact of experience that to give a body a bigger acceleration, a bigger push or pull is needed, and the greater is a body's inertial mass, the bigger seems to be the push or pull needed to impart to it a given acceleration. The ultimate justification for defining force as the product of mass and acceleration would be the success with which a useful science of mechanics can be erected on this basis.

In the above outline, an appeal is made to intuition at certain points in order to reinforce the presentation of the basic ideas. It should be noted that although such a procedure has obvious pedagogic advantages, it is strictly speaking not really essential to a presentation of Mach's programme. As is customary in elementary treatments, the outline omits discussion of the "mediate" determination of masses, on the grounds that this is a subtlety which can be left out in the initial stages; also, it deals explicitly with only one kind of interaction between bodies, though both conceptually and in demonstrations it would be easy to extend the treatment in this respect. Whether it is desirable to introduce the subject in this way is a question which we have yet to discuss.

/The.....

The method of comparison of masses could easily find a place in an alternative presentation (see below).

Feather (1959, 1965), Orear (1961, 1967) and many others use an approach to mechanics which, while being Machian in spirit, pays lip service to Newton's formulation. That is to say, Newton's Laws are used in their conventional rendering, but Laws I and II are treated as definitions, with I a special case of II, and Law III is used to give operational meaning to the concept of mass, which appears in Law II (stated as "force equals mass times acceleration") but is at that stage still undefined. This is done by applying the Third Law as it stands to a pair of bodies in isolation (yielding $M_A\ a_A = - M_B\ a_B$ and hence the Machian definition of mass-ratio), or by deriving from the Third Law the Principle of Conservation of Momentum and applying that to a pair of isolated bodies to deduce the same result. Thus this approach, unlike Mach's, explicitly invokes the concepts of force and mass before the latter concept is defined; whereas in Mach's treatment the content of Newton's Third Law is an elementary consequence of the definitions of mass-ratio and of force, in the above-described alternative treatment the sequence of events is reversed, the Third Law being used to establish the concept of mass. Since Mach's second experimental proposition is essentially a re-rendering of the Third Law, it is fair to say that the above treatment lays bare the historical origin of the Machian approach while reaching the same conclusions, the consequent slight diminution in formal elegance being compensated for by a corresponding diminution in austerity.

4.7.4 Critique of Mach's approach to mechanics

Criticisms of the Machian formulation of Newtonian mechanics have been directed at a number of its aspects, viz.

/the.....

the limited usefulness of the definition of mass-ratio, the doubtful desirability of defining force in terms of mass and acceleration, the "frame-dependent" nature of the definition of mass-ratio, the doubtful logical status of Mach's definitions of mass and force, and the underlying philosophy of his approach. To see what is meant by the "limited usefulness" of the definition of the mass-ratio, let us consider the application of the definition to an isolated system composed of an arbitrarily finite number of particles (cf. Pendse 1937, 1939, 1940). If Mach's definition is to be of some use for this system, then it should be possible to infer from measurements made on the accelerations of these particles when they interact simultaneously with one another what their relative masses are. It is easy to show that this is not possible for a system composed of more than seven particles, and that Mach's definition can only be "rescued" for such a system if an assumption additional to Mach's formulation is made (e.g. the conservation of linear momentum - cf. Jammer 1957). Bunge (1968) has pointed out that the invocation of extra assumptions of this kind is at variance with Mach's attitude, since he sought to develop mechanics in such a way that the concepts of mass and force were purely derived concepts and therefore not to be used explicitly in statements of the fundamental assumptions of the theory. (Feather (1965) - who, incidentally, nowhere mentions Mach's name - has shown that a set of Machian-type propositions concerning accelerations and their ratios can give rise to a "complete mechanics" for any system of an arbitrarily finite number N of particles, in the sense that these propositions ensure that there are precisely (N-1) constant, independent ratios of the component accelerations of the particles towards one another, i.e. there are (N-1) unique relative masses.

/However.....

However, his analysis does not deal with the problem of how the component accelerations themselves are to be inferred from measurements on the system.)

It is readily evident that Mach's definition of mass-ratios can only hold for inertial frames of reference (cf. Jammer 1960 for a detailed discussion). Ehrlichson (1967) and Mueller (1968) have maintained in reply to Bunge (1966) that Mach intended his definition to apply only to inertial frames, Mueller pointing out in addition that conventional methods of mass measurement are, in effect, referred back to inertial frames. Bunge (1968) claims that this is merely supporting evidence for his view (Bunge 1966, 1967) that because Mach's definition of mass only yields correct mass-values when referred to a particular set of frames, it cannot be valid as a definition. Since this objection also applies to another common definition of mass, viz. mass = $\frac{\text{Force}}{\text{acceleration}}$ (force being treated here as a primary concept,), it should be carefully considered. Bunge (1966) suggests that "in mechanics, we do not accept the positivist tenet that bodies acquire a mass upon being subjected to mass measurement operations but assume instead that they have a definite mass value all along." It is indeed true that "we" (where "we" refers to virtually all physicists) do think of bodies as "having a definite mass value all along", and that the mass-values we assign to them when observing them from an accelerated frame of reference are those determined with respect to an inertial frame. We therefore have the situation that we conceive of mass as an objective property, continuously possessed by a body, and independent of our choice of frame of reference (at least in non-relativistic situations); yet, if we follow Mach, we define mass - i.e. say what it is - by means of an operational procedure which only gives the

/right.....

right mass value when referred to the correct frames. We can only escape from the contradiction by modifying our interpretation of Mach's "definition"; that is to say, we must regard it not as a definition of mass but as a means of specifying a procedure of measurement of mass which is valid under certain circumstances. As such it is still of fundamental importance in Newtonian mechanics.

It does not help in this instance to use what might at first seem to be an alternative escape-route, viz. to regard Mach's definition as a constitutive definition instead of an operational one. We can turn it into a constitutive definition if we regard the accelerations specified in it as not being immediately observed quantities but as constructs instead (cf. Margenau 1942). Since one is presumably at liberty to specify one's constructs relative to arbitrarily but suitably chosen frames of reference, one could claim that the ratio of masses of a pair of particles was defined as "continuously possessed" by the particles by means of this constitutive definition, instead of being only brought into existence by the appropriate measurements of accelerations. But there is another objection to be faced: Bunge alleges that Mach's definition is an example of a "logical mistake" in that it confuses an equality with an identity. He says that the equation $\frac{M_A}{M_B} = \frac{-a_B}{a_A}$ "establishes an equality between two expressions which differ in meaning and cannot therefore be regarded as the two sides of a definition." The left-hand side, that is to say, is the ratio of the measures of the inertias of the particles, while the right-hand side is a purely kinematical quantity. He concludes: "The equality is numerical not logical: it does not authorize us to eliminate one of the sides in favo(u)r of the other."

/Bunge.....

Bunge has advanced another objection to Mach's procedure which strikes at the root of the entire Machian formulation. According to Mach, the development of mechanics in the sixteenth and seventeenth centuries produced, in reality, only one "great fact" : "Different pairs of bodies determine, independently of each other, and mutually, in themselves, pairs of accelerations, whose 'terms' exhibit a constant ratio, the criterion and characteristic of each pair." As we have noted, Mach embodied this "great fact" in the experimental propositions with which he sought to supplant Newton's formulation, and from it he defined the mass ratio of a pair of particles. According to him, "in the concept of mass" [as thus defined] "no theory of any kind whatever is contained, but simply a fact of experience."
This assertion has been challenged by Bunge (1966), who claims that Mach's "great fact" "is not a direct expression of phenomena; from observation alone we could not infer that the components of a system move under the sole influence of their mutual actions: only an empirical checking of the consequences of such an hypothesis, when embedded in some theory, is possible." He goes on to say that the great fact is in fact a deductive consequence of Newton's postulates for the very special case of an isolated system consisting of two interacting bodies. Now it is true that the "great fact" can very easily be deduced from Newton's Laws; what we have to consider here is whether Bunge is justified in claiming that Mach was guilty of self-delusion when he asserted that his "great fact" could serve as an alternative (and superior) postulate to Newton's Second and Third Laws.

At first sight, Mach's view does not seem unreasonable. We can imagine carrying out an experiment in which two isolated bodies interact with each other by collision or under the

/influence.....

influence of gravitational, electrostatic or magnetic forces and in which the accelerations are measured and found to bear a constant ratio to each other. But we have to concede that no-one has ever carried out an experiment of precisely that kind. Certainly the two-trolley experiment described above is not such an experiment, since the two trolleys obviously do not constitute an isolated system. Can we allow for our failure to isolate them by using Mach's own formulation of mechanics? Mach's experimental proposition "d" suggests that we may treat the interaction of the "cars" with each other independently of their individual gravitational interactions with the earth and of their "contact" interactions with the table (or air track) which supports them. If the latter two interactions cancel each other out, as they seem to, then we are left with only the accelerations of the cars resulting from their mutual interaction. But this description makes it obvious that we cannot interpret the results of this experiment without reference to another so-called "experimental proposition." Consequently Mach's "great fact" cannot emerge from an experiment like this as an <u>independent</u> experimental fact, even if we interpret the results of the experiment entirely within the framework of Mach's own formulation of mechanics.

In this regard it is instructive to compare Mach's "experimental proposition" with Coulomb's Law in electrostatics, which has long been accepted as a genuinely experimental proposition. To establish this Law <u>directly</u> by experiment, we could use a torsion balance or similar device to measure the repulsive force between two similarly-charged small objects separated by various distances from each other. In so doing, we would make use of our qualitative knowledge of electrostatics to minimize the disturbing effects of induced charges on other bodies and of the perturbation of the

/distribution.....

distribution of charge on each body by the presence of charge on the other. In interpreting the results of the experiment we would successfully be able to avoid the embarrassment of using the inverse square law itself to allow for disturbing effects, and we would obtain a result which would suggest within the limits of experimental error that the law of force between two charged objects was an inverse-square one. (If we wanted to establish the law more accurately than this we could use the method of Plimpton and Lawton (1936) to show that there was no electric field inside a hollow charged conductor, this result being a simple consequence of the inverse square law and the Principle of Superposition.) We should simply be unable to achieve a comparable "isolation" of a pair of bodies intended to serve as the test bodies in a Machian-type experiment. We might, for example, be able to persuade a pair of bodies to interact with each other electrically as in the experiment just described, but in any terrestrial experiment we should be unable to prevent each of them interacting gravitationally with the earth and other objects. In order to interpret the results of such an experiment from a Machian viewpoint, we should have to "disentangle" the electrical interactions from the gravitational ones - and this we could not do without foreknowledge of the Law of Gravitation or, as we have seen, without use of another "experimental proposition", such as Mach's "d". In neither case, would we have achieved the desired goal of establishing the result as an independent experimental proposition.

Mach's "experimental proposition", then, is nothing of the sort. Can it serve as a <u>postulate</u>, in company with his other "experimental propositions", from which the rest of mechanics can be deduced? We have already noted the limitations of the Machian formulation, whether its foundations

/be.....

be regarded as propositions or postulates, in respect of systems of more than seven particles. It seems that we must agree with Bunge that the best that can be said for Mach's experimental propositions and definitions is that they are deductions from the Laws of Newtonian mechanics and not replacements for them.

Mach's definition of force by means of the relation "F = ma" has come under strong attack from Bunge and others. This definition did not originate with Mach. The motivation for defining force in this way and so turning Newton's Second Law into a mere definition seems to go back to Bishop Berkeley (see our earlier quotation). Following in Berkeley's footsteps, we have Maupertuis (1756, quoted in Jammer 1957) - declaring that "we should always remember that the concept of force is but an invention to satisfy our desire for explanation" and accordingly regarding Newton's Second Law as a mere definition of force. Later, d'Alembert, de Saint-Venant, Kirchhoff and others besides Mach defined force as "mass times acceleration" in their expositions of mechanics.

As we shall see below, this definition is open to challenge on logical grounds. But apart from this, dissatisfaction with it has stemmed from the feeling that "F = ma" does really tell us something important about the physical world, which a mere definition could not. If "force" were simply another name for "mass times acceleration", we should neither be able to verify this relationship nor predict anything from it. Yet in physics we commonly work with "laws of force" for gravitation and electromagnetism and attempt to discover them for the "strong" and "weak interactions", as though there really were a concept called "force" which has independent properties (cf Feynman et al 1963). In using "force" in such circumstances our intuitive idea of it is

/surely.....

surely that it is something other than just "ma". We can resolve this dilemma in two ways. One is to decide to treat force as an independent, and perhaps in the last resort undefinable concept; to this idea we shall return in due course. The other is to persist in viewing "F = ma" as a definition of force, but to formulate our mechanics in such a way as to make it clear why it is useful in the non-relativistic limit to define force as mass times acceleration and in no other way. This is the approach adopted by Eisenbud (1958), who formulates a "Law B" as follows, as constituting the effective, though implicit, content of Newton's Second Law: 'The acceleration of a particle may be correlated with observable properties of the environment and particle together with the position, velocity and time relative to the chosen system". On the basis of this law, he is able to define an acceleration function for a particle in a given environment. If this acceleration function is multiplied by the mass of the particle, the resulting product characterizes a property of the environment which is independent of the particle. This product is then called the force function or, more simply, the force acting on the particle of mass m when placed in the given environment.

Jammer (1957) in the year in which Eisenbud submitted his manuscript for publication, summed up what were essentially the same ideas in the following way: Experience shows that the product "ma" for a test body placed amidst a particular constellation of bodies (i.e. in a particular environment), is a function Φ of the total configuration under discussion. (Jammer defines "configuration" in this context as "the amounts of the gravitational masses, electric charges, magnetic moments, and so forth, as well as the geometric constellations.") If the configuration is denoted by the

/symbol.....

symbol X, this fact of experience may be stated in the form $ma = \Phi(X)$. Should the original test body be replaced by another of mass m^1, it will be found that $m^1a^1 = ma$; in general, for any test body, the product of its mass and acceleration is a single-valued function of the configuration X. Thus Newton's Second Law is "an empirical statement with respect to the fact that just the product 'mass times acceleration' is a single-valued function of the configuration X; it is an analytic statement with respect to the nominal definition of force". According to Jammer, force plays the part of a "methodological intermediate" in contemporary physics. When a body is placed in a certain environment, it is observed to move in a certain way. By saying that the environment gives rise to a force which in turn acts on the body, we are able to study the kinematical aspects of the body's behaviour separately from the physical situation in which it finds itself, and we can link the two, if we wish, by means of the concepts of force and mass. On this latter view, force and mass are useful but still dispensable concepts. However, as Eisenbud points out, the full significance of these concepts is brought out when we consider "the third law of motion" which, in Eisenbud's formulation, becomes: "Fixed mass numbers may be associated with the material elements of the system so that the total rates of change of momentum and angular momentum of the system are equal, respectively, to the impressed force and torque functions." We shall not elaborate on this re-formulation of Newton's Third Law, except to remark that it is significant that its author finds it expedient to use the concepts of force, torque, momentum and angular momentum.

From some of the criticisms considered above of Mach's definition of mass, it will be apparent that there are certain dangers attached to defining force as the product of mass and

/acceleration.

acceleration. For one thing, this definition is valid only for inertial reference systems, a fact which compels us to treat the definition as constitutive rather than operational (see our corresponding comments on the definition of mass). For another, this definition does not serve us well if we are dealing with particles travelling with relativistic velocities; as is well known, if we are to preserve the concept of force within the framework of special relativity, we have to write Newton's Second Law in the form $F = \frac{dp}{dt}$, where p is the momentum, i.e. in a form resembling Newton's original statement. A <u>definition</u> of force cannot be valid in some circumstances and not in others; therefore we should, to be safe, frame our definition of this concept in such a way as to be all-embracing. Of course to define force as $\frac{"dp"}{dt}$ would meet this requirement as far as special relativity and classical mechanics was concerned, since $\frac{"dp"}{dt}$ reduces to "ma" in the non-relativistic limit. But there are other troubles to be faced: if F = ma is a logical identity as implied by its use as a definition, then, suggests Bunge (1966), "masses and forces could not be assigned (hypothesized) independently. In particular, it would be impossible to assign masses to unaccelerated bodies." To see the force of these remarks, let us contrast this supposed definition of force with that of linear momentum (momentum = mass times velocity), which would appear to have the same form, and to which exactly the same objection might at first be thought to apply. It is certainly true that the definition of momentum tells us that a body can have no momentum unless it has a velocity - but this is exactly what we require of the definition. On the other hand, it would not be true to say that a body with no velocity could not be assigned a mass; if mass has been defined earlier by some other relation, or if it has been agreed to leave it undefined, then we are not compelled to turn the equation round and say that mass is <u>defined</u> by

/the.....

the quotient $\frac{\text{momentum}}{\text{velocity}}$. Difficulties arise with "F = ma" as a definition because in this case there is uncertainty as to which of the quantities, F or m, is really being defined. If we take m as being defined by some procedure not involving accelerations or as being undefined (neither of which decisions would meet with Mach's approval), then we shall be able to hypothesize masses independently of forces, but not the reverse.

To conclude this critique of "F = ma" as a definition, we shall take up certain implications of the above-quoted discussions by Jammer (1957) and Eisenbud (1958) of the "force function" (Jammer's $\Phi(X)$). Both these authors point out that a "force function" exists which is a "single-valued function of the configuration X" (Jammer), characterizing in certain cases "a property of the environment which is independent of the particle" (Eisenbud). But if this force function, which is equal to the mass times the acceleration of the particle, can be "independent of the particle", is it not a little strange to define it in terms of properties of the particle rather than of the environment? It is, of course, impossible to give a general definition of the force function in terms of the environment, but perhaps it is safer to conclude from this that the force function is indefinable, in the strictly logical sense of the term, than to seek to define the concept in a way which is logically insecure.

As both Jammer and Eisenbud make clear, the relation F = ma derives from a fact of experience, viz. that "just the product 'mass times acceleration' is a single-valued function of the configuration X" (Jammer). Suppose we put a particle of known mass into a given environment. It experiences a certain acceleration, which we measure; using F = ma, we calculate the force experienced by the particle. We can now

/predict.....

predict the acceleration that another particle of known mass will experience if placed in the same environment. We can make this prediction with confidence, not because somebody has "defined" force as mass times acceleration - no predictive power can reside in a mere definition -but because we believe that if the environment stays the same, the same force will be exerted on the new particle, so that its mass multiplied by its acceleration will have the same value as that obtained with the original particle. In short, "F = ma", instead of being just a definition of force, should be thought of as stating a relation between a property of the environment or "configuration X", characterized by the concept of force, and a property of the particle placed in that environment, characterized by the concept of mass.

4.7.5 Mass, force and relativity

As a footnote to the above discussion of the Machian approach to mechanics, we consider what influence, if any, the development of the theories of Special and General Relativity ought to have on our attitude towards the fundamental concepts of classical mechanics within the context of that theory. Gledhill (1968) has put forward the view that as force does not have the same status in General Relativity as it has in classical mechanics, while mass seems to retain its status, we ought after all to be content with treating force as a derived concept in classical mechanics, defining it as "mass times acceleration" in situations in which it is useful.

If the general theory of relativity and classical mechanics overlapped, in the sense that the two theories to a large extent shared their basic concepts, hypotheses and methods of approach, then it is perhaps conceivable that the ideas and results of the more advanced theory should influence

/our.....

our way of regarding the concepts of the more primitive theory. Special relativity and classical mechanics do overlap in this way. The effect of the application of the postulates of Special Relativity is to make it necessary for the definition of momentum to be modified somewhat (i.e. "mass times proper velocity" instead of "mass times velocity"), and to make it plain that Newton's Second Law must be stated in the form in which Newton apparently gave it in order to reflect the correct relationship between the force exerted on the particle and its motion under the influence of this force. One consequence of this is the disclosure of a relationship between mass and energy unsuspected in classical mechanics. However, it is significant that the status of the concepts of force and mass is not altered. Although Taylor and Wheeler (1966) claim that "to discuss both momentum and force in the context of relativity is to complicate the story", it is still true that if we are to consider the behaviour of any one particle or system, it is expedient to discuss this in terms of the resultant force acting on it, just as it would be in classical mechanics.

General Relativity and classical mechanics cannot be said to overlap in the way that Special Relativity and classical mechanics do. General Relativity and classical mechanics are two very different theories, constituted in very different ways and approaching the problem of particle motion- especially particle motion in a gravitational field - from very different viewpoints. The two theories overlap only in that they share a common domain of validity, i.e. General Relativity predicts the same results as classical mechanics in the domain in which the latter theory makes predictions which are confirmed by experiment. In classical mechanics one introduces mass and force as new "primitive" concepts essential to the theory, and one works with the hierarchy of inertial

/frames.....

frames of the "Newtonian" kind. In this context, one is compelled to employ the concept of gravitational forces and, as Newton was the first to show, one can establish a theory of gravitational forces which works extremely well - up to a point. In General Relativity, the belief in gravitational forces is abandoned, the hierarchy of Newtonian inertial frames of reference is replaced by local inertial frames defined by the motion of freely falling particles, and "gravitational" effects are explained as "tidal forces" arising from the local curvature of space-time. "Equations of motion" in this theory bear no resemblance to those in classical mechanics. In view of these deep differences between the two theories, it is difficult to see how the ideas and approach of the "wider" theory can usefully influence one's way of regarding the concepts of the more limited theory <u>within the context of that theory</u>. "Situations in which force is useful" are likely to be those in which we can meaningfully apply the ideas of classical mechanics and/or Special Relativity, and in such situations "force" must surely take on the status which those theories accord to it. As the foregoing pages will have made clear, differences of opinion still exist concerning that status, but the relationship between classical mechanics and General Relativity does not seem to be of the kind that can contribute anything to the resolution of these differences.

4.7.6 <u>Other approaches to mechanics</u>

There seems to be no avoiding the conclusion that however attractive Mach's re-formulation of classical mechanics might seem initially, it cannot stand up to detailed critical examination. This is borne out by the number of attempts that have been made in more recent times to set up formulations in mechanics which, while Machian in spirit,

/avoid.....

avoid some of the pitfalls of Mach's original approach. Little purpose would be served here by detailed discussion of these alternative formulations. The best-known of them among those actively concerned with physics teaching is probably Eisenbud's (1958), from which we have already quoted in the course of the above discussion of Mach's definition of force. A number of authors have already acknowledged its influence on their work (cf Austern 1961, Weinstock 1961, and Halliday and Resnick 1966), showing that it has been and is being taken seriously. Although it represents a considerable advance on Mach's presentation, its underlying philosophy seems to be much the same, in that it places primary emphasis on "observational concepts" such as acceleration and treats mass and force as derivable from these. However, as will have been evident from its "Law C", quoted above, these concepts are stated or plainly implied in its fundamental laws, so that there is no unrealistic attempt to derive the whole of mechanics from alleged "experimental propositions" of a kinematic nature. Also, as we have seen, what is claimed to be the effective content of Newton's Second Law is made the substance of Eisenbud's Law B. But since Eisenbud still defines mass in terms of an acceleration-ratio (though not, we must in fairness add, of the accelerations of an isolated pair of bodies) and force as the product of mass and acceleration, his presentation remains open to the kinds of objection already considered. This being the case with his and other similar formulations of mechanics, one is led to wonder whether the whole enterprise of trying to establish mechanics in this way is not perhaps misguided, and if it is, what the alternatives are. To reach a decision on this point, let us pay some attention to the underlying philosophy of Machian-type approaches to mechanics which, as has been pointed out, is that <u>any</u> concepts, in particular those of mass and force,

/must.....

must be treated as observational; that is, concepts of any kind are properly definable only in terms of measurement operations. We have already noted the limitations of this point of view (cf. Section 3.2), abandonment of which would enable one to avoid the difficulties encountered by Machian-type approaches to mechanics. If, that is to say, we ceased to insist on the need to define force and mass by operational or other means, we should clearly be able to avoid the embarrassments provided by adoption of the usual "definitions" of those concepts. Nor would this be a case of throwing out the baby together with the bath water, the theory founded on these concepts and associated laws or axioms is readily capable of suggesting procedures by which these undefined, "primitive" concepts can be measured and so provide the linkage between these concepts and observations.

A possible approach, then, is to choose "mass" and "force" as among the basic undefined "primitives" of the theory (cf. Bunge 1967, Truesdell 1968) and to postulate suitable axioms or laws from which the theory can then be developed in terms of the "primitives". The choice of the new concepts of mass and force is not, of course, made out of the blue. The one is associated with our anthropomorphic ideas of inertia and the other with our anthropomorphic ideas of pushes and pulls and, as Newton was the first to show, these concepts can be linked together in such a way that (i) we can find a means of measuring them and (ii) a fruitful theory can be erected upon them. As axioms of the theory we can take Newton's own, or at any rate the usual modern versions of his Second and Third Laws. (For present purposes the First Law can certainly be treated as a special case of the Second.) These are not the only axioms we should need in a fully-developed "axiomatisation" of mechanics (for an example of a full list see Bunge (1967)), but they are the ones which have the most

/interesting.....

interesting physical content. Are Newton's Second and Third Laws "experimental propositions" in the Machian sense? If by this we mean "propositions directly verifiable by experiment", then the answer is undoubtedly no. We have referred earlier to the difficulties attendant upon any attempt at direct verification of the Third Law; similar, though less obvious difficulties arise in connection with direct verification of the Second, as we shall explain when examining teaching approaches to Newton's Laws. It is better to regard them as axioms or postulates whose consequences can be put to experimental test.

Before we address ourselves at last to the question of how "F = ma" is to be taught at the elementary level, we feel it appropriate to append the following quotation from Hesse (1964) to our discussion of rival approaches to the teaching of mechanics: "We need not regard any one view of the nature of force, or any one formulation of the laws of motion, as being uniquely correct. Mechanics is rich enough to contain most of the logical aspects of the other sciences - empirical, definitional, a priori, and conventional, and alternative formulations may be appropriate for different purposes." For our own immediate purpose, we shall adopt the axiomatic formulation sketched out in the previous paragraph, partly because by this means we can avoid arguments about the best way in which to define mass and force, and partly because it is easy to relate this approach to recently-evolved techniques of teaching "F = ma" in the laboratory and the classroom.

4.7.7 Attitudes to teaching Newton's laws

Traditionally, dynamics begins with Newton's Laws of motion, from which one proceeds swiftly to units of mass and force and "force equals mass times acceleration." Let us quote Rogers (1960) on this point: "In most elementary

/textbooks.....

textbooks, Newton's Laws are stated formally at the beginning of the chapter instead of at the end where they properly belong as summaries. And they are announced so firmly that students think that Newton got them from Heaven. He did not - he was merely re-wording the views of Galileo and others who had experimented and thought about motion. He gave them as working rules, partly based on experiment, partly definition and clarification of terms." We feel that the points made here by Rogers are valid, and plead earnestly that the subject should not be treated in the way he criticizes. On no account must the pupil be allowed to gain the impression that his first act in beginning the study of a fundamentally important branch of physics is the swallowing of an approved set of dogmas. The intention should be to let the pupil develop a feeling for what lies behind the Laws of Motion before he hears or sees them stated. He can do this by carrying out certain simple experiments and by taking careful note of certain facts from everyday experience. If properly guided, he should not be in the position of the unfortunate student who, as has been related in an article in the American Journal of Physics (Weinstock 1961), complained to his friends "Sure I know F = ma, but what's F, what's m, what's a?".

For the sophisticated student, i.e. one who has acquired some prior familiarity with Newton's Laws, it is indeed possible to begin with a simple assertion of these Laws, closely followed by a careful analysis of their meaning. This is the approach followed by Feynman et al (1963), who present a most illuminating discussion of the Laws, in the course of which these authors vent their dissatisfaction with the view of Newton's Second Law as a mere definition, and are careful not to produce a clear-cut definition of mass. But this is a presentation intended for the elite freshmen of "Caltech"; and even they, one would venture to suggest, would obtain

/little.....

little benefit from the discussion of Feynman et al unless they had previously made some experimental study of the relationship between force and acceleration.

4.7.8 Newton's First Law

Although, strictly speaking, we can regard the First Law as a special case of the Second, there are clear pedagogical advantages of dealing with the First Law as such. The concepts of motion-going-on-in-the-absence-of-force, motion-going-on-when-forces-are-balanced, and of inertia, the property of a body which is revealed by its tendency to persist in its state of rest or uniform velocity, can all be brought out in the course of the study of demonstrations and experiments illustrative of the First Law. This will provide the opportunity for a more formal study of inertia with pupils than was possible or desirable in the junior years. The pupil can see the concept of inertia evolving as a conclusion from controlled experiment.

The approach used can in some respects be closely parallel to the early historical development of the subject, with particular reference to Galileo's pioneering work. Galileo's evolution of a principle of inertia by a process of simple experimenting allied to hard and clear thinking was a major step in man's understanding of motion. Pupils can carry out for themselves the Galilean (thought?-) experiment of allowing a ball to roll down one inclined plane and up another facing it, from which they will very likely conclude that in the absence of friction, the ball would attain to its original height on rolling up the second plane, no matter what the inclination to the vertical of the latter might be. (The Nuffield Guide III recommends the use of a steel ball rolling on curtain rail "hills", very firmly supported, for this experiment.) Pupils should also carry out Galileo's "pin-and-

/pendulum.....

pendulum" experiment (cf. P.S.S.C. Laboratory Guide and Nuffield Guide III), which is a miniaturized, almost friction-free version of the previous experiment. Galileo proceeded from the results of these experiments to a celebrated extrapolation: that if the second "inclined plane" were made horizontal, the ball would roll on forever in the absence of friction. (As various authorities have pointed out, Galileo interpreted the term "horizontal plane" to mean "a plane parallel to the surface of the earth". In other words, for Galileo what we now call inertia meant a tendency to move with uniform motion in a circular path. See Toulmin and Goodfield 1960, Cohen 1961, Drake 1964 and Losee 1966.) If possible, pupils should draw this conclusion for themselves.

If time permits and if suitable apparatus is available, it is highly desirable to reinforce this conclusion by showing demonstrations of near-frictionless motion. This is very conveniently done with solid carbon dioxide: for example, if a block of dry ice of about 6 inches by 6 inches by 4 inches is available it could be allowed to coast on a large and carefully-levelled sheet of glass; or a metal ring covered by a lid, with some dry ice under the lid could be allowed to coast on a smooth table (cf. Nuffield Guide III). As an inferior but useful alternative the P.S.S.C. authors give a design for an "air puck" that can be made from a piece of plywood, a rubber stopper and a balloon (P.S.S.C. Teachers' Resource Book and Guide, Part Three). In recent years, so-called linear air troughs and tracks (Neher and Leighton 1963) and even "air tables" (Daw 1963) have been developed for the performance of fundamental mechanical experiments under well-nigh frictionless conditions. In their refined form, these items of apparatus are very expensive and cannot realistically be recommended for purchase by schools. However, much smaller

/length.....

(length less than 1 metre) and cheaper versions of air tracks are now obtainable; indeed air tracks of this type can be made cheaply in a reasonably well-equipped workshop (cf. Jardine 1966, Ritchie 1966). The basic requirement is a length of a yard or two of Dexion Speediframe (hollow tubing of rectangular cross-section) in which holes of 3/64 inch diameter are drilled at $\frac{1}{2}$ inch intervals along the top and the sides. The air supply is provided by a vacuum cleaner, and the "gliders" that ride on the cushion provided by the escaping air can be made from pieces of Perspex sheet of $\frac{1}{8}$ inch thickness. The disadvantages of air tracks for school use is that they require fairly sophisticated auxiliary equipment for quantitative experiments. To illustrate Newton's First Law, for example, it would be desirable to show that when a "glider" is acted upon by no net force, it moves with constant velocity, and unless the glider is able to travel an appreciable distance on its air track, the constancy of its velocity will not necessarily be obvious to those watching the experiment. (On the other hand, the tendency for its motion to persist will be obvious if the "glider" is allowed to rebound from elastic bands placed at each end of the track; Ritchie (1966) reports that a glider would make as many as 50 journeys on a 6-foot track of the simple type described above.) To show the constancy of the velocity of a glider on an air track, or a dry ice puck on a suitable surface, the techniques of stroboscopic photography could be employed (cf. Ritchie 1966, Dorling 1966, Nuffield Guides III and IV). Certainly a properly-equipped school laboratory ought to have a good 35 mm camera available or, better still, a Polaroid Land camera. The latter has the advantage that positive prints are obtainable within a few seconds of taking the photograph, but it is comparatively expensive. The article by Dorling (1966) describes techniques by means of which 35 mm film may be processed in a matter of minutes.

/To.....

To obtain multiple photographs taken at successive equal short intervals of time, the apparatus can be illuminated by means of a stroboscope with a high light output - such an item of equipment is only moderately expensive and is well worth the cost - or a "mechanical shutter" in the form of a rotating slotted disk can be placed in front of the camera. Stroboscopic photographs of the motion of a freely-moving "glider" or dry ice puck would show that the object underwent equal displacements in equal time intervals.

Some teachers fear that to the unsophisticated pupil, it might appear that the gliders on the air track are being driven by the air escaping from the holes along the track. Thus Gledhill (1968) argues that "It requires a fairly sophisticated pupil to see that the air will not in fact drive the glider." Obviously, if pupils are not convinced that the air merely provides an almost-frictionless "cushion" on which the gliders can travel, then demonstrations on the air track lose all their point. It is therefore important to let pupils see first of all that a glider originally at rest will remain at rest on the track if left to itself. Provided that the track is well aligned and level, and the weight of the glider is symmetrically distributed around its centre of gravity, this will be true to a good approximation. (Very slight departures from the horizontal can produce appreciable drift velocities: the manufacturers of the Stull-Ealing Air Track point out that on "an ideally straight track, a constant slope of just ten seconds of arc would sustain a drift velocity of from 5 - 10 cm/s".) Then one can show that if the glider is given a particular speed in a particular direction, it will sustain that speed for at least the length of the track; this should be done for various speeds and each of the two possible directions along the track. If there are pupils who still believe that the motive power for

/the.....

the glider is provided by the air, they now have to explain why, if air blowing out of the holes along the track does not start a glider moving, it can keep it moving in a given direction once it has been started, and how the air can push two gliders along the track towards each other. Perhaps they can be persuaded that the hypothesis that the gliders simply ride on the air provides a much more reasonable explanation of what happens.

Similar difficulties could arise with air pucks and dry ice pucks. It has been brought to the writer's notice that some science teachers were troubled by the idea that a dry-ice puck ought perhaps to move off on its own, because of the loss of mass brought about by sublimation. Those who are inclined to believe in this interesting violation of the Second Law of Thermodynamics ought to be encouraged to suggest why the sublimated CO_2 should decide to move off in a particular direction.

The demonstrations discussed above are concerned with the motion that takes place when there are no forces acting along the direction of motion of the vehicles under study. In order to banish any lingering, Aristotelian doubts about relationship between force and velocity, it is also highly important that pupils should be convinced that when no net force acts on an object, its velocity is unaltered. Of course, this fact plays a vital part in the demonstrations already described: the "gliders" or dry ice pucks rest on cushions of gas, they experience no net force in the vertical direction and have zero velocity in that direction. The attention of pupils should certainly be drawn to this aspect of the situation. The Nuffield Guide (IV) suggests an explicit demonstration of the fact that when balancing forces act on an object, it experiences no acceleration. The essence of such a demonstration is to show that when a pair of forces

/which.....

which are known to be equal are applied in opposite directions to an object, the latter will be unaccelerated. The difficulty here is how to know that the forces applied to the object are in fact equal in magnitude. The ready answer is to apply the forces by means of calibrated spring balances (e.g. by those calibrated in newtons and used during the junior years). In reality, the spring balances measure the forces exerted on them (cf. our earlier discussion) and in assuming that when they give the same reading, they apply the same force to the object, we are applying Newton's Third Law. This difficulty is inescapable; but since pupils should by this time be accustomed to interpreting the readings of spring balances in this way, and should have made the acquaintance of the Third Law, it need not be a source of worry to teachers.

Two variations of this type of demonstration are suggested in the Nuffield Guide (IV). In the first of these, it is shown that when an object is dragged across a rough surface, the forward force is equal to the frictional drag. In the arrangement suggested in the Nuffield Guide, the object in question is a brick resting on a plank (the "rough surface") which in turn rests on metal rollers to reduce friction between the plank and the table to a minimum. One spring balance is used to pull the brick across the plank at constant velocity; another is used to hold the plank stationary. Thus the forces acting on the brick are the forward pull exerted by the spring balance, and the frictional drag exerted by the plank; the forces acting on the plank are the forward drag exerted by the frictional force between the plank and the brick, and the backward pull exerted by the other spring balance. Each spring balance records the tension in it, resulting from the action of equal and opposite forces on its ends. Then if Newton's Third and First Laws are true, we may conclude that equality of the readings on the spring

/blances.....

balances shows that the forces on the brick balance. Therefore we cannot interpret the results of this experiment fully unless we assume the very Law whose validity we are seeking to establish, a fact of which the author of the Nuffield Guide is fully aware, as his footnote shows. However, we can agree with his summing-up that "... it is a demonstration of a real event in the physical world and it does supply some information to people who might have expected a different result. Therefore, it does convey some knowledge and should be shown."

The other arrangement suggested by the Nuffield Guide leaves out the effect of friction, the two opposing forces on the test vehicle both being supplied by spring balances. The test vehicle this time becomes the plank itself, once again riding on rollers, and spring balances are attached to opposite ends of it and so arranged as to pull the plank with constant velocity. This demonstration meets with similar difficulties of interpretation to those discussed in the preceding paragraph, but for the reasons indicated there, it is nevertheless a good idea to carry it out.

After this preliminary experimentation has been completed, it would be fitting to sum it up in a formal statement of "Newton's First Law" in a fully modernized rendering, such as "Every body perseveres in its state of rest, or of uniform motion in a straight line unless acted by a resultant (or net) force." Such a statement should preferably follow attempts by the pupils themselves to produce a suitable summing up. As to the interpretation of this Law, we should with school pupils make this the most straightforward possible, viz. that the Law means what it appears to say - that an object experiencing no resultant force is subject to no acceleration, and that if an object is undergoing acceleration, it must therefore

/be.....

be subject to a resultant force. We may use the Law, in other words, in order to provide us with a simple criterion for the presence or absence of a resultant force. As for the implied frame of reference relative to which the acceleration is specified, we trust that matter has been adequately dealt with in Section 4.6.

4.7.9 Newton's Second Law: introduction. An important preliminary experiment in the establishment of Newton's Second Law is that which shows that a constant force produces constant acceleration in a given object. The basic features of the experiment are the application of a steady pulling force to a suitable vehicle which moves under nearly frictionless conditions, and the measurement of the resulting acceleration. It has now become standard practice in many progressive school laboratories to use the method and apparatus for this experiment developed by the P.S.S.C. (cf P.S.S.C. Laboratory Guide), who have designed the now-ubiquitous robust, three-wheeled trolleys suitable for use in this and other mechanical experiments. As a "runway" for such trolleys, a table of more than ordinary smoothness and approximation to a plane surface is required. The Nuffield Guide (III) suggests a plank of 1 inch plywood or chipboard, 1 foot wide by 8 feet long, with lengths of slotted angle iron on each side to act as fences.

To provide the constant pulling force, a long thin rubber band stretched by a constant amount is used. One way of ensuring that the amount of stretch is kept constant is to use a rubber band or thread that is about two-thirds of the length of the trolley, attach it to a post at the rear of the trolley and stretch it so that the other end is just above the front of the trolley (cf Nuffield Guide III). The experimenter then walks (or runs!) along next to the trolley, keeping the rubber band stretched to the same length. Alternatively, if a longer rubber band is used, one may hook

/one.....

one end of it to the peg at the front of the trolley and the other over the end of a metre rule held parallel to the trolley; one then pushes on the rule and moves it along so that a particular graduation on the rule is kept opposite some fixed mark on the trolley (cf Brommert 1969).

We have mentioned earlier that the displacements undergone by the trolley in successive time intervals can be measured by attaching a length of ticker tape to the trolley and making marks on it at regular time intervals. The standard technique is to pass the ticker tape under a disc of carbon paper and over a piece of hard wood; an electric bell clapper is allowed to bang on the carbon paper. This method of marking displacements in equal time intervals supplants that long in use with "Fletcher's trolley" and other variants of it, in which a pen attached to a vibrator drew a wavy line on a strip of paper. The drawback of this method is thus summed up by the Nuffield Guide (III): "Few children see readily what the relationship is between the obvious wavy trace and a distance or a speed". The disadvantages of the bell-clapper-marking-ticker-tape method are that the vibrators (bell units) are sometimes temperamental in adjustment, and that the passage of the ticker tape under the carbon paper disc and the clapper produces a small frictional force which is not always negligible. Shuter (1965) has described the modifications to P.S.S.C. trolleys and timing systems that have been incorporated in the apparatus used in the first-year physics laboratory at Rhodes University. These include the use of rectified A.C. from the mains to operate the bell-timers, enabling the frequency with which the clapper strikes the paper to be precisely known (50 Hz in South Africa), and the addition of a vibrator unit to the trolleys. Measurements carried out in this laboratory have shown that the bell-timer unit used in this way produces a frictional force of about 0.1N, which is

//readily.....

readily comparable with the pulling forces normally encountered in experiments with P.S.S.C. trolleys. It has therefore been found expedient to use the bell-timers only in an experiment in which "g" is determined for an object which while falling "freely" drags a length of ticker-tape through a vertically mounted bell timer. In the trolley experiments, the vibrator is used to produce a wavy trace on a sheet of paper mounted on the table on which the trolley runs. In this case the friction is negligible compared with that produced by a bell timer operating at 50 Hz. In school laboratories, it is probably best to use bell timers for these trolley experiments, for the reason already mentioned. The presence of friction will not confuse the issue in the experiment demonstrating the constancy of acceleration for a constant pulling force, provided that it remains substantially independent of velocity. This appears to be the case with the trolleys now in use. In other experiments requiring the use of trolleys, we shall see that the presence of friction can be explicitly taken into account without causing undue complication.

It is desirable that pupils should not simply take for granted any figure quoted for the frequency of the bell timer. By means of a hand stroboscope they should be able to make their own measurement of this frequency.

4.7.10 Newton's Second Law: possible approaches to F = ma

There are two main objectives for the teacher in presenting "F = ma" :-

(i) the relationship "F = ma" must be established;

(ii) satisfactory procedures for measuring the two new concepts brought together by this relation, viz. mass and force, must be developed. It is now common practice to make (ii) follow from (i) in elementary physics teaching. Round about the turn of the century, however, many physics textbooks

/treated.....

treated mass as a quantity for which a method of measurement independent of Newton's Laws was available; that is, they defined mass in terms of weight (cf Jammer 1960). Recently (Armstrong 1965b, 1966 a) an attempt has been made to revive this viewpoint, and to suggest in addition that the terms "weight" and "mass" should be used synonymously. Since these ideas might be found attractive by the inexperienced teacher, we shall consider their implications in some detail before going on to discuss point (i) above.

Typically, a presentation of this type proceeds along the following lines (cf. Duncan and Starling 1918): The weight of a body is defined as the attractive force which the earth exerts on the body. Then the procedure of measuring mass with the aid of an equal-arm balance is described, and gravitational units of force are introduced. To introduce Newton's Second Law, the following argument is used (Duncan and Starling 1918): "Since the weights of two bodies are proportional to the masses, and since both bodies fall freely with equal accelerations, it follows that the forces required are proportional to the masses if equal accelerations are to be imparted to a number of bodies." (This statement is reproduced without acknowledgement in Gordon, Neser and Pienaar, no date; these authors tack on to it the remark that "This conclusion is a pointer to Newton's Second Law, which can be proved experimentally".) It is then said that "a laboratory experiment may be devised to illustrate another law, viz. the force which must be applied to a body of given mass is proportional to the acceleration required." For details of the experiment one is referred to a later page, where the experiment in question turns out to be concerned with the use of Atwood's machine. Finally, these facts are summed up in the equation "F = ma".

/As.....

As outlined here, this approach does have the merit of being logically consistent. It is, however, open to a serious objection on methodological grounds. It seems unfortunate that the assumption should be made at the outset that mass and weight are proportional. They are, of course; but if we can build mechanics without making use of what Jammer (1960) calls "a purely accidental aspect in classical physics", viz. the proportionality of weight and mass, then surely it would be better to do so. (As we have already noted, Newton himself found it necessary to establish experimentally that weight was proportional to mass.) And there is a further difficulty about defining mass in terms of weight; what is one to say about the mass of such an object as the earth (cf Dadourian 1939)? As for Armstrong's view that we should let what he calls "the measure of the body" be either "mass" or "weight" and so "[strike] a blow against senseless pedantry" and to use as a fundamental equation in place of "F = ma" the equation "$\frac{a}{g} = \frac{P}{W}$", where P and W are to be measured in the same units, we wonder how he can seriously believe that this approach is less confusing than the usual one, especially when he says that "the force of gravity on [a body] can still be called 'weight', just as .... the 'length' of a journey can refer either to distance or time." What his proposal amounts to is that whereas it is customary in physics to use two different words, mass and weight, to refer respectively to two different concepts which are measured in different units, we should now use either or both of these words to mean the same or different things, just as we please, measured in the same or different units, just as we please. Comment seems superfluous; but we wonder how, in terms of this proposal, one would cope with the term "weightlessness".
In short, though it is indeed possible to approach Newton's Second Law via mass measured in terms of weight, such an

/approach.....

approach is unnecessarily inelegant; and we might be well advised to stick to tne allegedly "senseless pedantry" of using tne terms "mass" and "weight" with tneir traditional meanings in physics.

In view of tne points just discussed, we recommend tnat in tne teaching of Newton's Second Law, the fundamental metnods of measuring mass and force should emerge from a study of tne law itself and not be treated as previously established. Since pupils will have been introduced in the junior years to measuring masses by means of balances, they should have it clearly explained to them that in this more advanced study of mechanics, they will discover why they were justified in measuring mass in this way.

It will have been deduced from our remarks in connection with various formulations of mechanics that we are not likely to advocate that "F = ma" should be taught to pupils as the defining equation for force. Indeed, even if one did accept that force was properly defined in this way, one could still argue that it would be unwise to present the concept in this way at an elementary level. The unsophisticated pupil will scareely be able to see the point of "defining" force in terms of mass and acceleration unless he has seen some experiments which suggest that these concepts could be linked like this. And if one is to perform experiments in this connection one might as well introduce "force equals mass times acceleration" as a relationship apparently established by experiment. This is the way in which the topic is treated in the P.S.S.C. and Nuffield Physics courses, and this is the approach we shall discuss here.

The teacher should be aware from the outset that the experiments he and his pupils will be conducting in classroom and laboratory are hardly experimental "proofs" of Newton's

/Second.....

Second Law; "proof" is altogether too big a word to use here. Instead, these experiments should be regarded as providing strong hints as to the correctness of the Second Law, the ultimate "proof" (if such a thing really exists) being the success of the theory of mechanics erected upon it. In the design and interpretation of these experiments certain assumptions are made; some of these should be made known or be obvious to pupils and others will not be. Since these assumptions help to reveal the uncertain status of these experiments as "proofs" of "F = ma", the teacher should certainly be aware of them and we shall endeavour to state them clearly in what follows.

There are various ways of investigating the relation between acceleration and force experimentally. In the most sophisticated of these, the first step is the establishment of a scale of force by an experiment in which a spring balance pulls an object whose acceleration is measured. The spring balance is calibrated in "force units" by making its scale readings proportional to the accelerations imparted to the object, the latter being treated as a "standard object". (Note that such a calibration requires no assumptions about the properties of the spring, while the reproducibility of the readings of the spring balance can be checked by experiment.) By experiment it is then shown that force as measured on this spring balance scale and acceleration are proportional for any body, although the proportionality constant is, in general, different for different bodies. (For textbook treatments using this approach, see Holton 1952, Arons 1965 or Halliday and Resnick 1966.)

This method of introducing Newton's Second Law is quite elegant, but it has a number of disadvantages for use in school. On the "theoretical" side, it would be difficult to

/persuade.....

persuade pupils of the virtues of initially setting up a force scale by setting the force proportional to the acceleration of the body chosen as standard. On the "practical" side, this approach requires some very careful experimental work which will have to be done under essentially frictionless conditions and therefore requires the use of dry ice pucks or "gliders" or air tracks. While one can imagine a skilled teacher carrying out these experiments and obtaining the desired results, one cannot for a variety of reasons see the pupils achieving this - and if "$F \propto a$" is not to emerge from experiments done by the pupils themselves, a great deal will have been lost. Incidentally, it should be realized that the assumption is made in experiments of this type that the spring balance exerts the same force on different bodies when the tension in it (given by its scale reading) is the same. This means, in effect, that we are assuming Newton's Third Law in order to interpret the results of the experiments. We shall find that there is no way of avoiding this assumption in interpreting any of the experiments by which we seek to establish the relation "$F \propto a$". (A recent attempt to <u>formulate</u> Newton's Second Law (Katz 1967) without reference to the Third Law does seem to be successful, but the author does not appear to have substantiated his claim that his approach makes it possible to <u>measure</u> force and mass without implicit use of the Third Law.)

An alternative experimental approach might be to use a spring balance calibrated by some other means to apply known forces to various objects. There is no objection in principle to using a spring balance calibrated by hanging different numbers of <u>identical</u> "weights" from it, as described in our treatment of the presentation of force to the junior standards (Section 3.8.3). On the other hand, there would be serious objections to using a spring balance calibrated by

/hanging.....

hanging "weights" from a weight-box from it, or to applying forces to the "test vehicles" by means of a string passing over a pulley and attached to a scale pan containing such "weights". These latter metnods would violate our principle of not wishing to make use of the proportionality of weight and mass in building up the fundamental principles of mecnanics. The disadvantage of using even a "correctly" calibrated spring balance in these experiments is that this makes the relationship between force and acceleration seem a little less direct than it is in the kind of experiment to be described below.

If we are not to apply forces by means of a spring balance or nanging weights, we need some other means of applying to a body forces having magnitudes that are known multiples of some standard force. This is easy enough if we apply these forces by means of different numbers of identical rubber bands or springs stretcned by tne same amount. We make tne assumption that wnen a trolley (say) is pulled by two identical rubber bands each stretcned by tne same, specified amount, the force exerted on it is twice tnat exerted by a single rubber band stretcned by tnis amount. Use of more rubber bands enables further multiples of tne "rubber band unit" of force to be obtained. Our experiment for investigating tne relation between force and acceleration can therefore take tne form of applying different numbers of "rubber-band units" (or "spring units") of force to the same trolley and measuring tne resulting acceleration in each case (P.S.S.C. textbook, laboratory guide and teachers' resource book and guide (1960), Nuffield Guides III and IV). In tnis experiment tne pupil is effectively applying tne force himself, and he has a direct "feel" for what is meant by twice, tnree times or four times tne same force. (For a description of the modification in

/use.....

use at Rhodes University in which springs are used to apply the force to the trolley, see Shuter 1965.) A plot of force versus acceleration should produce a ~~reasonable~~ straight line which does not pass through the origin, but cuts the force axis at a point slightly above it. This positive intercept on the force axis represents the total frictional force on the trolley; it should be made clear to pupils that if this force is deducted from the applied force, the plot of resultant force versus acceleration must be a straight line through the origin, showing that the two quantities are proportional. As an alternative to this treatment of friction, the technique of tilting the runway to compensate for friction can be adopted (cf Nuffield Guide III, Jardine 1966). If this were done, it would of course be best to discuss the technique with pupils beforehand, and if possible to have them suggest the correct method of friction compensation.

At this stage, the concept of mass has not yet been officially introduced, while force has been treated as a "primitive" fundamental concept for which no definition need be supplied. There are various ways in which we could now proceed to introduce mass and so arrive at "F = ma".

(i) Following on the experiment showing that a and F are proportional, an experiment can now be performed in which the same force is applied to different numbers of trolleys piled on top of one another and the acceleration is measured in each case. The results show that for a given force, the acceleration is inversely proportional to the number of trolleys. If the assumption is made that the mass of n trolleys is n times the mass of one trolley, this result may be expressed in the form that for a given force, the acceleration produced is inversely proportional to the mass moved; in symbols, $a \propto \frac{1}{m}$. When this result is combined with the result that

/F.....

$F \propto a$ for a given m, we have that $F \propto ma$, or $F = kma$. The units of mass and force are so chosen as to make the proportionality constant k equal to unity, and we finally have $F = ma$. This the approach followed in the Nuffield course (cf Nuffield Guide IV), by Jardine (1966) and Taylor (1963).

(ii) Instead of making the assumption that mass is additive, one can proceed by other means to find a method of measuring mass that is independent of a measurement of force, and then proceed to show that $a \propto \frac{1}{m}$ for a given force. Thus it can be shown by experiment that if the same force is applied to two different bodies, their accelerations will in general be different; and that if a <u>different</u> force is applied to the <u>same</u> pair of bodies, a new pair of accelerations will be obtained, but the <u>ratio</u> of the accelerations is the same as before. If we now <u>assert</u> that the ratio of the accelerations is <u>equal</u> to the inverse ratio of the masses, in accord with our intuitive ideas that the body undergoing the smaller acceleration has the bigger inertia, then it follows from the way in which we have now chosen to measure mass that $a \propto \frac{1}{m}$ for a given force. As before, combination of this result with $F \propto a$ yields $F = kma$ or $F = ma$ with correct choice of units. This is essentially the approach recommended for "abler and more mathematically inclined pupils", by the Association for Science Education (1963).

As an easier version of this approach, Gledhill (1968) has suggested that one might introduce mass by arguing simply that it is reasonable to take one's measure of the inertia of an object to be inversely proportional to the acceleration imparted to the object by an arbitrarily chosen force. Then one does not have to concern oneself with <u>ratios</u> of accelerations or masses. This approach then becomes very close to that dealt with as (iii) below. Sooner or later,

/however,.....

however, the pupil must come to appreciate the connection between the mass ratio and the acceleration ratio for a pair of particles subjected simultaneously or in turn to the same force. The reason is that our methods of mass measurement are based on the choice of a mass standard, and this means that practical measurements of mass must be based either directly or indirectly on comparisons with the mass standard. When the pupil is introduced to mass through dynamical experiments, as is suggested here, the form of comparison of masses that he meets first is that based on measurements of acceleration. Thus he literally will not be able to understand how mass is measured unless he understands this point. (iii) More direct than either (i) and (ii) is the approach in which, having shown that $F \propto a$ for a given body, one proceeds to carry out similar measurements to show that a is also proportional to F for, say, a trolley with an additional load placed on it, but that the proportionality constant in the relation $F = Ka$ is now different (larger, in fact). The proportionality constant K is now treated as a measure of the inertia of the body (or, as most textbooks using this approach would say, it is defined as the inertial mass of the body), since it is clearly dependent on it. Since "mass" is the name we give to the quantity we associate with the inertia of the body, we can regard the constant K as representing the mass of the body; and so we can replace K in the relation $F = Ka$ with m, where m now stands for mass. Hence we finally obtain $F = ma$. (In view of what was said earlier about the difficulties of defining mass and force, we would recommend that the relation $m = \frac{F}{a}$ not be taken as a definition of mass, but rather as a relation showing how it can be calculated from measurements of force and acceleration.) This is essentially the approach used in the P.S.S.C. textbook (1960, /1965).....

1965) by Arons (1965), Sears and Zemansky (1960, 1964) French and Hudson (1965) and many others.

There can be no doubt that on grounds of economy, approach (iii) is superior to approaches (i) and (ii). Approach (i) requires the assumption of the additivity of masses of similar objects and the performance of an additional experiment which shows that $a \propto \frac{1}{m}$ for a given F; approach (ii) requires the performance of an additional experiment to establish the constancy of the acceleration ratio for a pair of bodies acted upon by the same force, and the acceptance by the pupil of the numerical identity of this ratio with the inverse ratio of the masses; and both (i) and (ii) require the combination of two proportionality statements into one and then the final setting of a proportionality constant equal to unity. Although we described approach (iii) as though it required the performance of an additional experiment, this is by no means essential; one can proceed straight from the earlier result that $F = Ka$ to identify K with the mass of the object. No combination of two proportionality statements is needed, and no problems arise about setting a proportionality constant equal to unity in the final step. The additivity of mass is not assumed; instead it can be proved by experiment that mass, measured by $\frac{\text{Force}}{\text{acceleration}}$ is indeed additive - that, for example, the mass of two trolleys is twice that of one trolley.

In an overall relative assessment of the merits of the three approaches, we can perhaps agree that approach (ii) can be allowed to drop out of the picture, at least as far as an elementary treatment of the subject is concerned. To "define" a mass-ratio in terms of an acceleration ratio is an operation which is likely to appeal only to "abler, more mathematically inclined pupils"; pupils should certainly <u>know</u> that $\frac{M_A}{M_B} = \frac{|a_B|}{|a_A|}$ when the same force is applied to objects A and B

/(cf.....

(cf. our remarks in connection with (ii) above), but it would be better for them to see this fact deduced from "F = ma". It is between (i) and (iii), then, that a decision has to be made. In favour of (i), it can be said that the assumption concerning the additivity of the masses of similar objects is likely to strike a pupil as obvious, and that therefore the experiment which uses this assumption to show that $a \propto \frac{1}{m}$ for a given force is likely to appeal to him as having some point. The principal disadvantages of (i) are the extra mathematical reasoning it requires and the initial appearance of a proportionality constant k in the relation "F = kma". The latter arises, of course, as a result of the temporary use of "trolley" units for mass; for an entertainingly-written denunciation of the use of a proportionality constant in Newton's Second Law, see Dadourian (1939).
In respect of economy of experimental and mathematical procedure, approach (iii) is an easy winner; its major disadvantage is the intellectual hurdle presented by the identification of the mass with "K" in "F = Ka" - this step might be found a little puzzling by some pupils. Perhaps it is on this account that the planners of the Nuffield syllabus decided in favour of approach (i). It seems to us that approach (iii) ought certainly to be tried out; it might be found more easily assimilable than might at first be supposed. The discussion below will proceed on the assumption that this is the approach being followed in the introduction of mass; it can easily be accommodated to approach (i) if this is desired.

For the sake of completeness, and because the idea comes from one who has established himself as a notable contributor to the advancement of the science of mechanics and to the clarification of its history, we consider briefly an approach to mechanics put forward by Truesdell (1968) as "a practical one today":

/"1. .....

"1. The axioms satisfied by forces are motivated, and to some extent verified by experiments, in statics. These axioms are taken over unchanged in kinetics.

2. The principle

$$\vec{F} = 0$$

as necessary and sufficient for equilibrium [is] motivated, and to some extent verified by experiment, in statics. Following the path opened by James Bernoulli, we assume that motion has the effect of giving rise to apparent forces that are equal, per unit mass, to the negatives of the accelerations in those frames where $\left[\vec{F} = 0\right]$ is the statical equation. Application of this D'Alembert-Euler principle leads to $\left[\vec{F} = \frac{d\vec{p}}{dt}\right]$ as the general equation of [particle] dynamics, including both statics and kinetics."

Truesdell presumably intends this approach to be used at university level, but it is of interest here to see whether it is feasible to implement it at a lower level. Teachers might at first be attracted by the idea that this topic should be approached via statics, on account of the comparative ease with which quantitative experiments in statics could be done by pupils. But difficulties would almost certainly arise when the attempt is made to introduce the relation between force and acceleration by way of D'Alembert's principle. The author's own experience of learning to solve problems in circular motion with the aid of this principle is that he was able to write down the necessary equations while being very confused about what they really meant. The recorded experience of others suggests that he was not unique in this respect. (Cf. our discussion of the teaching of circular motion in Section 4.9.). The idea of introducing "fictitious", acceleration-dependent forces when changing one's frame of reference is a sophisticated one (cf also Section 4.6 in this connection), and it is hard to believe that even a very good teacher of

/physics.....

physics at school could make it convincing when introducing the relation between resultant force and acceleration. One suspects that an attempt to introduce "F = ma" in this way would have the effect of putting this relation into the category in pupils' minds reserved for ideas that have to be swallowed rather than understood - and this would be unfortunate. We therefore think it wiser for teachers to introduce Newton's Second Law by way of dynamics, as recommended above.

4.7.11 Units of mass and force

Once the pupil has seen that force and mass are linked as quantitative concepts by the equation F = ma, he is ready to consider the units in which they can be measured. He already knows units of force of a kind: rubber-band units. In principle, there is nothing wrong with setting up an official standard rubber-band unit of force - let us call it the "Pssc"- in terms of which all forces may be measured; the unit of mass would then be the Pssc-s.m^{-2}, for which the name "puck" might be appropriate. In practice, of course, one proceeds differently. A certain body is arbitrarily chosen to have unit mass, i.e. a mass of 1 kilogram, and the unit of force is defined in terms of this. Once the standard of mass is fixed, other masses may, in principle, be measured in terms of it by comparing the acceleration of the unknown mass with that of the standard when both are acted upon by the same force; from the Second Law it follows that the ratio of the masses is the inverse ratio of the accelerations. The unit of force, called the newton (N), is defined as that force which imparts an acceleration of 1 m s^{-2} to a mass of 1 kilogram. There are those who suspect that this method of defining the newton contains, in this context,

/a.....

a hint of circularity; is one not going round a circle, they say, if one first introduces one's measure of mass as $\frac{\text{force}}{\text{acceleration}}$ and then defines the unit of force in terms of the "product" of a mass unit and an acceleration unit? Not really; it is the confusion between the fundamental measurement procedure of a concept (some would say "the definition of a concept") and the definition of the unit in which it is measured that has given rise to this unfounded suspicion. Neither the unit of force nor that of mass have been officially defined at this stage, so that one is at liberty to define whatever unit one pleases first, after which the definition of the other must follow automatically from "F = ma" . By choosing a standard of mass first one is in effect laying down a conversion factor between the original rubber-band "Pssc" and the more convenient newton.

4.7.12 The inertial balance

In order to drive home the association of mass with inertia, it is desirable that pupils should carry out comparative determinations of mass which do not depend in any way on gravitational effects. As a first step in this direction, they should prove by means of the multiple-trolley experiment that the mass of n trolleys is n times the mass of one trolley. They could then use the principle of the additivity of the masses of identical objects to calibrate a so-called "inertia balance". The prototype of this device seems to have been designed by Schriever (1939); a cheap version of it has been popularized by the P.S.S.C. (cf P.S.S.C Laboratory Guide and Teachers' Resource Book and Guide, Part III) and is now widely used (cf Nuffield Guide IV, Jardine 1966, Brommert 1969, Livesey et al 1967). In a way, the name "inertia [or inertial] balance" is unfortunate, since nothing is "balanced" in this device. Rogers's (1960) name

/for.....

for it, the "wig-wag" machine, is much more apt. The machine consists essentially of a tray or other suitable form of "scale pan", attached to a pair of hacksaw blades so that it can execute sideways oscillations - "wig-wags", in fact. The period of oscillation of a given wig-wag machine is determined by the vibrating mass, so that a calibration curve for the device can be plotted by measuring its period for various known masses in the "scale-pan". In the recommended form of the experiment, the "known masses" are provided by known numbers of similar objects (e.g. two-inch C clamps). After a calibration curve has been plotted the device can be used to find the mass of some "unknown" object in terms of the units of mass employed in the calibration.

A detailed understanding of the dynamics of the wig-wag machine goes well beyond the requirements of a South African school syllabus, but a pupil does not have to have such an understanding in order to appreciate that a comparison of masses is possible with the machine. What could be pointed out to him is that different objects placed on the "pan" of the machine will experience the same force from the machine when the "pan" is displaced the same amount from its equilibrium position. From this he can see in a general way why the period of oscillation of the machine should be less when more massive objects are placed on it. But even should the working of the device remain obscure, the pupil is still able to see experimentally that there is a clear connection between the mass in the pan and the period of oscillation, and to realize that the action of the machine cannot depend on the "heaviness" of the objects used. Reports reaching the writer from various teachers suggest that pupils do not find the wigwag machine bewildering.

/4.8.13.....

4.7.13 Force as a vector

Experimental establishment of the vector nature of force is most easily achieved in the context of statics. However, since force has been introduced dynamically in this treatment, it would be a good thing for pupils to consider the general case of the application of a number of forces to a body. By this time they will already have made use of the ideas that parallel forces add by simple arithmetic (in the establishment of the proportionality of acceleration and force) and that if a pair of anti-parallel forces are applied to a body, the magnitude of the resultant force is given by the difference between the magnitudes of the two forces. From here it ought not to be too difficult to persuade them that, in general, a body would accelerate in the direction of the resultant, obtained by vector addition, of the forces acting on it (the Nuffield Guide IV comments encouragingly that "If we present simple geometrical addition as obvious, pupils will accept it"); and that the magnitude of the acceleration would be equal to the magnitude of the resultant force divided by the mass.

4.7.14 Problems on "F = ma"

One could scarcely have a simpler formula than F = ma, and calculations based on it are bound to be simple also. At first sight one might feel inclined to complain that with so simple a formula the only possible problems involving a single body are "plug-in" ones, i.e. problems which require of the pupil no more than that he should plug given numbers into a formula. In this connection, the following points must be made:-

(i) Certainly calculations based on F = ma are bound to be trivial. But this is of minor importance; the interest in such problems lies in deciding what must be substituted for F

/or.....

or m or a - especially F. Those with experience in teaching elementary mechanics know very well that the principal difficulties experienced by beginners in applying Newton's Second Law are in knowing what the relevant forces are. Therefore in setting problems of this kind special emphasis should be placed on giving pupils the opportunity to think about what forces are acting. Until pupils become properly conversant with Newton's Third Law, they may have some difficulty in sorting out which forces act on which body; however, if they follow the practice of drawing simple diagrams in which they attempt to show the forces acting on the body of interest, and are at the same time encouraged by the teacher to disabuse themselves of such Aristotelian notions as "the force of the body", they may well make progress in this direction.

(ii) It is possible to set searching and worthwhile problems in elementary dynamics that do not require any calculations at all. (See Brommert 1969, Chapter Four, for a number of problems of this type.)

4.8 Momentum as a vector

In this way does the Common Basic Syllabus for Physical Science specify the introduction of (linear) momentum. It is certainly convenient to introduce the concept of linear momentum in the wake of the discussion of Newton's First and Second Laws, although it would also be appropriate to defer this introduction until collisions are discussed. Rearrangement of the equation $F = ma$ in the form $Ft = mv - mu$ shows that if a constant force F is applied for a time t to a body of mass m, the result is a change in the quantity "mass x velocity" of the body equal in magnitude to the product of the magnitude of the force and the time for which it is applied. This implies that the greater the value of "mass x velocity" for a body, the larger the force will have to be in order to

/stop.....

stop it in a given time. Thus the quantity "mv" is seen to have some physical interest, and so it is dignified with a name of its own - (linear) momentum.

The foregoing paragraph outlines an approach to momentum that is an improvement on the bare introduction of momentum by means of its definition. The vector nature of the concept, while implied by its definition as the product of a scalar (mass) and a vector (velocity), is also shown by the relationship to force revealed by the equation "Force x time = change of momentum". Reformulation of Newton's Second Law as "force is equal to rate of change of momentum" is an obvious and desirable step to take in following up the introduction of momentum. It is in this form that the Law most closely resembles Newton's original statement and it is in this form, too, that we have to use the Law in order to retain it in the structure of the Special Theory of Relativity. (In this case, momentum has to be redefined by the expression $p = \frac{M_o v}{\sqrt{1-\frac{v^2}{c^2}}}$, where M_o is the rest mass of the particle in question.)

Although this is not of direct relevance to a school syllabus, it is perhaps worth disposing here of a point which arises in connection with the interpretation of Newton's Second Law in the form $F = \frac{dp}{dt}$, where p is the linear momentum of the system. The attention which this matter has recently received in the correspondence columns of the "American Journal of Physics" shows that there are still aspects of Newton's Second Law which require clarification in the minds of some. Consider a rocket: it is a system of variable mass, since its mass is continually diminishing as a result of the ejection of its exhaust gases (the same is true of conventional motor vehicles, but not to the same extent!). If we apply Newton's Second Law in the form $F = \frac{dp}{dt}$, where p is the

/momentum.....

momentum of the rocket, will we obtain the right answer for the resultant force on the rocket? Actually, we will not; the right answer will only be obtained if the "p" in $\frac{"dp"}{dt}$ is taken to refer to the total momentum of the system of rocket-plus-exhaust-gases, which is not a system of variable mass (cf Trigg 1966a, b, Halliday and Resnick 1966 and Wessner 1967).

4.9 Motion in a curved path as an introduction to circular motion; centripetal force; circular orbital motion

At present these topics are relegated to the status of optional extensions to the Common Basic Syllabus. This is to be regretted, in view of their intrinsic interest and their great practical importance. We shall assume here that any self-respecting syllabus would want to include these topics, and therefore we shall give them prominent treatment here.

If motion in a curved path is approached kinematically, as is implied by the above succession of syllabus headings, then it is easy to argue from the definition of acceleration that even when such motion takes place at constant speed, it must still be accelerated, since the direction of the velocity is changing all the time. However, this may not seem to be a very profound deduction, particularly to the pupil who has not yet been able to see the point of defining acceleration as rate of change of velocity rather than rate of change of speed. So the matter is best approached dynamically, i.e. from the point of view of the forces acting. Somehow the pupil must be convinced that even if a particle is travelling at constant speed along a curved path, a resultant force must act upon it. One may try to tell him that "this is obvious - it follows from Newton's First Law", but it would be best if this point were made by experiment; that the natural motion of, say, a

/dry.....

dry ice puck on a smooth table is straight-line motion, and that a sideways push or pull will deflect its course, or that a stone whirled round at the end of the string is pulled inwards by the string, and would initially move in the direction of the tangent to the circle if this restraint were suddenly removed. (Initially, because its motion would of course also be subject to the force of gravity.) By discussion of examples of this sort, the pupil can become properly aware of the fact that forces can cause changes of direction as well as of magnitude of velocity, and the point can be put to him that in the relation "F = ma", the acceleration "a" need not just be concerned with changes of speed. At this point it might be appropriate to discuss uniform circular motion: the pupil should be persuaded that in this case the acceleration is perpendicular to the velocity, both by means of a geometrical argument featuring the use of velocity vectors and by the argument that no change in the speed of the particle means that there can be no component of acceleration in the instantaneous direction of motion. In other words, a particle experiencing uniform circular motion is always experiencing an acceleration towards the centre of its path; and if Newton's Second Law applies to forces of constantly changing direction, then it follows that such a particle must be subject to a constant resultant force directed towards the centre of its orbit, i.e. to a centripetal force, to use Newton's terminology. Of course, no new category of force is implied by this choice of name and care should be taken to prevent pupils receiving the contrary impression. Perhaps it would be a good idea to omit the term "centripetal" altogether; it is certainly possible to teach the subject without ever using it (cf Hewitt 1963).

/This.....

This brings us to the vexed question of centripetal vs. centrifugal force in elementary teaching of orbital motion. Debates on this subject have flared up from time to time in the correspondence pages of the "American Journal Physics", the "Physics Teacher" and, most recently, in "Physics Education". (For a useful summary of the situation as at April, 1963, see Struthers 1963; for the discussion in "Physics Education" see Belham 1966a, b and MacLeod, Duffin and Constable, all 1966). In the Nuffield Physics Guide (V), an extended and entirely admirable discussion of the whole problem is given in a prefatory "Note to teachers"; in the main text, the author of the Guide virtually falls over backwards in his efforts to be fair to the advocates of the use of "centrifugal force", but opts firmly in favour of "centripetal force". With this choice we are in entire agreement; but before we indicate our reasons for this, let us briefly survey the situation. When we describe circular motion with respect to an inertial frame of reference, as has been done by implication in the above discussion, we are bound to say that a particle moving in a circular path is subject to a resultant force towards the centre of the path, as has been explained. On the other hand, when the behaviour of the particle is described from the point of view of a non-inertial frame of reference rotating with it, a fictitious "centrifugal" force has to be postulated to act on the particle. It is necessary in such a description because, viewed from this accelerated frame of reference, the particle is apparently in equilibrium and to account for this, we have to suppose that a "centrifugal force" equal and opposite to the centripetal force acts on the particle. This is, one might say, the "fundamental" meaning of the term centrifugal force" - it is a fictitious or "phoney" force which we have to invoke when we choose to

/view.....

view circular motion from a frame of reference rotating with the particle instead of an inertial frame. (Cf. Nuffield Guide V: "Centrifugal force is a delusion arising from living in the rotating system and trying to forget it".) It is in this sense, presumably, that journalists or even eminent scientists writing books of a popular character use the term when they make such statements as "For a satellite to remain in orbit around the earth, the earth's gravitational pull must be balanced by the centrifugal force" (cf Massey 1960, 1965). Another sense in which "centrifugal force" is used is as the name for the reaction force exerted by the orbiting particle on whatever agent is exerting the centripetal force; e.g. the force exerted by a stone on the string which is whirling it round. More confusingly, it is somethings thought of in this kind of situation as the force of the string on its pivot (e.g. the hand that is whirling it round). Either of these views of centrifugal force confuses the issue, since it is with the forces acting on the particle that we are concerned when studying its motion. Yet another way of introducing centrifugal force is by way of what Rogers (1960) calls the "engineer's headache cure" (it is somewhat more politely called the "novice's headache cure" in the Nuffield Guide V). This is in the application of D'Alembert's principle to circular motion; that is, one reduces the problem to one of statics instead of dynamics by introduction of a fictitious force (in this case the centrifugal force) to balance out the real resultant force acting on the particle.

Now the first and third ways described above of introducing centrifugal force into a discussion of circular orbital motion are, in their way, quite legitimate, provided that all concerned have a full understanding of what is being done.

But this is just the difficulty: are beginners in the subject

/likely.....

likely to be able to absorb successfully these different ways of viewing the same situation? The author's own experience both as a first-year university student and as a teacher of such students leads him to declare that use of the concept of centrifugal force and of D'Alembert's principle in the teaching of circular motion can be sources of long-lasting confusion, and are therefore best avoided. The efforts of the teacher should be concentrated on trying to make circular motion understandable from the viewpoint of inertial frames of reference and to counter as firmly as he can any malign influences arising from vague notions of centrifugal force that pupils may have acquired from outside reading. It is greatly to be deplored that "centrifugal force" is used so glibly and unthinkingly by many writers who ought certainly to have known better. Incidentally, the slick explanation of the behaviour of an orbiting satellite in terms of centrifugal force only "works" for circular orbits; as no satellite executes such an orbit, it seems more than a little dishonest to produce such an "explanation" as though it were valid.

When a pupil has acquired a thorough grasp of circular motion from the point of view of an inertial observer, he is ready to study it from alternative viewpoints. But this, we suggest, lies outside the scope of any South African school syllabus, present or future.

If circular orbital motion is to be studied <u>quantitatively</u> (and there is not much point in introducing it if it is not), then a decision has to be taken about how to introduce the formula for the centripetal acceleration, viz. $a = \frac{v^2}{r}$. There are two possibilities: either the formula must be derived, or it must be obtained as a deduction from experiment. Obviously the first course is the better to follow, other things being equal. If, as the Nuffield Guide (V) suggests,

/pupils.....

pupils are made to appreciate the need for such a formula by being shown that without it we cannot make calculations of satellite orbits or of the trajectories of charged particles in magnetic fields, then they might be ready to make the effort to understand the derivation of $a = \frac{v^2}{r}$. Perhaps the simplest derivation to use is the one using a vector diagram to show the change of velocity when the radius vector of the particle has traversed a certain angle. (cf Shortley/Williams and 1965). By comparing this vector diagram with the corresponding triangle made by the initial and final radius vectors for the particle and the line joining their ends, and noting that these triangles are similar, it is easy to obtain an expression for the average acceleration in the time interval under consideration. To obtain the instantaneous acceleration, one must proceed to take the limit of this expression as the time interval tends to zero. This is the biggest stumbling block in the whole derivation, and the Nuffield Guide (V) suggests by-passing it by "[avoiding] labouring, or even referring to, the need to proceed to a limit, or the method of doing so." Only the teacher can decide ~~here~~ how best to proceed here with his class; if, at any rate, it is agreed that the derivation of $\frac{v^2}{r}$ should not form examination material, the teacher will know that he need not approach the problem in the spirit of "swallow it or else...".

The experimental approach to " $a = \frac{v^2}{r}$ " depends on showing that the centripetal force is equal to the product of the accelerated mass and a quantity which turns out to be equal to $\frac{v^2}{r}$. Even if the pupil has had "$a = \frac{v^2}{r}$ " derived for him, he should do this experiment as a means of confirming his prediction that the force on a particle moving in a circle is given by $F = \frac{mv^2}{r}$. A number of simple ways of carrying out such an experiment are described in the Nuffield Guide (V).

/We.....

We describe one of these, made familiar by the P.S.S.C. (Laboratory Guide and Teacher's Resource Book and Guide, Part III), as being typical in the cheapness of the apparatus used and the simplicity of the method. A rubber stopper of mass 60 grams or thereabouts is whirled around on the end of a nylon line about 1.5 metres in length which passes through a vertically-held glass tube; the other end of the line is attached to a paper clip which acts as a support for a number of steel washers (preferably weighing about 60 grams each) which, in turn, supply by their weight the tension in the line. The frequency of rotation for a constant radius of orbit and different numbers of washers is measured, and from these results it can be deduced graphically that the square of the frequency (and therefore the square of the speed) is proportional to the tension when the radius is kept constant. Further sets of observations can be taken for different orbital radii; these will enable the dependence of the radius on the tension for constant orbital speed to be deduced.

A stopper of different mass can also be used, so that it can be shown that the centripetal force depends on the mass of the object on which it acts (this should come as little surprise to the pupil!). Taken together, these results should show that the measured force (= tension in the line = weight it supports) is equal to $\frac{mV^2}{r}$.

To give the pupil some immediate reward for his trouble, the teacher might allow him to do a quick calculation of the orbital velocity of a (hypothetical) low-altitude satellite, i.e. one orbiting close to the surface of the earth but somehow not significantly impeded in its motion by air resistance. For such a satellite, the centripetal acceleration must be "g", so by equating $\frac{V^2}{r}$ to "g" and substituting a given value of the earth's radius for r, it is easy to calculate a value for V. The general formula from which the orbital velocity

/of.....

of a satellite in a circular orbit of arbitrary radius can be found, can be derived after Newton's Law of Gravitation has been discussed.

4.10 Newton's Law of Gravitation and falling bodies

Under this heading it is required of the pupil that he should know what Newton's Law of Gravitation is, what it "explains" and how it "explains" it. What it "explains", of course, is that bodies (including satellites, real and artificial, and planets) "fall"; the "falling" is accounted for by the postulated existence of a gravitational force between any pair of particles in the universe. (The use of the word "particles" in the preceding sentence is important; any statement of the Law of Gravitation that begins "Every body attracts every other body in the universe"... is incorrect.) Naturally it was the "falling" of the moon and the planets that interested Newton and his contemporaries most, and it was Newton's demonstration that the motion of these heavenly bodies was accurately described by his Laws of Motion and Gravitation that constituted the crowning achievement of his career.

The method of presentation of these ideas will depend on the amount of time the teacher can devote to them. At one extreme, we have the approach recommended for use in the Nuffield Physics course (cf Nuffield Guide (V)), in which Newton's theory of gravitation is approached via a study of astronomy on the one hand and history on the other. Time and patience permitting, on the part of both teacher and pupil, this is wholly admirable; the story that is outlined here can be made exciting and enjoyable, and what will have been learned by the pupil will surely be of lasting value. But an immense amount of time would be required, for which there certainly would not be room within the confines of the present Common Basic Syllabus.

/If.....

If this "Nuffield" approach is not followed, one might depart from the historical development of the subject by discussing Cavendish's experiment to determine the gravitational constant, immediately after introducing the Law of Gravitation. (A film loop of this experiment is available.) The Law itself can be introduced purely as a postulate, or better by way of some argument that suggests that we need some Law of Gravitation that will tell us that gravity gets weaker as we go further away from the attracting object. (To justify <u>that</u> idea in turn, we can do as recommended in the Nuffield Guide and have pupils calculate the centripetal acceleration of the moon in its orbit: the result readily shows that "gravity is much weaker out there".) The value of the gravitational constant G as obtained from experiments of the Cavendish type can be quoted and used in one or two simple calculations so as to show how very weak gravitational forces between common terrestrial objects are. At this point, it would be appropriate to use Newton's Second Law of Motion and his Law of Gravitation to show that all bodies at the same place on the earth's surface fall with the same constant acceleration "g" in vacuo. The derivation is simple and instructive. Let m_e be the mass of the earth, m the mass of a freely falling object, and R the radius of the earth. Then

$$mg = G\frac{m\,m_e}{R^2}$$

$$\text{so that } g = \frac{G\,m_e}{R^2}$$

The basic equation above expresses the fact that the two alternative expressions for the earth's gravitational pull on the mass m are equivalent, the expression on the left-hand side coming from Newton's Second Law and that on the right-nand side from his Law of Gravitation. The expression for "g" enables the mass of the earth to be calculated if the values of "g", G and R are known. This calculation is well

/worth.....

worth performing for pupils.

It should be stressed at this point or when the Law of Gravitation is first introduced that there was no obvious reason why the gravitational force should have turned out to be proportional to the masses of the interacting particles. That it <u>is</u> so proportional, with the consequence that freely-falling bodies have the same acceleration at the same place, is a remarkable fact which helped to lead Einstein to his General Theory of Relativity, and should certainly not be glossed over.

A return could now be made to the historical development with a brief account of the achievement of Johannes Kepler, who showed that the facts of planetary motion could be summed up in three statements, now known as "Kepler's Laws". The pupil need not be expected to remember Kepler's Laws, but he should gain some idea of their significance. Kepler's discovery of these laws meant that the task of giving a satisfactory explanation of planetary motion was reduced to one of accounting for the existence and form of these Laws. This was the task that faced Newton and which he accomplished with such brilliance. It is impossible in a school course to show how the Law of Gravitation "explains" all of Kepler's Laws of planetary motion, but fortunately one may give a simpler example of the astronomical application of the Law. Virtually all the mathematical equipment is at hand for the calculation of the speed of an artificial earth satellite describing a circular orbit; the only assumption that must be made is that earth and satellite must be treated as point masses located at their actual centres. From the equation

$$m_s \frac{V^2}{R} = G \frac{m_s m_e}{R^2}$$

which yields $$V^2 = \frac{Gm_e}{R}$$

(m_s = mass of satellite, m_e = mass of earth, R = radius of

/earth.....

earth, V = magnitude of orbital velocity of satellite and G = gravitational constant), the value of V may easily be calculated from given data. Since V is also equal to $\frac{2\pi R}{T}$, where T is the period of the orbital motion, it is readily seen that Kepler's Third Law ($\frac{T^2}{R^3}$ = constant) follows at once from this result.

Newton's theory of gravitation is a classic example of an eminently successful scientific theory. The pupil should be made aware of the reasons for the superiority of Newton's theory over its predecessors in this field, viz. the very much greater simplicity and elegance of its formal structure, and its impressive fruitfulness. An appreciation of what constitutes a good theory is fundamental to the understanding of any science, and discussion of Newton's theory of gravitation provides an excellent opportunity for laying the foundations of such an appreciation in the pupil. First-rate accounts of Newton's work viewed from this standpoint can be found in a number of books (cf Rogers 1960, Arons 1965).

4.11 Weight

4.11.1 The concept of weight

Now that the sensation of "weightlessness" has become part of the experience of at least some of the earth's inhabitants, interest has been stimulated in just what we mean when we use the word "weight". In what sense is the spaceman in his capsule "weightless", for example?

The most familiar definition of the weight of a body is undoubtedly "the force which the earth exerts on the body" (cf Brommert 1968). Recently at least one pair of authors has sought to widen this definition by making it read "the resultant gravitational force exerted on the body by all other bodies in the universe" (Sears and Zemansky 1964); this makes it possible to talk about the weight of an object on different planets without having to re-define the quantity each time a

/transfer.....

transfer is made from one planet to another. For purposes of calculation, the formula $W = mg$ is invariably used, where W is the weight of an object of mass m and g is commonly called "the acceleration due to gravity". It is important to realize that weight in this last sense, which is the quantity measured by a spring balance, is not equal to the quantity referred to in the first two definitions quoted. The quantity "g" is not, in fact, equal to the gravitational force on a body divided by its mass. Instead, because it is unavoidably measured in an accelerated frame of reference (i.e. that fixed with respect to the earth's surface) "g" is the apparent acceleration relative to this frame of reference. In other words, an object hanging from a spring balance is not really in equilibrium; it is accelerating towards the earth's axis of rotation with acceleration $\frac{v^2}{R}$ where v is the magnitude of the velocity of rotation of the spring balance round the axis and R is its perpendicular distance from the axis. Therefore "g" is the vector difference between the true "acceleration due to gravity" and the acceleration a_R towards the earth's axis. It is the variation in magnitude of a_R with latitude that is mainly responsible for the observed variation of "g" over the surface of the earth; variations arising from the fact that the earth is not spherical, from differences in altitude and from local variations in density play a comparatively minor role (cf Sears and Zemansky 1964).

We have not yet finished our list of definitions of weight. King (1962), seeking an operational definition of weight, proposed to define it as "the force exerted by the body on its support", and thereby initiated a friendly debate between himself and Sears (cf Sears 1963a, 1968, King 1963, 1967) which shows every sign of continuing. Sears prefers

/to.....

to reserve the term "weight" for a force exerted on the body. By implication, he defines "apparent weight" of a body as "The forces [sic] due to the phenomenon of gravity acting on a body as observed with respect to a specified system in which the body is at rest". The wording is actually that of Iona (1963a), for whom this is a definition of "weight".

How does all this affect the treatment of "weight" in terms of the present syllabus? There is no doubt that in solving most problems concerned with terrestrial objects one has to deal with "weight" as given by W = mg (where "g" has its usual significance). The important point, then, is to establish this formula and to show how and when it may be used. It appears harmless enough to start with the first definition of weight given above, to claim for the moment that weight is measured by a spring balance, and to derive the relation W = mg from Newton's Second Law and the fact that all bodies at the same place on the earth's surface fall in free space with the same uniform acceleration "g". One can then point out that "g" is not quite what it has been assumed to be, and indicate briefly why this is so. The question of what would be meant by the weight of an object on the moon or on some other planet than the earth should be raised with pupils; they should not experience any difficulty in modifying or widening their original definition to take this new situation into account.

It will be impossible and indeed undesirable for the teacher to avoid discussing the now-familiar phenomenon of "weightlessness" in the classroom. It should be made clear at the outset that in this connection, "weight" means "the reading on a bathroom scale (or a spring balance)", i.e. it

/is.....

is King's "weight" or Sears's "apparent weight" that is under discussion. Then the essential point about "weightlessness" could be explained with reference to the analogy of the man standing on a bathroom scale in a freely-falling lift. If pupils can appreciate that a satellite and its occupants are likewise freely falling, they should not then have any trouble in understanding "weightlessness". When it is not obscured by references to "centrifugal force" or by the prevalent superstition that "gravity is so weak up there", the issue is seen to be simple and clear.

The above approach to "weightlessness" would not recommend itself to Helsdon (1967a,b), a mechanical engineer who is dissatisfied with the present "definition" of force as "F = ma" (where a is the acceleration relative to "fixed space") which, he claims, is not compatible with the conventional view that a body at rest in a gravitational field is acted on by <u>two</u> equal and opposite forces. He proposes the following new definition of force: "The force acting on a body is equal to its mass multiplied by its acceleration relative to free-fall, where free-fall is its motion if it were not in contact with other bodies (liquid, solid or gaseous)". It follows at once that a freely-falling body such as a satellite experiences no force and, as he says, "the force of gravity disappears." One might think that this is simply an attempt to introduce the general-relativistic definition of an inertial frame into elementary mechanics, but his description of an iron bar accelerating in a solenoid as being in "free fall" belies this. (He nowhere mentions general relativity in his articles.) Among the advantages claimed by Helsdon for his proposed approach is that it entails a much simpler description of the motion of a freely-falling body than "the conventional one in which it is assumed that the body throughout its motion is acted upon by two equal, opposite and mutually

/cancelling.....

cancelling forces, due to gravity and inertia." But who but the most muddle-headed and perverse of physics teachers would want to describe free fall to an elementary physics class as being the result of opposing forces due to gravity and inertia? Helsdon's attempt at re-definition of force seems to spring from too narrow an operational view of force. Perhaps a consistent teaching programme could be based on it, as he attempts to show, but since the conventional approach is better in accord with our intuitive ideas about force and does not give rise to serious conceptual difficulties, it seems best to stick to it. The flurry of adverse comment provoked by Helsdon's original communication (Parry, Colbeck, Armstrong, all 1967), appears to indicate that thoughtful teachers feel the same way about the matter.

4.1.1.2 Weight and gravitational and inertial mass

A matter requiring more serious consideration is the question of whether the concept of "gravitational mass" should be introduced in a school syllabus and the proportionality of weight and mass be further described as "the proportionality of gravitational and inertial mass". In at least one high-school physics course (PSSC) the topic is dealt with in this way; and when, some time ago, the present writer raised the matter with an "elite" group of private school teachers, they were overwhelmingly in favour of this approach.

The introduction of the concept of "gravitational mass" (or, if this term is thought to be question-begging, "gravitational charge" (Arons 1965) or "gravitance" (Schriever 1939)) is usually achieved in the following way. In describing the phenomenon of gravitation quantitatively, we can say that any particle has a property, to which we can give one or other of the names just mentioned, which determines the extent of its gravitational attraction <u>for</u> or <u>by</u> other particles.

/(Strictly.....

(Strictly speaking, we should distinguish at first between the concepts of active gravitational mass - the gravitation-creating or field-creating mass - and passive gravitational mass - the gravitation-responding or field-responding mass; cf. Bondi 1957. But it is easy to show from Newton's Third Law and his Law of Gravitation that these two kinds of gravitational mass must be equal.) We do this by analogy with what is done in describing electrostatic forces, where the phenomenon of electrostatic repulsion or attraction is discussed in terms of the possession of charges by the particles concerned. Next, we consider how gravitational mass is to be measured. Again we can proceed by analogy with what is done in electrostatics. First we can say that two particles have equal gravitational masses if, when placed at the same distance from a third particle, they experience the same gravitational attraction. Further, we can say that it follows from the Law of Gravitation that if two particles placed alternatively at the same distance from a third particle experience unequal gravitational forces, the ratio of their gravitational masses is equal to the ratio of the gravitational forces. The first of these propositions, which is obviously a special case of the second, is the basis of the comparison of gravitational masses by means of the equal arm balance; the second, which we can write in the form $\frac{M_{G_A}}{M_{G_B}} = \frac{W_A}{W_B}$, is often used (e.g. by Rogers 1960) as a definition of the ratio of gravitational masses, though no doubt here we ought to be careful not to confuse an equality with a logical identity (cf Section 4.7.4). For an actual measurement of the gravitational mass of an object, we need to compare it with some standard object ; if that standard object is chosen to be the standard kilogram, then by the principles outlined above we can determine the gravitational mass of any other object in kilograms,

/if.....

if we agree that the units of gravitational mass are to be kilograms, like those of (inertial) mass.

The above-described procedure for the measurement of gravitational mass would appear to be acceptable to all who have treated the subject, with one exception: Burniston Brown (1960a,b; see also Lustig 1960). According to Brown, once we have made our choice of units of force, mass and distance, the choice of a unit for gravitational mass is not an arbitrary one. In particular, we are not at liberty to choose the standard kilogram for this unit. Brown's method of defining the unit of gravitational mass is to follow a similar procedure to that used for defining the electrostatic unit of charge: he writes the equation

$$F_{\text{gravitational}} = \frac{m_{GA}\, m_{GB}}{r^2},$$

where m_{GA}, m_{GB} are respectively the gravitational masses of particles A and B, and regards this as <u>defining</u> the unit of gravitational mass.

It is interesting now to compare the treatment of Brown and his opponents of the proportionality of gravitational and inertial mass. On the conventional approach, one reasons thus: Newton's experiments with pendulums, and sophisticated modern repetitions of this experiment, show that the weight of an object may be expressed in the form $W = mg$, where g is the same for all freely-falling bodies in the same part of the earth's surface, and m is the (inertial) mass of the object. Hence for a pair of objects A and B we may write $\frac{W_A}{W_B} = \frac{m_A}{m_B}$. But we already have that $\frac{W_A}{W_B} = \frac{m_{GA}}{m_{GB}}$; therefore $\frac{m_A}{m_B} = \frac{m_{GA}}{m_{GB}}$, or, rearranging this equation,

$$\frac{m_A}{m_{GA}} = \frac{m_B}{m_{GB}}$$

This last statement makes it obvious that gravitational and inertial masses are proportional to each other. It follows at once that if the same units are used for both, inertial and

/gravitational.....

gravitational masses are equivalent. Of course, the basic fact is the proportionality: the "equivalence" arises from our apparent freedom of choice of units for the two kinds of mass.

Brown, on the other hand, argues as follows:- Experiment shows that the gravitational force between two particles is proportional to the product of their inertial masses, when they are separated by a given amount: i.e. we have experimentally that

$$F_{\text{gravitational}} = G \frac{m_A m_B}{r^2},$$

where G is the gravitational constant (= 6.67×10^{-11} N-m^2/kg^2, if SI units are used).

If we compare this with the equation which defines gravitational mass, viz.

$$F_{\text{gravitational}} = \frac{m_{GA} m_{GB}}{r^2}.$$

we see that it follows that $m_{GA} = G^{\frac{1}{2}} m_A$,

so that the inertial mass is greater than the gravitational mass by a factor of about 1.2×10^5. (Brown, incidentally, works this sum out in CGS units, and so obtains a ratio of "about four thousand".) Therefore, says Brown, inertial and gravitational mass cannot be equal. It is difficult to follow this reasoning. Brown seems to feel that gravitational mass must be defined by setting the proportionality constant in his defining equation equal to unity, but he is nowhere able to account for the special magic that apparently attaches to this choice of constant. In short, the case he makes out for his viewpoint is unconvincing.

There is a school of thought which maintains that there is no real need to express the proportionality between weight and mass in the form of the statement that "gravitational and inertial masses are proportional" (Gledhill 1965, Bunge 1967). Gledhill's argument may be summarized by saying that as Newton

/showed.....

showed long ago that weight was proportional to inertial mass, there is no need to complicate the present-day exposition (and the teaching) of the subject by first introducing the additional concept of gravitational mass and then showing that it is, after all, proportional (or equal) to inertial mass. One kind of mass was good enough for Newton: why should it not be good enough for us? Bunge's argument is similar: he puts the essence of his arcitude as follows:- "This statement [that the inertial and gravitational masses are equal] is misleading insofar as it suggests that there are two different masses, only they happen to be numerically equal rather than identical. Such an interpretation is bolstered by the operationalist demand to 'define' every physical concept by means of a set of laboratory operations." This last remark does not seem entirely fair. The concept of electric charge does not arise from any "operationalist demand to define every physical concept by means of a set of laboratory operations", but because of the need to give a quantitative account of a certain physical phenomenon. One can argue that the concept of "gravitational mass" or "gravitational charge" was invented in order to fill a similar need. However, it is certainly true that "experiment has never detected any difference between these hypothetically different masses" (Bunge), so that on the grounds of economy of concepts, one might well dispense with "gravitational mass". Of course, neither of these authorities dispute the fundamental significance of the proportionality between weight and mass in helping to lay the foundations of the General Theory of Relativity. Dingle (1967a,b) would not appear to share the general conviction regarding the importance of this fact: he speaks of Newton's experimental discovery of it as being "of exactly the same character as the discovery that the

/velocity.....

velocity of sound in a gas depends on the ratio of the principal specific heats of the gas", and regards Einstein's theory of gravitation, so far as it concerns this point, as being an alternative but not a superior formulation to Newton's. Since the dependence of the velocity of sound in a gas on the ratio of the principal specific heats can be deduced from the appropriate theory, whereas the proportionality of weight and mass cannot be deduced from the Newtonian theory of gravitation, Dingle's comparison does not seem to be valid. His other point, more fully stated, is that Einstein's theory only appears preferable to Newton's in elegance if we insist on translating Newton's theory into field form. Otherwise, apart from the question of which agrees better with experiment, the one is as good as the other. This is a question which is perhaps not appropriately debated here; but mention of it does serve to shed additional light on the complexities that lie behind the experimental fact that has been under discussion for these last few pages. Of the implications of all this for the teaching of the subject, the following can safely be said:-

(i) It would be best to omit all mention of "gravitational mass" in a high-school science syllabus; first, on account of the controversial nature of the concept itself and second, because the concept <u>can</u> be dispensed with in teaching "W = mg" and multiplication of mass concepts could cause confusion.

(ii) The apparently accidental nature of the constancy of "g" for all freely-falling bodies at a given place should be stressed.

(iii) Point (ii) and "W = mg" must be used to justify to pupils that the equal-arm balance can legitimately be used to compare masses.

4.12 Newton's Third Law and Conservation of Momentum

Although the Third Law will have been discussed to some extent in the junior years, it is important to revise it now

/and.....

and perhaps also to apply it to more complex situations than those considered earlier. It would be instructive, for example, for pupils to make a careful study of the forces acting when a drawing pin is being pushed into a board: what forces act on the person's thumb, on the pin and on the board, and how are these related when the pin is (a) at rest (b) accelerating and (c) moving with uniform velocity? Then there are the familiar "difficult" cases of the impossibility of lifting oneself up by one's own "bootstraps" and the horse-and-cart problem. Problems of the latter kind should not be left out of classroom discussion on the grounds that they are too "difficult" for pupils to understand. One of the advantages that Newton's Third Law confers is that it enables one to treat the motion of a system without regard to the forces that different parts of the system exert upon one another. It is important for a pupil to realize this, and to recognize that problems such as that of the motion of the horse and cart can be understood both from this point of view and in terms of the forces acting on the horse and the cart individually. The forces exerted on each other by horse and cart have no effect on the motion of the system, considered as a whole, since they are equal and opposite forces acting on the same system. The horse-and-cart system will accelerate forward if the forward push exerted by the ground is stronger than any forces resisting the motion such as friction. In terms of the forces on the individual constituents of the system, the horse will accelerate forwards if the forward push on it from the ground is more than sufficient to overcome the combined effect of the backward pull of the cart and the frictional forces. The cart will accelerate forward if the pull exerted on it by the horse is more than sufficient to overcome frictional forces. (This and other "action and reaction" problems are discussed in detail and

/with.....

with great clarity by Rogers (1960).) In the last resort, it comes to this: that if the teacher himself has properly understood the Third Law and can communicate his sense of confidence to his pupils, none of these problems need seem difficult at all.

Pupils should see the Principle of Conservation of Momentum derived from Newton's Third Law for the special case of a "straight-line" collision between two bodies. The derivation is entirely straightforward and needs no comment here. It would help pupils to appreciate the vector nature of momentum if a two-dimensional collision could be demonstrated and analysed in class. Multiflash or stroboscopic photographs of billiard-ball or dry -ice puck collisions would serve excellently for this purpose (cf PSSC 1960, 1965, Jardine 1966, Ritchie 1966).

The pupil should do simple experiments on "straight-line" collisions or "explosions" as <u>illustrations</u> of the Principle of Conservation of Momentum. For example, it is easy to carry out an experiment in which the ratio of the magnitudes of the velocities of two P.S.S.C. trolleys is measured after the trolleys have been "exploded" apart by a spring (cf. P.S.S.C. Laboratory Guide, Shuter 1965, Experiments 7.1 and 7.3 in Jardine 1966, and the Nuffield Guide (IV)). Also one can drop objects on to a moving trolley to demonstrate that momentum is also conserved when the initial momentum of the system is not zero. In some quarters, doubt has been cast upon the logical status of experiments of this type; Armstrong (1966b), for example, suggests that such experiments are superfluous since the Principle of Conservation of Momentum is a logical deduction from the Third Law and is in no need of experimental confirmation. Now it is quite true that experiments of the kind mentioned above do not "prove" the

/Principle.....

Principle of Conservation of Momentum: they merely illustrate it. If, for example, the relative masses of the colliding or exploding trolleys had in the first place been determined by a "reaction" experiment of the type described in Section 4.7.3, there is no doubt that subsequent use of those trolleys in an experiment to "prove" conservation of momentum would amount to a "Zanzibar" experiment. Suppose instead, that we perform our experiment with vehicles of different masses, using several piled-up trolleys as one vehicle and a single trolley as the other. Then our assumption that the ratio of the masses of the vehicles is equal to the ratio of the number of trolleys can only be checked by an experiment in which the same force is applied to the two vehicles in turn, or in which the vehicles are weighed on a suitable balance. But in either case the Third Law enters implicitly or explicitly into the interpretation of the results of the experiment, so that we are finally in the position that the interpretation of the results of any experiment we use to prove Conservation of Momentum depends on the Third Law, from which the Principle is deduced. And the Third Law, as we have remarked earlier, (Section 3.8.4) does not by its nature seem capable of direct proof.

Nevertheless, as we have already asserted, pupils should carry out experiments which <u>illustrate</u> conservation of momentum; they should also carry out experiments in which the Principle is <u>applied</u> in order to deduce some result. In the experiments on "collisions" or "explosions", they can work out the total initial and the total final momentum of the system and show that the results are consistent with conservation of momentum within the limits of experimental error. This is really a test of their experimental ability, and they should realize this; but the important point is that they have been

/conducting.....

conducting observations on a system to which conservation of momentum can be said to apply. This kind of experience will aid them in their recognition of situations in which the Principle can profitably be applied, and in view of the fundamental importance of the Principle, the more such experience they gain, the better.

Practical examples of the application of the Principle are legion. The most spectacular is rocket propulsion, the general principle of operation of which ought certainly to be made known to the pupil. Naturally any suggestion of mathematical complication should be excluded from the discussion of this example; there are various simple analogies (e.g. a machine-gun being fired from a trolley) that can be used to make the essential points clear (cf. P.S.S.C. 1960).

Any treatment of conservation of momentum would be incomplete if it did not stress the universal validity of the Principle. No violations of it have ever been found. We run into difficulty when we seek to apply Newton's Third Law to certain cases in which the interaction between particles is patently not instantaneous, in contrast to what is tacitly assumed about the forces to which the Third Law refers; but even in these cases momentum is found to be conserved. The Principle has been successfully applied in the analysis of all manner of reactions in the realm of nuclear and elementary particle physics. Although experiment and theory both tell us to expect deviations from "classical" behaviour when we study particles moving at speeds close to that of light, these particles still act in accordance with momentum conservation, provided we define momentum as $\frac{M_0 V}{\sqrt{1-\frac{v^2}{c^2}}}$.

The interactions between a pair of "current elements" or between a pair of isolated moving charges provide celebrated examples of apparent violation of the Principle of Conser-

/-vation.....

vation of Momentum. The supposed violation in the case of current elements need not be a cause for concern, since isolated "current elements" can have no real existence; they are mathematical fictions used to facilitate calculation of the magnetic field set up at a given point by a given current-carrying circuit, or of the forces exerted upon each other by current-carrying circuits. The latter forces do conform to Newton's Third Law and the Principle of Conservation of Momentum, as one would expect. However it is possible to "rescue" Conservation of Momentum in the case of isolated current elements if one takes into account the momentum associated with the electromagnetic fields of the current elements, and the interaction between a pair of moving charges can be similarly dealt with and shown to satisfy Conservation of Momentum if the momentum of the field is taken into account (cf Page and Adams 1945). The forces exerted on each other cause the charges to accelerate and to radiate away the momentum that is unaccounted for in a naive approach to the problem.

4.13 Can one "derive" the Principles of Conservation of Linear and Angular Momentum?

We have recommended above (Section 4.12) that the pupil should be introduced to conservation of linear momentum by way of Newton's Third Law, the latter being treated for the moment as more fundamental. The teacher can reveal later that even in situations in which the Third Law does not seem to hold, the Principle of Conservation of Linear Momentum retains its validity. It is of interest to enquire whether or not this Principle can be derived in any other way from other physical principles, besides the naive "derivation" from the Third Law mentioned above. A related question is whether what Truesdell (1967, 1968) calls the "Principle of Moment of Momentum" (the rotational analogue of Newton's Second Law) and the associated "Principle of Conservation of Angular

/Momentum".....

Momentum" are to be treated as independent or derived laws: both are sometimes presented as consequences of Newton's Laws of Motion.

The Principle of Conservation of Linear Momentum can easily be proved for a macroscopic system from the assumption that the vector sum of all the internal forces acting on the system must be zero. This assumption can be justified on the further assumption that all internal forces can be grouped into pairs to which Newton's Third Law applies, and that the lines of action of any pair of forces must coincide. As we have seen above, this kind of assumption is made in reference to the "macroscopic" forces one encounters in problems like that of the horse pulling the cart and leads to results that are consistent with experience. To extend it to forces of interaction between the microscopic "particles" composing the system is perhaps to go further than is either necessary or desirable. As Gerjuoy (1949) remarks, "... the force acting on any one of three contiguous atoms depends on the simultaneous positions, orientations and polarizations of the other two, and can only with extreme artificiality be regarded as resulting from pairs of equal and opposite forces between the various atom pairs"; and we have already seen that Newton's Third Law cannot be straightforwardly applied to the interaction of a pair of moving charges. Gerjuoy (op. cit.) has shown that the vector sum of the internal forces must vanish provided that these forces do no net work during a translation of the system through space. In other words, the internal energy of the system must not change during such a translation, if no energy is supplied from external sources: this is consistent with experience. In the same paper, Gerjuoy has also shown that if the internal forces of material bodies do no work in a <u>rotation</u> of the body in space, the

/vector.....

vector sum of the torques associated with the internal forces must vanish and the Principle of Conservation of Angular Momentum follows. (So, too, does the Principle of Moment of Momentum, although Gerjuoy does not say this explicitly).

On this approach, Conservation of Linear Momentum and Conservation of Angular Momentum are independent principles, each being derived from another, more fundamental principle, viz. in the one case the "indifference" of the internal energy of the system to translation through space, and in the other case the indifference of the internal energy of the system to rotation. More recently Noll (cited by Bunge 1967 and Truesdell 1968) has produced an argument in this connection which appears to be much the same as that of Gerjuoy. As a "basic principle of mechanics" he "[lays] down the demand that the mechanical power be frame-indifferent; i.e. invariant under time-dependent rigid transformation of Euclidean (or a more general 'neo-classical') space-time. Any such transformation may be decomposed into a translation and a rotation. Considering translations yields the principle of linear momentum (i.e. Newton's Second Law); considering rotations yields the principle of moment of momentum" (Truesdell 1968). Unfortunately we have not had access to Noll's full treatment of this topic, so that it has not been possibly properly to compare his approach with Gerjuoy's. In any event, he and Gerjuoy do concur in establishing the Principle of Moment of Momentum (and with it the Principle of Conservation of Angular Momentum) as an independent principle. The first person who seems clearly to have recognized its status as such in classical particle and continuum mechanics was Euler. (Cf. Truesdell (1968): "The general principle of moment of momentum, as independent of the principle of linear momentum and as applicable to every part of every body, was first proposed by Euler in 1775. He was led to it through studies of elastic

/lines,.....

lines, culminating in 1771 in his general theory of the plane linear continuum".)

It is possible to derive the above-discussed conservation laws from considerations even more abstract and fundamental than those mentioned above. Noether (cited by Kaempffer (1967)) showed some time ago that the symmetry properties of physical systems were associated with conservation laws, in particular that symmetry with respect to "translation in time" implies that energy must be conserved, symmetry with respect to translation in space implies conservation of linear momentum and symmetry with respect to rotation in space implies conservation of angular momentum. (These results are perhaps most easily derived in the context of quantum mechanics: cf. Feynman et al 1965.) But none of these approaches achieves a genuine reduction in the number of physical principles employed: we need just as many symmetry or invariance principles as the conservation laws we want to derive so that, as Kaempffer (1967) remarks, we might just as well have postulated the latter in the first place. There seems to be no way of avoiding this situation. There are other possible ways of deriving these laws - for example, one can derive the Principle of Moment of Momentum from another principle enunciated by Noll, viz. the "Principle of material frame-indifference" ("The properties of materials - 'actions of bodies on each other' - are the same for all observers" - cf Truesdell 1968) - but in no case do we seem to be able to derive more than one of them from the same basic assumption (it must be remembered here that Noll's "frame-indifference of mechanical power" principle is really two-pronged.).

On the whole, these facts are of little direct use to the teacher, except inasmuch as they assure him that he cannot progress further at elementary level in elucidating conservation

/of.....

of linear momentum than we have previously indicated, and that they make clear that the "Principle of Moments" (which will be the limit of pupils' acquaintance with the Principle of Moment of Momentum)is in fact an independent principle.

4.14 Dynamics via momentum

Ogborn (1965) has suggested that it might be pedagogically advantageous to approach dynamics through simple collision experiments instead of through the relationship between force and acceleration. He describes a series of experiments in which the collision between two similar trolleys is first used to show that the changes of velocity of the two trolleys are equal and opposite, and a further series in which the "explosive interactions of bodies with varying mass-ratios" (e.g. a simple trolley and a set of piled-up trolleys) are used to show that "the changes of velocity are in the ratios of the 'sizes' of the bodies, are opposite, and the larger body suffers the smaller change." To interpret the results of the second set of experiments the concept of mass is introduced:" 'Size' or 'heaviness' should begin to seem to be accidental accompaniments of the essential property that decides how big changes of speed will be, and we may introduce the name 'mass' for this property." The law embodied in these phenomena is then formulated in the terms "The changes of (mass x velocity) are equal and opposite for interacting bodies". At this point one presumes that the concept of momentum is explicitly introduced, although Ogborn does not actually say so.

The relation between force and acceleration is introduced by means of the following argument:

In interactions, explosive or collisional, pushes or pulls acting for some time produce changes of velocity; therefore perhaps "a steady force would produce a continual change...

/or.....

or a constant acceleration". The familiar type of experiment in which a constant force is applied to the trolley by means of an elastic thread kept at constant stretch, is used to confirm this hypothesis. Then it is suggested that since the interaction experiments show that the forces acting in these interactions produce equal and opposite changes of momentum, it is reasonable to suppose that equal and opposite forces act throughout on the interacting bodies; and, "since a constant force produces a constant rate of change, we may put the result in the form: 'Force equals rate of change of momentum, or mass times acceleration'". This last statement is described as a "force definition;" pupils carry out a "final check on the adequacy of this definition" in which trolleys are pulled with several identical and similarly stretched threads, and in which various numbers of stacked trolleys are pulled with the same thread.

One of the points made by Ogborn in favour of this "momentum" approach is that the introductory experiments depend on the measurement of velocity or speed, and not on the much more sophisticated concept of acceleration. Also "the outcome [of these experiments] is not easily foreseeable from the preliminary ground work;" Ogborn feels that in the usual sequence of experiments concerning force and acceleration, "there is a danger that the end result will be so anticipated that the element of discovery when it is reached will be minimal". There is certainly something in this, although we are inclined to think that Ogborn's argument may apply more appropriately to the needs of O-level pupils, with which it is really concerned, than to South African pupils, who would reach a formal study of dynamics possibly at a later stage than their British counterparts. The sophistication of the concept of acceleration should be less bothersome to sixteen- or seventeen-year olds than it might be to younger pupils.

/If.....

If Ogborn's "momentum approach" were to be adopted, however, certain details of its accompanying arguments should perhaps be modified. First, it is not essential that the introduction of mass should be done by working with stacked trolleys and talking at first in terms of the "sizes" of the interacting bodies. Instead, the concept of inertia can be introduced at this point, and it can be suggested that the measure of inertia, or mass, can reasonably be taken to be inversely proportional to the velocity acquired in the "explosion" by the body concerned (cf. the treatment of mass recommended above, Section 4.7.10). Second, it is not necessary to view dynamics in quasi-Machian terms, as Ogborn does, and accordingly assert that force is <u>defined</u> as rate of change of momentum. One could simply say, following the argument used by Ogborn, that these experiments suggest that the pushes and pulls exerted in explosions or collisions are proportional to the rates of change of momentum which they bring about. Then one could proceed to carry out the experiments concerned with force and acceleration, not to "check on the adequacy" of one's alleged definition of force, but to see whether the relationship between force and acceleration suggested by the momentum experiments receives additional experimental support.

Ogborn is at pains to stress the "interaction" aspects of his approach. He rightly draws attention to the interest of modern physicists in the results of different types of interaction, and represents his initial two-trolley experiments as the simplest way in which the interaction between the two bodies can be investigated. The teacher can certainly take a hint from this: whichever way he chooses to introduce force, he should point out that the concept of force gives one a way of characterizing the interaction between a particle and its environment (cf Section 4.7.4). This can pave the

/way.....

way for other, more sophisticated ways of representing this interaction (e.g. potential energy) that the pupil may meet later on. At the elementary level, force remains the simplest way of characterizing this interaction and the simplest to work with. It can still be very useful in certain situations concerned with elementary particle physics: in order to account for the orbit of a charged particle in a magnetic field, it remains advantageous to describe what happens in terms of the Lorentz force acting on the particle.

We do not think that Ogborn's approach necessarily brings the idea of "interaction" into prominence better than do the more familiar approaches to dynamics. Certainly Ogborn's interaction between two trolleys is no simpler in principle than the interaction between a rubber band and a trolley. Thus on this point Ogborn's approach has no special advantage.

Ogborn's case for the pedagogical simplicity of his approach does not seem entirely convincing. Introduction of the concept of mass in the context of "collisions and explosions" prior to the study of "force versus acceleration" means that in the study of force versus acceleration, one would have to use the comparative measures of the masses of trolleys as established in the collision experiments, and then show that the product of mass times acceleration was proportional to the applied force. Contrast this with the approach to mass listed as number (iii) in Section 4.7.10, where the value of the mass is simply given by the proportionality constant which relates force to acceleration. Thus simplicity in the introduction of mass and the quantitative study of interactions is achieved at the expense of some loss of directness in the treatment of Newton's Second Law.
One does not wish to be dogmatic about which approach should ultimately be regarded as superior; it would be useful to have reports on the success of classroom trials of Ogborn's approach.

/4.15....

4.15 Work, energy, conservation of energy and power

4.15.1 Background

Before we discuss work and energy in the context of the present syllabus, we shall give some account of these concepts as they are understood by present-day physicists. We do this for a number of reasons:-

(i) Recently-introduced school syllabuses in physics lay great stress on energy and its conservation.

(ii) Some teachers, bewildered perhaps by the glib and scrappy treatment accorded to energy in some of our older textbooks, have requested enlightenment on this issue.

(iii) In some quarters, the prominent role given to energy in the new syllabuses has given rise to concern (cf our remarks in Section 3.9.1 and the references given there); the critics of this approach to elementary physics feel that energy is being made to appear what it is not, viz. something perilously like the "caloric" of the earlier theories of heat.

The first and perhaps startling fact to appreciate about energy is that in classical mechanics one could dispense with this concept altogether; all mechanical problems could, in principle, be solved without reference to energy. Why, then, is so much fuss made about it? To answer this question we must first examine how energy makes its appearance in mechanics. For simplicity, we deal only with non-relativistic situations for the moment.

We start with the definition of work. A constant force $\vec{F}$ whose point of application is displaced an amount $\vec{\Delta s}$ is said to do an amount of work equal to $\vec{F}.\vec{\Delta s}$ or $F\,\Delta s\cos\theta$, where F and Δs are the magnitudes of $\vec{F}$ and $\vec{\Delta s}$ respectively and θ is the angle between the vectors $\vec{F}$ and $\vec{\Delta s}$. From this definition it is easy to show that the work done by the resultant force on a particle of mass m in accelerating it from rest to a speed

/v.....

v is equal to $\frac{1}{2}mv^2$, regardless of whether the accelerating force is constant or not.

There are various ways of interpreting this result. In earlier decades it would have been fashionable to approach the matter as follows:- The particle could do an amount of work equal to $\frac{1}{2}mv^2$ in being brought to rest; it has a capacity to do this amount of work as a result of its motion. If we call the capacity to do work energy, then we may say that the body possesses energy by virtue of its motion, and give the appropriate name kinetic energy to this kind of energy. Nowadays, if the practice of innumerable modern textbooks (e.g. Sears and Zemansky 1964, Arons 1965, Kittel et al 1965, Halliday and Resnick 1966, to mention only a few) may be taken as typical, one simply defines the quantity "$\frac{1}{2}mv^2$" to be the kinetic energy of a body of mass m and having a velocity of magnitude v. In any event, we now have the derived result that the work done on a particle by a resultant force is equal to the increase in kinetic energy of the particle. (We have supposed the particle to have been accelerated from rest, so that its kinetic energy increases from zero to $\frac{1}{2}mv^2$.) This result is known as the "work-energy" or "work-kinetic-energy" theorem for a particle. It is important to note that the derivation of the familiar expression "$\frac{1}{2}mv^2$," which we have called the kinetic energy of the particle, is based directly upon Newton's Second Law, and discloses nothing essentially new about the dynamics of the particle.

It is because energy turns out to be a conserved quantity that it has attained its present status in physics. Historically, this aspect of energy first revealed itself in studies of perfectly elastic collisions. Huygens had used the principle that kinetic energy (or vis viva, as it was then called) is conserved in collisions between "hard bodies" to prove

/that.....

that the magnitude of the relative velocity of two hard bodies before impact is the same as that after impact (cf Theobald 1966), and of course the theorem can be worked the other way round, by proving from conservation of momentum and the assumption that the relative velocity of the colliding bodies is merely reversed by the collision, that Σmv^2 (or $\Sigma\frac{1}{2}mv^2$) must be conserved in this type of collision. Now, as Theobald (op. cit) correctly remarks, the impact of elastic masses can be dealt with in "Newtonian terms", i.e. without reference to conservation of kinetic energy, provided certain assumptions are made about the nature of elasticity. Unfortunately his own demonstration of this is unnecessarily clumsy, if not confused. His "assumption about elasticity" is that "the compression and subsequent recovery of the [colliding] bodies are identical processes though in opposite senses." This "assumption about elasticity" amounts to saying that if one of the colliding bodies suffers a diminution of velocity ΔV during the first stage of the collision in which the bodies momentarily attain a common velocity, it will suffer a further diminution of velocity of the same amount during the second stage in which the bodies separate. (For the other body, substitute the word "increase" for "diminution"). From this it follows easily that the relative velocity of the colliding bodies is reversed by the collision. If ΔV_1 is the initial diminution in velocity suffered by body (1), originally travelling with velocity V_1 and ΔV_2 is the initial increase in velocity of body (2), originally travelling with velocity V_2 , then

$$V_1 - \Delta V_1 = V_2 + \Delta V_2 \quad \ldots\ldots\ldots\ldots\ldots (1)$$

and, from the "assumption of elasticity",

$$V_1' = V_1 - 2\Delta V_1 \quad \ldots\ldots\ldots\ldots\ldots (2a)$$

$$\text{and} \quad V_2' = V_2 + 2\Delta V_2 \quad \ldots\ldots\ldots\ldots\ldots (2b)$$

where V_1', V_2' are respectively the velocities of (1) and (2)

/after.....

after the collision.

From these equations one readily obtains that

$$v_2' - v_1' = -(v_2 - v_1),$$

which is all that needs to be proved. Theobald, however, although having written down equations (1), (2a) and (2b) in his own notation, finds it necessary to deal with the special case for which $\frac{m_1}{m_2} = -\frac{v_2}{v_1}$, and then uses conservation of momentum to show that "the masses will recede from each other with the velocities with which they approached." Nevertheless, the point is established that while kinetic energy may be a convenient concept to use in the analysis of perfectly elastic collisions, it is not essential; we can dispense with it if we make use of an alternative approach.

Two related queries are raised by this study of perfectly elastic collisions. These are:-

(i) Can the "assumption of elasticity" be re-stated in other, possibly more fundamental terms?

(ii) Although the total kinetic energy of the system is the same after the collision as it is before it, during the collision it is less than the initial amount. Can something significant be said about this temporary loss of kinetic energy?

We can answer these questions with the aid of the concept of potential energy, which we now introduce in the context of a gravitational field.

Let us return to the particle whose motion we considered in stating the "work-kinetic energy" theorem. We shall now suppose that the force responsible for accelerating this particle is entirely gravitational in origin. In this case it would be relevant to consider the properties of the gravitational force field. It is a familiar feature of a gravitational field that the work done by an external force in transporting a body at constant velocity from one point in the field

/to .....

to another, depends only on the positions of the end-points of the path and not on the shape of the path itself. This is another way of saying that the work done by the gravitational force when a body is transported from one point to another in a gravitational field is likewise independent of the path followed. If the gravitational field is approximately uniform, as it is in any small region near the earth's surface, then the proof of this result is elementary. A general proof may be found in any up-to-date first-year university textbook (e.g. Halliday and Resnick 1966, Sears and Zemansky 1964). Here we call attention to the fact that the proof does not follow from Newton's Second Law and the definition of work alone; the inverse-square law of gravitation has to be invoked as well.

Because the work done by gravity on a particle taken from one point to another is a unique quantity, independent of the path followed, it is possible to define a quantity called the gravitational potential energy of the particle (strictly speaking, of the particle-earth system). Using the idea of energy as "capacity to do work", we can argue as follows:- A body lifted from one level to another can do work in returning to its original level, and so gains increased capacity for doing work, or energy, as a result of being lifted. As this energy is "stored" (in this case in the gravitational field) it is spoken of as being potential energy. Alternatively, we can simply proceed to define a quantity we call gravitational potential energy in some such manner as the following (cf. Sears and Zemansky 1964): "The negative of the work done by the gravitational force when a body is taken from an arbitrary reference point A to some other point B is called the gravitational potential energy of the body (or body-earth system) relative to point A". "A" may be chosen to be anywhere we please (it is convenient in certain applica-

/tions.....

tions to banish it to infinity) since only differences in potential energy concern us in practice.

Referring once more to the particle whose behaviour we were considering earlier, and which we are now supposing to be subject to the gravitational force alone, let us imagine that it returns from the point B to the point A mentioned in the above definition. The work done by the gravitational force on the particle is equal to its gain in kinetic energy, according to the work-kinetic-energy theorem. But according to our definition of gravitational potential energy, the work done by the gravitational force is also equal to the loss in gravitational potential energy of the particle (or particle-earth system) as it falls from B to A. That is, the increase in kinetic energy of the particle is exactly counter-balanced by its decrease in potential energy. It is easy to see that this will be true no matter between what points in the gravitational field we consider the motion of our particle; it is a consequence of the way in which we chose to define kinetic energy and potential energy. We sum up this result by saying that for a freely falling body, the sum of its kinetic and potential energies is constant.

Of course this is none other than the Principle of Conservation of Energy. As derived here, it is clear that the Principle follows directly from the way in which one chooses to define kinetic energy and gravitational potential energy. This being the case one feels bound to ask what physical fact, if any, is expressed by the Principle. The answer is that applied to motion in a gravitational field, the Principle is a way of stating that the work done by the gravitational force on an object is independent of the path followed by the object. Any force field that possesses this property is said to be conservative, and the force exerted by the field is called a conservative force (cf Nuffield Guide IV, Feymman et al /1963, Kittel

/et.....

et al 1965). Some forces with which we are familiar appear to be conservative (e.g. the electrostatic force between charged objects), while others evidently are not. The sum of kinetic and potential energies of a falling raindrop is not constant, for example; presumably this is another way of saying that the frictional force exerted by the air on the raindrop is not conservative.

The introduction of the concepts of "conservative field" and "potential energy" now makes it possible to discuss the perfectly elastic collision of a pair of bodies in more detail. Suppose we assert that when two bodies collide and deform each other in virtue of the forces they exert on each other, these forces depend only on the extent of the deformation. (This is the sort of assumption that we make about a spring, viz. that the restoring force it exerts depends only on the amount by which it is stretched or compressed). Then first of all, it follows easily that the total kinetic energy of the system is the same after the collision as before. For during the period of contact between the colliding bodies, each body must travel the same distance, and provided that no friction-type forces act during the collision (which is implied by our assumption above), it follows from Newton's Third Law that each body must do the same amount of work on the other in the collision, and from the work-kinetic-energy theorem that the loss in kinetic energy sustained by the one body must equal to the gain in kinetic energy of the other. Second, during the intermediate stages of the collision when the bodies deform each other and then return to their original shape, the total kinetic energy at first decreases then returns to its original value. The mutual deformation of the bodies is accomplished by each doing work against the elastic restoring forces of the other. This lowers the total kinetic energy of the system, because in this process the forces the

/bodies.....

bodies exert on each other do not simply decelerate the one and accelerate the other. When the bodies return to their original shape, the forces now act so as to increase the kinetic energy of the system to its original value. If we define a "potential energy" for the two bodies in contact as being the negative of the amount of work done by their elastic forces when the bodies are deformed, then we can interpret the temporary loss of kinetic energy during the collision as being due to the storage of this energy as potential energy in the two-body system. It is then possible to say that in this type of collision, the sum of kinetic and potential energies of the colliding bodies is constant throughout. So again we have Conservation of Energy, which arises from our ability to define an appropriate potential energy for the system; and this, in turn, arises from the "conservative" property of the forces of interaction between the bodies, which in this case simply amounts to the fact that "the forces are the same on the way out as they are on the way in", as the P.S.S.C. textbook expresses it. (This book contains an admirably lucid and detailed analysis of perfectly elastic collisions.)

At this stage, it might seem that the introduction of the concept of energy serves two main purposes:-

(i) It facilitates the solution of certain problems in mechanics concerned with varying forces (e.g. the problem of finding the velocity acquired by a pendulum bob when it passes through its midpoint).

(ii) It enables the "conservative" property of certain forces or force fields to be embodied in a principle of conservation.

If this were all that was achieved by the use of energy and its conservation, the concept would scarcely deserve the emphasis which it receives today. But of course this is not all. In the middle of the nineteenth century it was realized

/primarily.....

primarily as a result of Joule's painstaking experimental investigation of the relationship between work and heat that the Principle of Conservation of Energy could be extended to embrace processes in which heat was generated by friction. (The Principle had been enunciated in this more general form by Mayer and Helmholtz, but it needed Joule's experimental work to establish it firmly.) This striking result led to the belief that non-conservative frictional forces were a macroscopic manifestation of microscopic processes controlled by conservative forces. This belief accorded well with the kinetic picture of matter, in terms of which the energy taken up by a body when it gets hotter is the increase in kinetic energy of the random motion of the molecules composing the body and in the mutual potential energy of the molecules. (Note that when we talk about "the mutual potential energy of the molecules", we are implicitly assuming that the forces of interaction between them are conservative.) The success of the Principle of Conservation of Energy in making clear the nature of "heat" resulted in its adoption as a Principle of general and universal validity, according to which the total energy in any isolated system was constant. There was no longer any question of energy "appearing" or "disappearing"; instead, it was realized that energy stored in one way could be made available or stored in some other way. As a natural consequence, it became customary to speak of "chemical energy", "electrical energy", "magnetic energy", in referring to the different possible "sources" of energy.

The Principle of Conservation of Energy has received such firm experimental support that an apparent violation of it in any new experimental situation would be regarded with the greatest suspicion. Evidently it expresses a fundamental truth about the physical world. Is there any way in which this truth could be expressed without direct reference to the

/concept.....

concept of energy? Indeed there is: we could say that all fundamental forces are conservative (cf Feynman et al/1963, Kittel et al 1965). There is no doubt, however, that this fundamental truth is most useful to us when we express it as the Principle of Conservation of Energy. In many situations, especially those which arise in atomic or nuclear physics, we may have little or no direct knowledge of the forces which act; in such cases we can still apply the Principle of Conservation of Energy and obtain valuable information.

One of the major consequences of Einstein's development of his Special Theory of Relativity was a discovery of a hitherto unsuspected relationship between energy and mass. How has this contributed towards the understanding of the nature of energy? By way of answering this question, it is necessary first of all to examine Einstein's discovery in some detail. As is well known, Einstein predicted that any amount of energy E would have a (relativistic) mass m (see our discussion of "relativistic mass" in Section 3.7.2) where the relationship between E and m is given by the celebrated equation

$$E = mc^2$$

c being the speed of light. A particularly interesting aspect of this prediction was that a particle of rest mass m_o must have associated with it an amount of "rest energy" equal to m_oc^2; this raised the possibility that energy "stored" as rest energy might become available as radiant energy in certain circumstances (i.e. that "matter might be converted into energy" in common parlance), or even that the reverse process might take place ("energy into matter"). All aspects of Einstein's prediction have received experimental confirmation, so that mass-energy equivalence must now be regarded as an experimental fact. This is not the place to talk of its applications in the field of nuclear physics; our immediate concern is with the light it sheds on the nature of energy.

/Does.....

Does the intimate relationship between mass and energy make the latter more or less "real" than it has seemed to be in our earlier discussions? Does the "equivalence" of mass and energy mean that they are really the same thing?

To some extent, the answer to the first of these questions hinges on the answer to the second, so let us consider the latter first. Conflicting statements are to be found in the literature on this subject: "In view of the fact that in relativity there is only one conservation law of mass or energy ('massergy'), the rigorous answer .....undoubtedly is: mass and energy are identical, they are synonyms for the same physical substratum" (Jammer 1960). "The question often arises whether mass really is energy. The answer clearly is that it is not. Matter has two fundamental properties, one of which is inertia, measured by its mass, and the other of which is energy, measured by its ability to do work, its ability to undergo changes... the properties of mass and energy are different, although their numerical measures are proportional" (Theobald 1966). "[Mass and energy] are physically equivalent quantities which correspond to one another. It is somewhat like the correspondence between the height of a mercury column and the air pressure" (Resnick 1968). "The numerical equivalence of m and E... is as little an equivalence as the linear relationship between force and displacement in Hooke's Law...The theorem concerned [$E = mc^2$] ... just states a relation between two properties, E and m, of a material system" (Bunge 1967). Now there is an important terminological point here which must be clarified before any attempt can be made to resolve this issue. If we follow the lead of a number of present-day workers in relativity and use "mass" to mean "rest mass" (cf. our discussion in Section 3.7.2and the references given there), there is no doubt tnat mass and

/energy.....

energy are different, even though for convenience they may be measured in the same units. Taylor and Wheeler (1966) express the matter thus in relativistic language: "The distinction between mass and energy is this: mass measures the magnitude of a 4-vector and energy measures the time component of the same 4-vector. Any feature of any discussion that emphasizes this contrast is an aid to understanding. Any slurring of terminology that obscures this distinction is a potential source of error or confusion." In simpler language, "the value of energy depends upon the inertial frame of reference from which the particle (or system of particles) is regarded. The value of the rest mass is independent of inertial frame" (Taylor and Wheeler, op. cit). (We may interpolate here the remark that only if we use the term "mass" to mean "rest mass" and if we measure energy in mass units have we any justification for using the unfortunate expression "the conversion of mass into energy". But this expression carries with it such misleading implications that it is best avoided.)

However, if we concern ourselves with the relationship between "relativistic mass" and energy, the issue is not so plain. Proponents of the identity of "relativistic mass" and energy would seem to argue that because when you measure the one you are, in effect, measuring the other, the two must be identical. Opponents of this view point out that we ordinarily conceive of mass and energy in two different ways, mass in terms of the inertia of the system, for which we can lay down a definite and direct procedure of measurement, and energy in terms of "work" defined by the "line integral" of the force, for which we also have definite and direct procedures of measurement. Therefore, they say, these are essentially two different properties; the fact that theory

/predicts.....

predicts and experiment confirms that there is a constant ratio between them does not affect the issue. This is not the kind of question about which one would choose to be dogmatic, but perhaps the (admittedly tentatively expressed) analogy suggested by Resnick and quoted above might be of some assistance here. The height of a mercury barometer may be and often is taken as a measure of atmospheric pressure, but no one would think of identifying the barometric height with atmospheric pressure. In the same way, while we may take the relativistic mass of a system as a measure of its total energy and vice versa, this does not compel us to identify the two concepts. Bohm (1965) has given the following reasoned summing-up of the matter: "...In Einstein's theory, mass and energy are not regarded as originating in essentially different ways. Rather, they are to be thought of as two different but related aspects of a single total process of movement. In such a movement, there is a relatively invariant capacity to do work, to interact with other systems, and to set them in motion, at the expense of movement in the original system. This is called energy. In addition, such a system has a certain inertia, or resistance to acceleration, as well as a certain gravitational attraction to other bodies. (Bohm appears here to be using the words "system" and "body" synonymously.) Both of these are proportional to a property called by the name of mass...It follows from Einstein's theory that whenever a system possesses a certain kind of energy, contributing a part ΔE to its total energy E it has all the properties (i.e. inertial and gravitational) that physics ascribes to a corresponding contribution $\Delta m = \frac{\Delta E}{c^2}$ to its mass, which is part of the total mass $m = \frac{E}{c^2}$". Bohm also attempts to clarify the difference between "rest energy" and energy in its more familiar manifestations by associating

/rest.....

rest energy with "inward movements", in contrast to the "out ward movement" to which the more familiar manifestations correspond.

Whether or not one accepts this particular view of mass-energy "equivalence", the relationship between mass and energy revealed by the theory of relativity does seem to compel acceptance of the view that "energy" is as "real" as "mass", whatever "real" may mean in this context. Burniston Brown (1965) has argued that while mass is a measure of "something physical (e.g. matter)", energy is something quite different in nature -a mere result of performing mathematical operations on certain measures, undeniably useful for "book-keeping" in certain instances, but obviously not "something physical". Brown is presumably speaking from a pre- (or anti-) relativistic standpoint - indeed other writings of his (cf Burniston Brown 1967) would seem to confirm this - because it is difficult to see how this particular view of the difference in nature between mass and energy can be sustained in the light of the proven "equivalence" of these concepts. The truth is rather that mass and energy are both what have been called "constructs" (cf. Margenau 1942, 1950), i.e. quantities obtained by performing mathematical operations on the results of certain measurements. Neither is a <u>direct</u> manifestation of sensory experience, i.e. neither is "immediately given" (cf. Margenau 1950).

If we agree that energy is as "real" as mass, the question still remains as to what the meaning of "reality" is in this case. The problem has been dealt with in detail by Theobald (1966), who reaches the conclusion that since energy is a theoretical concept which plays a fundamental role in a well-confirmed theory (indeed, in a wide variety of well-confirmed theories), it must be considered to be "real"

/(with.....

(with a small r). Indeed it is in this sense that we speak of the "reality" of any physical concept. However, its "reality" emphatically does not imply that energy is something substantial, like the caloric fluid of old, that can literally be possessed by a body or system of bodies, can literally flow from one body or system to another, or can literally be transferred or transformed from one form to another. (In this connection cf. Dingle and Samuel 1961 and Armstrong 1965a. The former reference includes an extremely lucid but fruitless attempt by Dingle to convince the late Viscount Samuel of the "reality" of energy in the sense in which we have used the term above, and of the validity of the description of physical processes in terms of energy transfer.) Many physicists habitually use this mode of speech when referring to energy, but it may safely be assumed that they have a perfectly good understanding of what they are talking about. Whether their meaning is always apparent to their listeners or readers is another matter, which will be discussed in its proper place below (see the end of Section 4.15.2).

To sum up: energy is a "construct" - a particular physical quantity in terms of which one may formulate a Principle of Conservation which states a physical fact of fundamental importance. It is "not a substance, fluid, paint or fuel which is smeared on bodies and rubbed off from one to another" (Arons 1965). The frequently-used expression "flow of energy" must not be taken literally, but is nevertheless a convenient way of referring to a process in which energy is "lost" by one system and "gained" by another. It can aid one's thinking about such processes, just as the use of imaginary "field lines" can aid one's understanding of problems in electricity and magnetism.

/4.15.2.....

4.15.2 The teaching of the concept of energy

In earlier decades (i.e. prior to the nineteen-sixties), it was customary for work and energy to be taught in elementary courses according to the following pattern (cf Duncan and Starling 1918, Stead 1948, Smith 1953, Ference et al 1956, Weber et al 1959; see also Constant 1963, Stollberg and Hill 1965 for more recent treatments of this type.):
(i) to introduce the concept of work, appropriately defined;
(ii) to define energy as the capacity for doing work;
(iii) following on these definitions, to introduce the concepts of kinetic and (gravitational) potential energy.
More recently, there has been a tendency in first-year college or university textbooks to place the emphasis more firmly on energy, in recognition of the fact that the purpose of the introduction of "work" is to pave the way for energy as a more fundamental concept. In such treatments (cf Feather 1959, Sears and Zemansky 1964, Arons 1965, Kittel et al 1965, Halliday and Resnick 1966, Alonso and Finn 1967), the initial step is also to define work (if it is a "calculus" text, work will be defined as the line integral of force, i.e. $\int \vec{F}.\vec{ds}$) but no attempt is made to furnish a general definition of energy. Instead the work-kinetic-energy theorem is derived (cf. our discussion above), with kinetic energy defined (in this non-relativistic approach) as "$\frac{1}{2}mv^2$"; "potential energy" functions for different conservative force fields are then defined in terms of work done by these force fields (again cf. our discussion above), from which the conservation of the sum of kinetic and potential energies in conservative fields immediately follows. This approach helps to bring out the real significance of energy as something which is a conserved quantity. In other recent treatments (e.g. Atkins 1965, Orear 1961, 1967, Feynman et al 1963 and Cotts and Detenbeck 1966), the approach is even more radical; energy is introduced

/before.....

before work, although again no attempt is made to provide a general definition of energy. Feynman et al discuss energy as a conserved quantity long before getting round to a formal definition of work. They begin their chapter on energy conservation with the question, "What is energy?" The answer they give will not satisfy the definition-hungry teacher or instructor, but it nevertheless demands quotation: "It is important to realize that in physics today, we have no knowledge of what energy is....However, there are formulas for calculating some numerical quantity, and when we add it all together it gives '28' - always the same number. It is an abstract thing in that it does not tell us the mechanism or the reasons for the various formulas." Atkins's approach is to state the formulae for various "forms" of energy early in his chapter devoted to the subject, on the grounds that this is the best way to combat the belief that energy is a "sparkling fluid" that can be transferred from one body to another.

The studied avoidance of the definition "energy is the capacity for doing work" in all these new treatments of the concept may come as a surprise to those who, like the present writer, were brought up on it. (Incidentally Theobald in his exhaustive study of the concept (1966) nowhere mentions this definition: the nearest he comes to it is in his reference, quoted above, to the energy of matter as being "measured by its ability to do work, its ability to undergo change".) Possibly the chief reasons for its omission are that there is much more to the concept of energy that can be conveyed by the phrase "capacity to do work" (contrast this with momentum, which can be accurately and completely defined in as few words), and that it is consequently downright misleading to parade this phrase as a definition. (As the Second Law of Thermodynamics teaches us, "capacity for doing work" would not be an adequate definition of "heat" energy.)

/The.....

The new approaches to energy and work just outlined, while suitable for use in first-year courses at universities or colleges, are too formal and abstract to commend themselves for use at school. The P.S.S.C. and Nuffield treatments of the subject represent satisfactory attempts to approach it in the same spirit but in a manner readily assimilable by schoolchildren. Here, too, the emphasis is on energy as a primary concept, measured in units of work, ultimately to be shown to satisfy a Principle of Conservation. We have seen in Section 3.9 of this thesis how energy and work are introduced, together with the idea of the transformation of energy. In the senior school, it is appropriate to give a more formal definition of work than that which was supplied earlier; and also to give a quantitative treatment of kinetic and potential energy. At first, the teacher could remind pupils that in the earlier standards they found that the quantity called "work", measured by "force times distance", was used as a measure of the amount of energy transferred from something to something else. He could then define "work" formally as "the product of the force and the component of the displacement of the point of application of the force in the direction of the force." It follows from the true definition that we may use the expression $Fs\cos\theta$ for the work done by a constant force, where F and s are the respective magnitudes of the force and the displacement and θ is the angle between them. Perhaps it would be as well for the teacher to stress that while one can be led to the concept of work by consideration of the processes in which "energy transfer" apparently takes place, the actual <u>definition</u> of work makes no reference to energy. (The statement that "work is a measure of energy transferred", while correct, is not a <u>definition</u> of work.)

It is usual when discussing "work" to say that although one may get very tired while holding a heavy suitcase for a long time, one has done no "work" in the physical sense of the

/term....

term because one has not moved the point of application of the force one exerts. However, it can be convincingly argued that one does do work in these circumstances (cf. Feynman/1963 et al, Nuffield Guide (I)). The muscles of our arms are striated or skeletal muscles whose individual fibres contract momentarily on receipt of a nerve impulse, and then relax. While one's arm holds a heavy weight, a large number of the muscle fibres will be in a state of contraction at any given instant while the others are relaxed. It is through the incessant bringing into action of different muscle fibres that work is done inside one's body.

Attention should now be focussed on the two basic types of energy, viz. kinetic and potential energy. Kinetic energy should be introduced by consideration of the work done on a body by a constant resultant force in accelerating it from rest to a velocity of magnitude v. This work is easily shown to be $\frac{1}{2}mv^2$, and is clearly the work that can be done by the body when some opposing force brings it to rest. The "relativeness" of kinetic energy is a point worth discussing after the concept has been introduced, i.e. the kinetic energy of an object depends on the choice of reference frame (cf. Nuffield Guide (IV)). Someone riding along a road at constant velocity has a lot of kinetic energy relative to us if we are stationary at the side of the road; but if we ride alongside him with the same velocity, his kinetic energy relative to us is zero.

This presentation of kinetic energy could be reinforced with simple qualitative class experiments in which pupils give kinetic energy to suitable objects by pushing them or by allowing them to suffer collisions with other objects already in possession of kinetic energy (cf. Nuffield Guide (IV)). Elastic collisions between two similar objects under near-

/frictionless.....

frictionless conditions show up the transfer of kinetic energy very neatly, especially if one of the objects is initially at rest. Of course, what is actually seen to happen is that the objects exchange velocities in the collision, and this could equally well be interpreted as an exchange of momentum. But this does not negate the value of these experiments in this context. The pupil is now required to think about the work done by an accelerating force and its relation to $\frac{1}{2}mv^2$, and to interpret what he sees in these simple experiments in the light of these ideas. It is inevitable that he should also think about momentum, but this is all to the good; he will shortly be faced with the discovery that kinetic energy is not conserved in all kinds of collisions, whereas momentum is. If a modern version of the eighteenth-century debate of "mv versus mv^2" springs up in the classroom, the teacher can stress that both momentum and kinetic energy are useful concepts and that the particular virtues of kinetic energy will become apparent in due course. Pupils who have been through the introduction to energy recommended for the junior years of the secondary school should be able to accept this view easily enough.

Although the P.S.S.C. textbook chooses to introduce the concept of potential energy in the course of the study of elastic collisions and presents a very elegant treatment of this topic, the average senior school pupil would probably find potential energy easier to grasp if he met it first of all in the shape of gravitational potential energy. In the treatment of this topic, the points to stress are:-

(i) When a body is lifted from one level to another, work must be done on it.

(ii) If the lifting force applied to the body exactly counterbalances its weight, the net work done on the body by all the

/forces.....

forces acting on it is zero. (Contrast this with the situation used to illustrate kinetic energy.)

(iii) However, the work done against the gravitational force by the lifting force is "recoverable", because one could arrange for the body to do this amount of work on some other body if it were allowed to descend to its original level. Therefore the act of lifting the body has given it additional capacity for transferring energy to something else (i.e. additional "capacity to do work"). We could say in this case that the body has received an increase in gravitational potential energy. Strictly speaking, it is the system body-plus-earth that has received the increase in gravitational potential energy. It is perhaps a pity that the well known formula for differences in gravitational potential energy, $\Delta W = mgh$, should be misleading in this respect, since it <u>apparently</u> depends only on the mass of the body lifted and the height through which it was lifted. But it should be remembered that $g = \frac{G m_E}{R^2}$, where the symbols used have the appropriate meanings; in fact the full formula should read $\Delta W = \frac{G m m_E h}{R(R+h)}$, which makes it plain that the increase in potential energy depends on the masses of both the body and the earth as well as on the separation of their centres. For small values of h (i.e. in practically all cases of relevance to pupils) this formula reduces to $\Delta W \cong \frac{G m m_E h}{R^2} = mgh$, since $g = \frac{G m_E}{R^2}$. It is not suggested that pupils should see the above <u>general</u> expression for potential energy in a gravitational field - the simple expression for the very nearly uniform gravitational field in which they live will be quite adequate - but the teacher should make some effort to convince pupils that this potential energy is "stored" in the "invisible spring" represented by the gravitational field. The underlying idea is an important one, since it will be

/encountered.....

encountered again when the pupil meets energy associated with electric and magnetic fields. One way in which this idea might be successfully conveyed is the "experiment" suggested by Rogers (1960) (see also Nuffield Guide (I)), in which the pupil lifts a weight and at the same time closes his eyes and imagines himself stretching the "invisible spring" connecting the weight and the earth.

(iv) It should be stressed that the work done against the gravitational force in lifting the body from one level to another does not depend on the path followed; if it did, we could not obtain a unique answer for the difference in potential energy between the two levels, and it would therefore be impossible to define gravitational potential energy in any meaningful way. This point is not usually mentioned in elementary treatments of energy. It is difficult to see why not. After all, as we pointed out earlier, the conservation of energy depends on this fact, and it can be very easily proved for a uniform gravitational field. The path along which an object is lifted can be approximated by a sequence of small alternating horizontal and vertical segments; along the horizontal segments of this path, the gravitational force does not work and no work is done against it, while the total work done against the gravitational force along the vertical segments of the path is simply equal to the weight multiplied by the total length of the vertical segments, i.e. the height through which the object is lifted.

As the idea of potential energy is at least as important as that of kinetic energy, it is desirable that other examples of it should be discussed before the pupil is required to work out problems on work and the conservation of energy. The obvious example is the potential energy of a compressed or stretched spring. Time will probably not permit a

/qualitative.....

quantitative study of this case in the classroom, but a general consideration of it will help to impress upon the pupil those features which characterize potential energy in general. The matter is thus excellently summed up by MacLachlan and McNeill (1964b): "....[A system has] potential energy ... when a force external to the system is required to prevent the action of a force internal to the system. Or, if a continuing force exists that can produce motion when a restraint is removed, then the object on which the force is acting possesses potential energy."

Of course, the most important result to emerge from this study of mechanical energy should be the conservation of mechanical energy in the absence of friction. The well-known "derivation" of this principle should be given for the case of interchange between gravitational potential energy and kinetic energy, and examples of this type of interchange as well as of the interchange between elastic potential energy and kinetic energy should be discussed. It would be instructive for pupils to carry out simple qualitative or quantitative experiments in this connection: the Nuffield Guide (IV) describes several experiments of this type. Perhaps the simplest of them all is one in which the pupil observes the swinging of a pendulum and is asked to comment on what he sees from the point of view of energy conversion. (There is also Galileo's modification of this experiment, in which the motion of the supporting string is intercepted by a pin or peg and the bob is seen to rise as nigh as it did in the absence of the pin.) These experiments and others, in all of which friction inevitably plays its part, can serve as a most useful introduction to the Principle of Conservation of Energy in its more general form.

We have referred above to the use of the expression "flow of energy", and the possibility that this might lead the

/uninitiated.....

uninitiated to conceive of energy as a real fluid. We do not think that this constitutes a serious danger in the introductory teaching of energy. By this time the pupil will have had considerable experience of working with abstract physical concepts such as velocity, acceleration, mass, force and momentum. As energy is introduced on the same basis as that on which these concepts were, he should be able to appreciate that energy is no more a "thing" than are these other concepts. The useful feature of energy is that when it "disappears" from one system, we find it turning up in another; also, if within a given system, energy of one"form" becomes less, this is compensated for by a corresponding increase in the amount of other "forms" of energy within the system. It is usual (and convenient) to speak of a transfer of energy within one system or from one system to another as a "flow of energy". If the nature of energy as an abstract physical concept is properly grasped by the pupil, he should have no difficulty in recognizing this usage for what it is, viz. as a "manner of speaking" not to be interpreted too literally. We cannot think that damage can be done to anyone's understanding of physics in general or energy in particular by the tendency to refer to "the flow of energy", or to this or that system possessing a particular "form" of energy. No fundamental truth is obscured by this looseness of language (cf. in this respect the pseudo-definition of mass as "quantity of matter"); on the contrary, it can help the beginner to understand the idea of energy transfer and so lead to an understanding of conservation of energy.

4.15.3 Power

It is easy to justify the introduction of the concept of power or the rate of doing work on the grounds that it is frequently of practical interest to know how fast energy is

/being.....

being transferred by a given machine, or how quickly it is being "used up". Every pupil knows that the more "powerful" the engine is in a given car, the faster it will go under given conditions. The unit of power can be introduced immediately after the concept has been defined. In a modern syllabus, only two units of power require formal definition, viz. the joule per second, or watt, and its multiple the kilowatt. The horse-power can simply be stated to be equal to 746 watts (to three significant figures). The energy unit known as the kilowatt-hour, must be introduced because of its practical importance. This will help to bring home an aspect of energy to the pupil - that energy is "what you pay for".

Sample calculations to illustrate the use of "power" should not and need not be elaborate. Examples requiring the calculation of the power expended by someone running up a flight of stairs or climbing a mountain in a given time are of some interest and could well be done in the classroom. Also interesting are problems dealing with motor vehicles travelling at constant velocity along a horizontal road. If the total frictional force is given (not the coefficient of friction and the mass of the car) and the speed of the vehicle, the power developed by the engine when the vehicle travels on a level road is simply the product of these two quantities.

CHAPTER 5: ADVANCED HEAT, KINETIC THEORY AND THE PHYSICS OF MATTER IN BULK.

5.1 Introduction

Once the principles of mechanics essential to the rest of the course have been introduced and elaborated upon, there are a number of possible ways of proceeding In the Common Basic Syllabuses for both Physics and Physical Science, it is recommended that the foregoing material be followed immediately by a wider study of the conservation of energy, leading to the First Law of Thermodynamics. In the Physics syllabus, this is then followed by "The Particle Structure of Matter", and "Intermolecular Forces", which are followed in turn by a study of "Gases, Solids and Liquids", the section concerned with gases dealing among other things with elementary kinetic theory of ideal gases, kinetic interpretation of temperature, Kelvin and Celsius temperature scales and "ideal gas laws". (In the Physical Science syllabus, the sections concerned with intermolecular forces and with matter in bulk are transposed to the "Chemistry" part of the syllabus.)

To some extent this is a logical sequence, at any rate as far as the placing of the First Law of Thermodynamics is concerned, for the pupil by this time knows enough about temperature and the measurement of "heat" to be able to understand the significance of Joule's experiments. Also, as we have already remarked, "conservation of energy" in the wider sense finds a ready interpretation in terms of the kinetic theory of matter. On the other hand, we think it unfortunate that the interpretation of temperature in terms of average molecular kinetic energy should precede a study of the ideal gas "laws" or indeed of the Kelvin temperature scale;

/this,.....

this, we think, is putting the cart before the horse, at any rate in the senior school. We shall ourselves discuss temperature and the gas laws before the First Law of Thermodynamics and the kinetic theory of matter, but we appreciate that there is much to be said for a compromise approach in which, say, the First Law is discussed immediately after mechanics, and Boyle's Law is then given a naive derivation from elementary kinetic theory. We would still insist that the Kelvin scale of temperature should be properly established before temperature is "interpreted" in terms of average molecular kinetic energy.

5.2 Temperature and the gas laws

5.2.1 The traditional approach

The traditional approach to these matters (cf Gordon and Neser, no date, Prinsloo et al 1965) begins with an experimental investigation of the variation with temperature of the volume of a given mass of gas kept at constant pressure. (In this case "temperature" means "temperature as measured by a mercury thermometer" - a fact which the textbook author may well not bother to state.) A graph is plotted of gaseous volume versus temperature and extrapolated backwards to give an "absolute zero." Depending on how sophisticated the textbook authors are, the emergence of the absolute zero will be interpreted in various ways. Thus Gordon and Neser (op. cit.) confidently announce that "In terms of the Kinetic Theory of heat...it does mean that at - 273°C there can be no molecular motion," and likewise Prinsloo et al tell us that "By die temperatuur sal geen stof meer enige warmte besit nie: alle molekules van alle stowwe sal, volgens die molekulêre teorie, stilstaan." ("At this temperature no substance will have any more heat; all molecules of all substances will, according to the molecular theory, cease

/move.".....

move".) One might well ask, "Which 'kinetic theory' or 'molecular theory'?" The next step is to enunciate "Charles's Law" in some such form as "At constant pressure the volume of a fixed mass of gas increases by $\frac{1}{273}$ of its volume at 0°C for each degree Celsius rise in temperature" (cf. Tyler 1951). "Absolute" (or Kelvin) temperature is then defined as Celsius temperature plus 273 degrees, and Boyle's Law and the form of Charles's Law just quoted are combined to give the ideal gas equation.

The principal objection to this approach is that, explicitly or implicitly, the temperatures specified in it are temperatures that are supposed to be read on a mercury thermometer. In extrapolating a volume versus temperature graph towards an absolute zero, we are moving into a region where, as some pupils will realize, it is quite impossible to take readings of temperature with a mercury thermometer since the mercury will have solidified. What meaning, then, can we attach to a temperature of, say, -100°C, that appears on the temperature axis of the graph? Warren (1961, 1965) has forcibly criticized this kind of extrapolation and drawn attention to its dangers, yet it continues to be applied: in particular, it is employed in so carefully-thought-out a course as the Nuffield Physics course (cf Nuffield Guide (III), in which it is recommended that pupils should plot a pressure versus mercury-scale-temperature graph for dry air and extrapolate it in the usual way). Clearly the subject needs to be discussed further, and since the Nuffield approach to it is characterized by a sense of responsibility that is conspicuously absent from other approaches, we shall use it as our point of departure.

The Nuffield Guide makes the clear statement that "Pupils should emerge with a clear idea that, judged by a mercury

/thermometer,.....

thermometer, gas pressure runs down a straight line as temperature falls, a straight line directed towards zero pressure at a temperature somewhere between -250°C and -300°". In the immediately preceding paragraph, it is suggested that "If liquid air is available, it is very interesting to dip this 'gas thermometer' in liquid air and see what it says about the boiling point of air." What can it say about the boiling point of air? If we interpret this question as being an instruction to read off from the graph the temperature corresponding to the pressure of the "gas thermometer" when it is dipped in liquid air, then we obtain from this graph a temperature which, judged by a mercury thermometer, would seem to have no meaning. Now this criticism cannot be shrugged off with the remark that it is merely the perverse application of "the operationalist demand to 'define' every physical concept by means of a set of laboratory operations" (if we may quote Bunge (1967) out of context). For here we are directly concerned with the measurement of temperature, and to talk of temperatures of the order of -100°C or less as though they were the results of measurements we must have a precise idea in mind of the temperature scale in terms of which these measurements are expressed. If, at this stage, the only temperature scale we (or our pupils) have in mind is that defined by the liquid-mercury-in-glass thermometer, then temperatures having the values just quoted have literally no meaning. Yet many pupils and teachers do accept that these temperatures have a meaning, and the Nuffield Guide suggests that one should find out what the "gas thermometer" says about the boiling point of air. How is this to be explained? Surely the answer is that consciously or unconsciously, temperatures below the freezing point of mercury are treated here not as mercury-scale temperatures but as temperatures

/on.....

on a gas scale whose graduations are identical in size with those of the mercury scale in the region where the two overlap. We suggest that it would promote understanding of temperature and its measurement if this switch of temperature scales were brought into the open and discussed at the appropriate time; specifically, we recommend that the idea of a temperature scale defined by the properties of a gas should be introduced prior to any discussion of "absolute temperature" and "absolute zero." Such an approach would not, we think, confront teacher or pupils with any major difficulties: we outline it below.

5.2.2 An improved approach

First, the principles studied in the junior years for the setting-up of a temperature scale should be reviewed. Then the question of a standard thermometer setting up a standard temperature scale can be raised. It can be pointed out that we may construct a variety of different Celsius thermometers based on different thermometric properties, according to the principles previously outlined, but that we must not expect that these different thermometers will agree with one another except at the fixed points. Which thermometer is right? Whichever one we choose as standard. Physicists have found it very much to their advantage to use as a standard a type of thermometer that uses as its thermometric property the change of pressure of a gas kept at constant volume. To see why they made this choice, we need to know more about the way in which gases behave.

At this point, Boyle's Law can be established. Nowadays, the idea of carrying out experiments to "verify" Boyle's (or any other) Law is thoroughly discredited, and teachers will want the Law to emerge from measurements made by the pupils themselves. Still, it would be dishonest to pretend

/that.....

that pupils are "discovering" anything that is other than very well known, and which indeed may be very well known to them also. The Nuffield Guide (III) suggests that pupils be told of Boyle's original investigations but not of his results, and that they then be asked to try what Boyle tried. "Some of you know what Boyle discovered, and in that case you will be trying your skill against the apparatus, to see whether you agree with the relationship that Boyle found." As to the method of trying it, many schools will possess the familiar type of Boyle's Law apparatus in which the applied pressure has to be calculated by adding or subtracting a difference in mercury levels to atmospheric pressure. This presents no real difficulty to senior school pupils and cannot be held to obscure the essential simplicity of the experiment. If this apparatus is not possessed by the school, there is a case for trying the Nuffield apparatus, in which the gas sample is enclosed in a wide glass tube with a volume scale beside it and pressure is measured on a Bourdon gauge graduated in absolute pressure; the pressure is changed by driving in oil from a reservoir.

When finding the "PV" products, there is a strong temptation for pupils to write down the answers as four-figure numbers, since these are the numbers obtainable from logarithm tables. The implied four-figure accuracy is quite unjustified, and teachers should seize the opportunity to issue the appropriate warnings here. Finally, an important point to make concerning Boyle's Law is that it does not <u>exactly</u> describe the behaviour of real gases, but that the lower the pressure, the more closely does the gas seem to "obey" Boyle's Law. An <u>ideal gas</u> can then be defined as a gas that would obey Boyle's Law exactly under all circumstances.

/Next,.....

Next, pupils can study gaseous expansion. The first point to be made would be that at constant temperature, the volume of a gas depends on the pressure. Therefore investigation of the change of volume with temperature should be carried out at constant pressure if the results are to be easily interpreted. If the appropriate apparatus is available, the standard experiment to investigate the change of volume of a gas kept at constant pressure with change in mercury-scale temperature can then be carried out, leading straightforwardly to the result that in terms of the mercury temperature scale, the gas expands uniformly with temperature. (It would be good teaching tactics to inject constant reminders into the discussion that the temperatures spoken of are, in fact, mercury-scale temperatures.) The teacher can now make the point - there would be no time to establish it in the classroom - that if different gases were used in turn in the apparatus, they would in general be found to behave in the same sort of way. But a further interesting result emerges when observations are made on the thermal expansion of different gases. This is the fact that at a constant pressure, all gases expand between the lower and upper fixed points by the same fraction of their volume at the lower fixed point.

[These facts about the variation of gaseous volume with temperature were independently established by J.A.C. Charles in 1787 and by J.L. Gay-Lussac in 1802. In view of the way in which "Charles's Law" is often rendered nowadays (cf. the quotation above), it is interesting to note Gay-Lussac's own statement of his findings (Magie 1935):

1. "All gases, whatever may be their density and the quantity of water which they hold in solution, and all vapours expand equally between the same degrees of heat.

/2.....

2. For the permanent gases the increase of volume received by each of them between the temperature of melting ice and that of boiling water is equal to $\frac{80}{21333}$ of the original volume for a thermometer divided into 80 parts, or to $\frac{100}{26666}$ of the same volume for the centigrade thermometer." It is possibly significant that Gay-Lussac thought fit to state his second conclusion in terms of the expansion between the lower and upper fixed points, and not as a fractional expansion per degree.]

The variation of pressure with temperature of a gas kept at constant volume should next be studied, leading to the conclusion that at constant volume the pressure of a gas increases uniformly with the mercury-scale temperature, and the pressure of all gases increases from the ice to the steam point by the same fraction of the pressure at the ice point. At this stage, instead of following the extrapolation procedure criticized above, the teacher can suggest to pupils that the facts of gaseous behaviour suggest not only that a gas might be an excellent substance to use in a thermometer, but also that a temperature scale defined with a gas thermometer might be made the standard. The implications of the use of such a thermometer as a standard can then be studied in detail. First, the process of establishing a Celsius scale with a constant volume gas thermometer can be described or, better, demonstrated. That is, the pressure of the gas should be measured at the two fixed points, and the results plotted on a graph of pressure against temperature. By definition, the temperatures of the two fixed points are 0°C and 100°C. (Or rather, that of the upper fixed point would be 100°C if the pressure were one standard atmosphere. This is not a matter about which a great deal of fuss need be made; if the pressure is not one standard atmosphere, then either the

/teacher.....

teacher could say "let us call the temperature of this fixed point 99.2°C" (or whatever would be the appropriate value) or he could decide that the temperature to be assigned to the upper fixed point might just as well be called "100°C", since the percentage error resulting from such a decision is of little consequence here.) The straight line joining these two points is the graph that defines the temperature scale for this gas thermometer. With the aid of this graph we can now use the gas thermometer to measure temperatures on its own scale.

Even now, one need not hasten to consider the implications of extrapolating this temperature-defining graph back to where it cuts the temperature axis. Instead, one could consider such questions as, "If we had a number of different gas thermometers using different gases in each, would they each give the same temperature reading when required to measure the temperature of the same system?" Experimentally one would find this to be true to a good approximation. Even more important is the fact that the agreement between gas thermometers gets better and better when the gases in them are used at lower and lower pressures. For example, suppose we have at our disposal a pair of constant volume gas thermometers, one containing oxygen and the other hydrogen. To begin with, we adjust the thermometers so that at the ice point, the pressure is 760 mm Hg. Then we use the thermometers to measure the temperature of a water bath. We will find that the agreement between the thermometers is nearly but not quite exact. Then we remove enough of the gas from each thermometer so as to make the pressure in each 380 mm Hg at the ice point, and we repeat the measurement of the temperature of the bath. The agreement between the thermometers will be better than before; if we make the pressure at the

/ice.....

ice point lower still, the agreement will be still further improved. Since it was found earlier that a gas behaved more like an ideal gas the lower its pressure, the foregoing results can be stated as follows:-

Gas thermometers using different gases agree with another better as the gases in them approach the condition of an ideal gas.

For this and other reasons, scientists have decided to use the properties of an ideal gas as the basis of their definition of a standard temperature scale.

At this stage of the argument, it would be appropriate to look once again at the (real or hypothetical) graph of pressure versus temperature which defines a "real gas" temperature scale. If the zero of this temperature scale is re-defined to be the point at which the straight-line graph cuts the temperature axis, then the pressure of the gas in the thermometer will be directly proportional to the temperature. (The teacher should stress that this is a result of the definitions by which this temperature scale is set up, not a deduction from experiment.) If one was to set up a Celsius scale using a constant-volume apparatus from the school laboratory, one would find that the graph of pressure against temperature would cut the temperature axis at around $-273^{\circ}C$. Could the thermometer have been an ideal thermometer containing an ideal gas, the graph obtained would cut the axis at 273.15 Celsius degrees to the left of the ice point. (This, it should perhaps be mentioned, is a deduction from the behaviour of real gases.) Therefore the pressure of an ideal gas is directly proportional to a new temperature, T, where T is defined by

$$T = 273.15 + t,$$

t being the temperature in degrees Celsius.

/The.....

The pupil is now ready to be introduced to the modern definition of the Kelvin temperature scale, based upon a single fixed point (cf Sears and Zemansky 1964, Zemansky 1964, Brommert 1965). He can first of all be told that (i) it is possible to set up a temperature scale using only one fixed point, as will shortly be shown to him; and (ii) that scientists decided upon the new, single, fixed point (the "triple point" of water) on the grounds of its superior reproducibility. To explain point (i), all that is needed in principle is to set up the following temperature-scale-defining equation:

$$\frac{T}{273.16} = \frac{X}{X_3}$$

where X is the value of the chosen thermometric property at some temperature T other than the triple point, and X_3 is its value at the triple-point. Thus the temperature of the triple point is arbitrarily assigned the value of 273.16 kelvin (or 273.16K), and the above equation enables any other temperature to be found if the value of X_3 and the appropriate value of X are known. Pupils should note that if a temperature scale is defined in this way, then the changes in temperature are proportional to the changes in thermometric property, just as before; but that in addition, the temperature itself has been made directly proportional to the value of the thermometric property.

For the scale now officially adopted as standard, the thermometric property is, in effect, the pressure of an ideal gas kept at constant volume. (In practice, ideal gas scale temperatures are calculated from suitable formulae into which are substituted the readings taken with a real gas constant volume thermometer.) By specifying the temperature of the new fixed point to be 273.16 kelvin, the international

/committee.....

committee responsible for laying down the new scale has ensured that the temperatures of the ice and steam points remain 273.15 K and 373.15 K respectively, just as they were on the old "absolute temperature" scale defined by the addition of 273.15 "degrees" to the Celsius scale.

To complete this discussion of the standard temperature scale as such, the teacher must tell pupils of the existence of the true Kelvin scale of temperature, i.e. the thermodynamic scale of temperature. The story he can tell can take the form of the following simple assertions:- "Lord Kelvin showed last century that it is possible to define a temperature scale that is independent of the properties of any substance. The reasoning he used forms part of the branch of physics called 'thermodynamics', which is the advanced study of heat as a form of energy; and so we call this speical temperature scale the 'Kelvin thermodynamic scale'. Kelvin also used thermodynamics to show that the ideal gas temperature scale was identical with his thermodynamic scale, so when we use a gas thermometer to measure 'ideal gas scale temperatures' for us, we are also measuring Kelvin scale temperatures."

The enterprising pupil may wonder why our definition of a temperature scale is so chosen as to make changes in temperature proportional to changes in thermometric property, and whether, having made this choice, there is any compelling reason for our having chosen the Kelvin thermodynamic scale as the standard temperature scale. The answer to the first question is that this is the simplest choice we can make, but that in this, as in other situations in physics, we were subject to no compulsion. As for the other question, the answer is that the choice was dictated by convenience (cf Rogers 1960). The temperatures on the Kelvin thermodynamic scale are the

/temperatures.....

temperatures which appear in the formulae of thermodynamics; thus, for example, Stefan's Law of Radiation may be written in the form

$$R \propto T^4,$$

which implies that the energy radiated by a body is proportional to the fourth power of the Kelvin temperature. Also, the mean kinetic energy of the molecules of an ideal gas is proportional to the Kelvin temperature; but we have no right to reason backwards from this fact and claim that we are therefore compelled by Nature to define our temperature scale so as to achieve this proportionality.

On the subject of temperature <u>scales</u>, there remains to be considered the apparent coincidence of the mercury-in-glass and Kelvin gas scales over the range of validity of the former. It is the agreement between these scales which paved the way for the establishment of the ideal gas scale and which justifies the use of linearly-graduated mercury-in-glass thermometers in elementary laboratories. There is nothing more to be said about this agreement than that it is a purely fortuitous circumstance, resulting from the properties of mercury.

One of the advantages of repeated reference to the standard temperature scale as an "ideal gas" scale is that one does not thereby commit oneself in advance to peculiar theories about what happens to real matter at the absolute zero. It is true that an ideal gas must be expected to exert no pressure at the absolute zero, and that the average kinetic energy of its molecules must correspondingly be zero at the absolute zero; but since these predictions concern an unrealizable "ideal gas", they need cause us no alarm. Pupils should be made aware of this, and more especially of the fact that modern (quantum-mechanical) theory does <u>not</u> predict the vanishing of molecular kinetic energy of <u>real</u> matter at the absolute zero. Only a system of particles which did not interact with each other nor with anything else could have no

/kinetic.....

kinetic energy at the absolute zero.

Pupils are bound to have an interest in whether the absolute zero is attainable or not; after all, the definition of our standard temperature scale in a particular way does not necessarily imply that it will be. If we use a gas thermometer containing helium at low pressure, we can measure temperatures down to about 1 K with it. But temperatures several orders of magnitude lower than this can be attained and measured; by making ingenious use of the magnetic properties of atoms and their nuclei, physicists were able in 1960 to produce temperatures of as low as about 1.2×10^{-6} K in a specimen consisting of copper wires (cf. Zemansky 1964, Mendelssohn 1960). As for the actual attainment of the absolute zero, it is now believed to be a law of nature (The Third Law of Thermodynamics) that while in principle we may be able to come as close as we please to the absolute zero, it is impossible actually to reach this temperature (cf. Zemansky, Mendelssohn op. cit.).

As here presented, the following relationship is seen to hold by definition for the pressure and temperature of an ideal gas:

$$P \propto T \text{ at constant volume.}$$

Using this relation and Boyle's Law, the ideal gas "law" or better, equation of state, viz. $PV = RT$, is seen to follow at once, and is ready for use in subsequent comparison with the predictions of elementary kinetic theory. It has traditionally been made the excuse in the classroom and examination room for the setting of surpassingly dull problems on real gases. We would suggest that pupils should certainly know that this equation can be applied with some success to the description of the behaviour of real gases under conditions of moderate pressure and temperature (their own

/laboratpry.....

laboratory experiments should convince them of this) and be induced to think of reasons why departures from "ideal gas behaviour" must occur. As to problems set on this kind of thing, which usually take the form of plugging in five of the values for P_1, V_1, T_1, P_2, V_2, T_2 and calculating the sixth, the fewer of these set the better. "Gas law" calculations are of most use and interest to the chemists, who can be left to handle this matter for themselves.

The fundamental experiments needed to support the above treatment of temperature and the gas laws have been described or mentioned at the appropriate places in the text. We would only add here that the Nuffield experiment concerning what the gas thermometer has to say about the boiling point of air should certainly be done if liquid air is available, but only after pupils have had a gas scale properly defined for them. If time permitted, interest could be added to this section by the introduction of other types of thermometers (e.g. resistance thermometers or thermocouples) which are more accurate than mercury-in-glass thermometers and more convenient to use than gas thermometers.

This completes the recommended treatment of temperature and the gas laws. It is somewhat longer than the usual one, and it cannot be claimed for it that it is at all times exciting. On the other hand temperature is an important fundamental concept, and the treatment of fundamental concepts should not be skimped, even at school level. There is evidence to show that a treatment of this standard of rigour can be successfully presented to senior school pupils. Victor (1968) has described an alternative treatment to the one outlined by us, based on deducing the ideal gas scale from the limiting values of PV products (cf. Sears and Zemansky 1964, Zemansky 1964), and he claims that this is readily understood by his pupils and much easier to make convincing than the traditional approaches.

/5.3.....

5.3 Relation between heat and work

In the Common Basic Syllabus for Physics and Physical Science, an "historical approach" to this topic is recommended. This is acceptable, because such an approach also happens to be a logical one and one likely to prove of interest to pupils.

Processes in which energy is "lost" as a result of friction seem always to be accompanied by the evolution of heat; is there not a connection between energy and heat? It was this question which James Prescott Joule asked himself and which he answered so decisively in the affirmative following his brilliant experimental study of the problem. (We have it on the authority of Tait, incidentally, that "Joule" rhymes with "rule"; see Thomson 1933, Allen 1943.) After appropriate preliminary discussion along these lines, some account could be given of Joule's "paddle-wheel-rotated-by-strings-pulled-by-falling-weights" experiment or of his "honeymoon" experiment in which he measured the difference in temperature between the top and bottom of a waterfall. It is not necessary to spend a lot of time discussing the "caloric" theory of heat that prevailed until the results of Joule's work so effectively demolished it; there is no room in the course for lengthy consideration of theories that are well and truly dead. Nor is it necessary to devote much attention to Rumford's pioneering work, interesting and significant though it was.

Regarding laboratory work in support of this part of the syllabus, we cannot do better than quote Rogers on "Joule Experiments in the Laboratory" (Rogers 1960): "Joule's work was a miracle of careful experimenting, wrought with outstanding instruments. Ordinary experiments are spoiled by heat losses. You should do laboratory experiments on energy conversion, to gain a clearer understanding of Joule's work; but

/hardly.....

hardly to add reliable testimony to the great pile of evidence that he started - your own work is likely to make you sympathize with Joule in his difficulties and admire his skill". Rogers's own suggestion for an experiment in this category is the familiar one in which a cardboard tube containing lead shot is inverted many times and the rise in temperature of the shot determined. Since the purpose of performing such an experiment is "to gain clearer understanding of Joule's work" and nothing else, its obvious crudity is not of serious consequence. It is important for pupils to try to think of all possible sources of error and consider how they could be corrected.

The evidence for the extension of the Principle of Conservation of Energy to include other "forms" of energy besides mechanical is not of the kind that could or should be provided by the results of any single experiment. Historically it was not so; and it should not be so for our pupils either. Somehow, they must be convinced of the validity of extending the Principle in this way, but it would take them and their teachers a substantial fraction of a lifetime to repeat the painstaking experimental work of Joule and others in order to provide their own evidence. The realistic solution to this problem is that which is recommended in Rogers's book (1960) and in the Nuffield Guide (IV): to give pupils a survey of the numerical results of various determinations of "the mechanical equivalent of heat", together with some account of the methods used, and to invite pupils to draw their own conclusions. The chart and data given by Rogers (1960) and the Nuffield Guide (IV) will serve excellently for this purpose.

Care must be taken when assisting pupils to sum up the results of these experiments and arrive at their "verdict".

/The....

The principal points to emerge should be the following:-

(i) Fundamentally, Joule (and his successors) showed that whenever work was done on a system with no other result than the evolution of "heat", the ratio of the amount of work done (measured in mechanical units) to the amount of "heat" developed (measured in now-obsolete units of "heat") was always the same. This ratio has become known as the "Mechanical Equivalent of Heat."

(ii) One may conclude from Joule's results, as did Joule himself, that although in these experiments no kinetic or potential energy was transferred to the given system when work was done on it, energy in some form was nevertheless transferred, manifesting itself by the rise in temperature of the system. Thus the idea of work as a measure of energy transfer is generalized to include cases in which heat is developed as a result of work done: in other words, the Principle of Conservation of Energy is generalized to include "heat" as well as kinetic and potential energy.

(iii) The fundamental result of Joule's experiments may now be re-phrased as follows:

Under given circumstances the same amount of energy is always required to raise the temperature of a given substance by a given amount, no matter how the energy is transferred.

(iv) Up to this point, the pupil will have been accustomed to thinking of "heat" as that which is gained by a body when its temperature rises (or even, in certain circumstances, when its temperature stays the same). "Heat" in this sense is now seen to be a form of energy, and this fact necessitates some revision of terminology. When a system is "heated" by whatever means it gains energy; because this energy is somehow associated with the internal constitution of the system (of kinetic and potential energy in this respect) it is logical to call it internal energy, which is the approved

/modern.....

modern term. The term "heat" is now properly used in a much more restricted sense. "Heat is the energy transfer between two systems that is connected exclusively with the temperature difference between the systems" (Sears and Zemansky 1964). To illustrate: when body A, which is at a higher temperature than body B, is placed in contact with B, A loses internal energy and B gains internal energy until the two bodies attain the same temperature. The energy in transit from A to B in this process is referred to as heat; once the energy has been taken up by B it has become part of B's store of internal energy and it is no longer correct to call it heat.

(v) Since internal energy and heat (in its correct sense) are forms of energy, there is no point in using units for them which are different from those introduced for kinetic and potential energy. In short, there is nothing to be gained by having the calorie in addition to the joule, and this fact is now officially recognized. By a decision of the Ninth General Conference of Weights and Measures in (Paris, 1948 , the official unit of internal energy and heat is the joule. The 15°C calorie is now defined as being equal to 4.1855 joules (cf Evans 1964, Kaye and Laby 1966), and the thermochemical and "International Table" calories are respectively defined as 4.184J and 4.1868J (Mc Glashan 1969).

5.4 The particle structure of matter

5.4.1 Introduction - intermolecular forces

Discussion of the gaslaws and the extension of the Principle of the Conservation of Energy form an appropriate preparation for a more advanced consideration of the particle structure of matter. Pupils at this stage will already have a reasonable understanding of the particle model of matter.

/This.....

This must now be extended to include the making of quantitative predictions from this model and comparing these with experiment.

From their knowledge of the properties of matter in bulk, pupils can easily deduce that if matter consists of tiny particles in ceaseless motion, then these "molecules" must maintain fixed mean positions relative to one another in the solid state and that they must also be packed quite close together; that in the liquid state the "molecules" must, on average, be still much the same distance apart and still cling quite well together but now be able to slide over one another with a certain amount of freedom; and that in the gaseous state the "molecules" must move about with considerable freedom and scarcely exert any influence upon one another except when they collide. Before these ideas are taken much further, pupils should be assisted in guessing the general pattern of variation of the force between a pair of "molecules" with their separation. (We place the word "molecules" within quotation marks, since by it we intend to mean atoms, ions or true molecules - whichever is appropriate to the substance under consideration.) Now by this time pupils may or may not have learnt in chemistry of the different kinds of interionic, interatomic or intermolecular forces - ionic, covalent and metallic bonds and Van der Waals forces. What is important for them here is not a detailed grasp of the nature of these forces - though it is hoped that teachers will not allow pupils to be misled by the suggestion made in a recently-published South African textbook (Macphail 1964) that intermolecular forces are gravitational - but an appreciation of the features that are common to all these types of binding force (cf Helm 1968). These features can to a large extent be inferred from the behaviour of matter in bulk.

/(i).....

(i) Solids and liquids resist <u>compression</u>: this suggests that when their "molecules" are close enough together, they exert strong repulsive forces upon one another. Also, if we seek to explain the pressure exerted by a gas on the walls of its container as the effect of collisions between the molecules of the gas and those of the wall, then we must suppose that during those collisions strong repulsive forces must act between the molecules. We can also reasonably infer that when the molecules of a gas collide with one another, as they surely must, strong repulsive forces act during the collisions.

(ii) Solids may be <u>stretched</u> by the application of forces; so may liquids if care is taken to remove all air bubbles from them. This suggests that molecules exert attractive forces upon one another when separated from one another by an appropriate distance.

(iii) To stretch a solid more and more, we have to apply increasingly bigger stretching forces to it; when these forces are big enough, breakage occurs. This suggests that the <u>attractive</u> force between neighbouring molecules increases with increasing separation up to a certain limiting value, then decreases as the separation is made even bigger. That the "limiting value" cannot be very large in terms of "molecular" diameters is clear from the constancy of the surface tension of thin films as those films are made thinner and thinner (cf Nuffield Guide I). Evidently the "tension" can only come from the two outermost layers of the film. When solids or liquids are <u>uniformly</u> compressed, application of increasing stress simply results in increasing compression; this suggests that as the "molecules" are squeezed closer and closer together, the repulsive force between any pair of them continues to increase indefinitely.

/When.....

When one puts all these conclusions together, the result is the familiar curve of intermolecular force versus intermolecular separation. This shows a negligible force at comparatively great distances, an attractive force at first increasing with diminution of distance, attaining a maximum, then dropping to zero as the intermolecular distance reaches the "equilibrium" value usually denoted by r_o; and then, as the distance decreases yet further, a rapidly increasing repulsive force. One may think of this curve as being the superposition of two curves, one showing the variation of attractive force and the other that of repulsive force with distance, with the repulsive force "winning" at distances less than r_o and the attractive force "winning" at greater distances. Not only is the behaviour shown in the graph what we should expect from macroscopic observations; it is also in accord with our physical intuition. We would not expect the forces to be purely attractive, for if they were it would be difficult to see how the molecules could attain stability of equilibrium. This point is worth plenty of emphasis, though perhaps the pupil is in less of need of it than his teacher might be, who has perhaps been nurtured on a story of intermolecular forces which concentrates only on the attractive forces.

With the aid of these ideas about intermolecular forces it is now possible to embark upon a closer examination of the gaseous, solid and liquid states than was possible earlier.

5.4.2 Simple kinetic theory of gases

Under this heading, pupils and teacher are expected to try their hand at building a simple theory of gaseous behaviour from simple assumptions about molecules. Recently (Warren 1965) a responsible critic has poured scorn upon typical attempts of this kind. He has written: "In view of the

/properties.....

properties which are usually attributed to molecules it is necessary to question the purpose in teaching [elementary kinetic] theory. Perhaps it is the intention to give the student some idea of the properties of molecules and to show how the behaviour of matter in bulk can be explained in terms of the interactions of its constituent particles. If this is the object the conventional approach is completely unsuitable since...the properties that can be attributed to molecules are mostly incorrect. An alternative purpose could be to conduct an intellectual exercise showing how a particular mechanical model can be used to interpret a set phenomena, without regard to the facts that exactly the same results can be obtained with completely different assumptions and that the properties assumed for this model are inconsistent with clearly demonstrable properties of molecules. Such an object might be legitimate, but if it is pursued it is obviously essential to make quite clear to the student that this in fact is the purpose of the analysis."

In the nineteen-sixties, everyone seems to be agreed that there is some virtue in teaching elementary kinetic theory of gases. There may be some confusion as to the purpose of this, as Warren has implied, and certainly it is essential for the intending teacher to be clear about why he is teaching this topic. Let us therefore attempt a clear statement of the reasons why it is thought valuable to teach elementary kinetic theory.

Surely the situation is this. We have "hypothesized" molecules in an early stage of our teaching and have helped our pupils to look at some of the evidence which points to the existence of molecules. We have shown that if there are molecules, it is possible to say a good deal in semi-quantitative terms about the forces between them. Now we want to

/try.....

try to show that we can make quantitative predictions from a simple theory of molecular behaviour based on what we believe we have found out about molecules. As everyone knows, we can go quite far with a simple theory of this kind, obtaining not only Boyle's Law but a formula which enables us to form a good idea of the average speed of molecules and to relate the average molecular kinetic energy to the Kelvin temperature. To be sure, this can only be done at school level by considerable simplification; but if pupils are aware of the nature of the simplifying assumptions and why they are made, this will do no harm. Pupils will have seen use made of the important and essentially correct idea that the pressure of a gas on the walls of its container arises from the impacts of its molecules with the walls, and this is in itself an interesting application of the fundamental principles of mechanics. In sum, we can say that the first of the aims mentioned by Warren is our principal aim: we are attempting to show how the behaviour of matter in bulk can be explained in terms of the interactions of its constituent particles.

But we are also "[conducting] an intellectual exercise showing how a particular mechanical model can be used to interpret a few phenomena", in the sense that any attempt to construct and use a theory is an "intellectual exercise".
If we do these things, keeping our eyes and those of our pupils open as to what we are doing, then we will have strengthened pupils' belief in the reality of molecules and perhaps allowed them to experience some of the intellectual excitement that is attached to the development and testing of a theory. Certainly we should be careful, and we would therefore do well to take note of the warnings of Warren and others about the presentation of this theory.

/First.....

First, we want to be sure about the assumptions we make on which we proceed to construct our theory. In stating these, we should distinguish between those which are fundamental to the theory and those which have to be introduced in order to simplify the treatment sufficiently to make it palatable to pupils. (In treating the problem by means of the techniques of classical mechanics, we are in any case oversimplifying; properly speaking, we ought to apply the methods of quantum statistics to the problem. Fortunately for the progress of our teaching at this point, application of the more fundamental theory leads to the same results as the classical treatment.) The assumptions we must regard as fundamental to the theory are the following:-

(i) A gas is composed of a very large number of small particles called molecules.

(ii) The molecules are in ceaseless, random motion.

(iii) The molecules occpuy negligible volume; that is, the mean distance between molecules is large compared to the linear dimensions of the molecules.

(iv) The range of intermolecular forces is small compared with the mean distance between molecules. (This amounts to saying that the molecules exert negligible attractive forces upon one another, and only exert appreciable repulsive forces upon one another during collisions.)

(v) The duration of intermolecular collisions is short compared with the mean time elapsed between one collision and the next.

(vi) The total translational kinetic energy of the gas molecules remains constant; i.e., it is unaffected by intermolecular collisions or by radiation from the walls of the container.

Some comments are in order concerning this last assumption. We must obviously lay down some condition about the

/total.....

total translational kinetic energy of the molecules of the system under discussion, as whatever is assumed here is likely to affect the predictions of the theory. For example, in terms of this model of a gas, the pressure is likely to depend on some kind of average of the speeds of the molecules that are rushing about in the container, and the behaviour of this average must depend on whether the total translational kinetic energy fluctuates, rises, falls or stays the same. It is clearly simplest to assume that it will stay the same; it is also physically plausible, as we might reasonably expect that the total translational kinetic energy of the molecules of the gas in a constant-volume container that is isolated or in thermal equilibruim with its surroundings will remain constant. In any event we are at liberty to postulate that the gas will be subject to this condition, and to see what conclusions will follow from this. One advantage which assumption (vi) confers is that it enables us to neglect collisions between the gas molecules themselves when we calculate the pressure exerted by gas.

One way of ensuring that (vi) must be satisfied is to specify that the collisions between molecules must be <u>perfectly elastic</u>; in fact the corresponding assumption to (vi) is more often than not stated in this form. In the Nuffield Guide (IV), it is asserted that "Gas molecules at room temperature bounce against each other with perfect elasticity." In elaboration of this, it is remarked that "In an inelastic collision a molecule may be torn apart into separate atoms or...raised to an excited state, or ...ionized; or there may be a loss of energy by radiation. All these are hopelessly improbable events among molecules of gases at ordinary temperatures as we know them." We may agree with these statements if they are considered to apply to monatomic gases only. But diatomic or polyatomic molecules may readily be raised to excited <u>rotational</u> energy states at

"ordinary.....

"ordinary temperatures" by collisions with one another; it is only the excitation of vibrational or electronic energy states by this means that are "hopelessly improbable events" at ordinary temperatures for some gases of this kind, though by no means for all. The evidence for this is provided by the observed specific heats of these gases, which show that their molecules must have internal energy of some sort (rotational kinetic energy, as we now know) in addition to translational kinetic energy. When energy is supplied to a gas at constant volume to raise its temperature, energy is transferred from the walls of the container to the molecules of the gas through collisions between the gas molecules and possibly also by absorption by the gas molecules of energy radiated by the walls (for simplicity, we shall neglect the latter possibility). The gas molecules can share the added energy among themselves by their collisions with one another; and if these molecules are diatomic or polyatomic, some of this energy must go into their store of internal energy, as we have seen. But this means that some at least of the intermolecular collisions must be inelastic. If the temperature of a diatomic or polyatomic gas is kept constant, and inelastic collisions take place, then we must suppose that their effect is balanced by superelastic collisions, i.e. collisions in which there is conversion from internal energy to added translational kinetic energy. (On this point, cf Warren 1965.) In other words, we are assuming (vi) above.

This kind of discussion would only be found bewildering by pupils being introduced to elementary kinetic theory, so that we should, in effect, confine our introductory discussion to monatomic gases by making the simplifying assumption that collisions between the molecules are perfectly

/elastic.

elastic. (It is worth elaborating with pupils the point made above that when energy is added to a monatomic gas, the only way in which the added energy can be shared among the molecules is by molecular collisions; in building our kinetic theory we are trying to provide explanations for just such phenomena as these.) Further, we should assume for convenience that the collisions between the gas molecules and those of the walls are perfectly elastic; this will ensure that the component of momentum of a gas molecule normal to the wall will be reversed by the collision. Again, this is not an essential assumption; all that we really require to prove the desired result is that on average, the same number of molecules leave the wall per unit time with a given normal component of momentum as reach it with this value of the normal component.

It is when we come to consider the velocities of the molecules that we have to make the most drastic simplifications. Pupils will not be able to understand the subtleties of the process of averaging by which the complexities of "molecular chaos" are reduced to simplicity; they need a much simpler approach which will still convey to them the essence of the matter. The best that can be done here is summarized by the Nuffield Guide (IV): "...we are going to pretend that we can divide the molecules into three regiments, one flying along the length [of the box], the other...etc. Although that is very artificial, we had better use it here - for all but an exceptionally bright pupil - because the resolving of velocities into vector components will certainly obscure the whole business." The artificiality will be apparent enough to pupils, who should nevertheless be able to appreciate that the results derived by this approach ought to have some semblance of the truth about them. They may be worried about the effect of intermolecular collisions and how they

/should.....

should be allowed for. (Our initial assumptions took note of the existence of such collisions.) Conservation of momentum takes care of this problem for us, ensuring that the momentum transferred to the walls is unaffected by the momentum exchanged amongst the molecules themselves. If we assume the molecules to be mere point masses, travelling along the three co-ordinate directions only with the same speed and making perfectly elastic collisions with one another, this idea is particularly obvious; the molecules will bounce apart with their velocities interchanged, and because the duration of molecular collisions is assumed to be negligible, the resultant effect will be as though individual molecules simply travelled back and forth unimpeded between opposite walls of the container.

Ultimately, then, we may make the following recommendations for an elementary derivation of $PV = \frac{1}{3} Nm \overline{v^2}$:-

(i) Assume that the molecules make perfectly elastic collisions with one another and with the walls of the container.

(ii) Assume that the container is box-shaped, though not necessarily cubical; although it is not possible at this stage to generalize the result to a container of any shape, we need not restrict the proof to a cubical container when it is so easily extended to a rectangular parallelepiped.

(iii) Assume that all molecules travel with the same speed, and that they are divided into three equal-sized groups, each travelling perpendicular to one pair of walls of the box.

(iv) Assume (though this is a deduction from conservation of momentum rather than an assumption) that intermolecular collisions need not be taken into account.

With these assumptions, the derivation of $PV = \frac{1}{3} Nm \overline{v^2}$ proceeds very simply indeed:-

/Let.....

Let the dimensions of the container be a, b and c.

Consider a molecule, mass m, speed v, travelling between the walls that are separated by a distance a.

This molecule makes $\frac{v}{2a}$ impacts with one of these walls per second.

At each impact it transfers an amount of momentum 2mv to the wall.

If there are N molecules altogether, then the total momentum transferred to this wall per second = $\frac{N}{3}$ 2mv $\frac{v}{2a} = \frac{Nmv^2}{3a}$

This is the force exerted on the wall.

Since the area of the wall is bc, the pressure on the wall must be given by $P = \frac{Nmv^2}{3abc}$

or $PV = \frac{1}{3} Nmv^2$,

since V, the volume of the container, is equal to abc. (The proof can be slightly extended by calculating the number of impacts in a time t and dividing by t at the appropriate stage later in the proof.) According to reports given in the Nuffield Guide (IV), the major difficulties that pupils experience in connection with this proof concern the momentum transferred to the wall by molecular impact and the interpretation of the transfer of momentum by many impacts over a period of time as a force. Of these, the first is reckoned to be the more troublesome. The solution recommended for it is for pupils to try simulating collisions between molecules and wall by collisions between themselves, especially those in which one pupil stands still and another collides with him and bounces back with equal and opposite velocity. This should help to convince pupils that momentum is transferred to the wall- that the wall experiences a force as a result of the collision.

Senior school pupils should be capable of grasping the idea behind the slight modification of the above formula by

/which.....

which it becomes $PV = \frac{1}{3} Nm\overline{v^2}$, where by $\overline{v^2}$ we mean the mean-square-velocity of the molecules. The point can be put as follows:- "In a more rigorous treatment in which we take into account the possibility that the molecules can have (almost) all possible speeds and directions, we arrive at the same formula, except that we replace v^2 with $\overline{v^2}$, where $\overline{v^2}$ means the average value of the squares of all the molecular speeds; i.e. we would calculate $\overline{v^2}$ by taking each molecular speed, squaring it, adding all these squared quantities and then calculating their mean value." In its alternative form $P = \frac{1}{3}\rho\overline{v^2}$, where ρ is the density of the gas, the kinetic-theory formula for pressure is most useful for calculating the root-mean-square-velocity of gas molecules, and pupils should accordingly be familiar with this version of it.

If the order of presentation recommended above is followed, pupils will have become acquainted earlier with the "ideal-gas" equation and will know that at moderate temperatures and pressures the behaviour of real gases is described by it to a good degree of accuracy. It is therefore a simple matter now to identify the expression $\frac{1}{3} Nm\overline{v^2}$ with the right-hand side of the ideal gas equation, viz. "RT", and so obtain the result that $RT = \frac{2}{3} \times$ (total translation kinetic energy of the gas molecules), and hence that the mean translation kinetic energy of the gas molecules is directly proportional to the Kelvin temperature. While this is an interesting and valuable result, it is important that too much should not be made of it (cf our earlier comments on the kinetic-theory "interpretation" of temperature). In particular, both pupils and teachers should note that the proportionality between Kelvin temperature and average molecular kinetic energy of translation is exact only for an ideal gas. Also, appropriate stress should be laid on the fact that for an ideal gas

/or.....

or a real gas under near-ideal conditions, it is between translational kinetic energy and temperature that this simple relationship has been established; the energy associated with the internal motions of the molecules is not associated with their linear momentum and hence plays no part in determining the pressure exerted by the gas.

This work on the elementary kinetic theory of gases should be supplemented both by demonstrations with simple mechanical models which simulate the behaviour of molecules in gases, and by experiments or demonstrations which support some of the additional predictions of the kinetic theory. In the former category, it is now standard practice to use two-dimensional models of gases (and also of liquids and solids) in which ball-bearings or similar objects are shaken around in a tray and collide with another and with the walls of the container. A strikingly successful and indeed beautiful version of this model has been devised by Meiring (1968), whose apparatus is both cheap and easy to make. The tray consists of a perspex platform on which is mounted an aluminium ring; it is agitated by the driver units of a pair of 4" loudspeakers, working in push-pull. With this model, which can easily be displayed with an overhead projector, it is possible to demonstrate a number of the features of molecular behaviour in the gaseous, liquid and solid phases of matter and also the transitions between these phases.

5.4.3 Some predictions of the kinetic theory of gases

Concerning experimental support for the predictions of the theory, it is not enough for us to say: "Our theory predicts that the product of pressure and volume of a gas should be constant at constant temperature. Real gases are observed to behave this way under certain conditions; therefore our theory can be regarded as 'verified'". While it would be

/an.....

an exaggeration to say that our theory was in fact built in such a way as to ensure that it would predict this result, there is an element of truth in this remark. To be able properly to judge the validity of the fundamental ideas of the theory, we need to study some of its new predictions and see whether they find experimental confirmation.

An obvious candidate for this study is provided by the prediction from the formula $P = \frac{1}{3}\rho\overline{v^2}$ of the root-mean-square velocity of the molecules. By using known values of atmospheric pressure and density it is easy to calculate the predicted value of about 500 m/s. Direct measurement of this quantity is impossible in a school laboratory. In standard experiments to measure the distribution of molecular velocities, the time taken by molecules to travel a known distance in vacuo is measured. A simplified account of Zartman's (1931) method is given by Rogers (1960); Halliday and Resnick (1966) describe the more accurate experiment of Miller and Kusch (1955). In principle, these methods are quite simple, but it is doubtful whether a teacher should take time to describe one of them when there is available to him convincing indirect evidence of the high values of molecular speeds. Such evidence is provided by experiments on bromine diffusion and measurements of the speed of sound. The premier advocate of the use of bromine diffusion experiments in this connection is Professor Eric Rogers (1960,1962); certainly no-one who has seen him perform these experiments and noted their impact on "lay" members of an audience can fail to be convinced of their immense effectiveness. As performed by Rogers and recommended by the Nuffield Guide (IV), the bromine should first be allowed to diffuse into air, which it does very slowly - no obvious evidence of high molecular speeds here! - and then, upon suggestion from pupils, into a

/vacuum.

vacuum. The latter step cannot give us a numerical estimate of molecular speed, but it does demonstrate most dramatically that molecules must travel at speeds far greater than "ordinary". (The technique of performing these experiments and the precautions that must be taken with a "dangerous" substance like bromine are fully described in the appropriate section of the Nuffield Guide (IV).) For numerical data on molecular speeds we turn to measurements on the speed of sound. The relevant experiments have been described in Section 3.18.1 of this thesis, and the argument by which we go from the speed of sound to an estimate of average molecular speed was outlined in Section 3.18.2. It can now be urged that if this argument and elementary kinetic theory are valid, we should find that the speed of sound is lower in a gas of high density than in a gas of low density at the same temperature; and that at constant temperature, the speed of sound should be independent of the pressure. One can point out that both these predictions are confirmed by experiment; unfortunately one will not be able conveniently to conduct the relevant experiments at school.

It is possible in elementary physics teaching to explore other predictions of the kinetic theory, as Rogers and the Nuffield Physics Course (cf Nuffield Guide IV) have shown. Appropriately analysed, the data obtained from the experiment in which bromine diffuses into <u>air</u> can be used to yield an estimate of molecular mean free path. If the reasoning is carried somewhat further, an estimate of molecular size can be obtained, and compared with that obtained from experiments on thin liquid films. The chain of reasoning that has to be used to establish these results is essentially straightforward but, even with the "short cuts" suggested (cf Nuffield Guide IV), it is still quite long, and it will take a very capable

/teacher.....

teacher to be able both to carry his pupils with him to the end of it and to leave them with some feeling of enthusiasm for the results obtained. That this can be done there is no doubt; nor is there any doubt that, in the right hands, it is worth doing. Here is some good, solid physical reasoning from a modern (or fairly modern) theory, used to derive results which are of interest. The whole exercise offers an excellent introduction to the way in which physicists have to argue from their data to reach significant conclusions. The matter is discussed in great detail in the appropriate section of the Nuffield Guide (IV), where the following advice is offered to those intending to follow this argument through with their pupils: "Whichever method the teacher chooses, he will find that young pupils <u>can</u> follow the argument if he can succeed in giving them a glimpse of where they are going and a rough idea of how they are going to get there, because they will then want to follow the detailed story. (As practice for that, teachers will find that trying the argument out on another adult is very helpful.)"

The Common Basic Syllabusesin Physics and Physical Science do not at present specify that estimates of molecular mean free path and size from elementary kinetic theory should be discussed with pupils. We think that the case for the inclusion of this topic is strong, but that those who would have to teach it would first have to have it taught to them by extremely competent teacners. In the present state of science teaching in South Africa, it would be utterly unrealistic to attempt to introduce this topic without intensive preparation for the teachers concerned.

5.4.4 <u>"Heat" and temperature</u>

If teachers follow the order and method in which we have recommended that "temperature" and "heat" be taught, there

/should.....

should be little room for confusion of these concepts and little need to spend time on trying to establish their different nature. The study of the elementary kinetic theory of gases provides an opportunity for distinguishing between these concepts at the microscopic level; or rather, since pupils are now supposed to know the difference between "heat" (in its correct sense) and "internal energy", we can say that kinetic theory enables one to interpret the internal energy of a gas in kinetic terms and to contrast this interpretation with that of temperature, which has already been studied. For an (ideal) monatomic gas, the internal energy is simply the <u>total</u> translational kinetic energy of all the molecules; for an (ideal) polyatomic gas, the internal energy is the total translational kinetic energy of all the molecules <u>plus</u> the total energy, kinetic plus potential, associated with the internal motions (rotational and vibrational) of the molecules. (The wording of the Common Basic Syllabuses in relation to this point is somewhat unhappy:"relation of heat to energy of molecules.") This is a simple idea which should meet with no difficulty in presentation. The concept of the "heat capacity" of a gas follows directly from this, though in a school syllabus it need not and cannot be taken very far. The following are important points to make or bear in mind:

(i) The term "heat capacity" means the amount of <u>energy</u> needed to change the temperature of the given mass of gas by unit amount. This energy is often but not necessarily supplied to the gas by heating.

(ii) Since gaseous volume depends so markedly on the temperature and pressure, care must be taken to specify the conditions under which the "heat capacity" and "specific heat capacity" of a gas are measured. For convenience of specification, the conditions of "constant volume" or "constant pressure" are chosen. Heat capacity at constant pressure

/must.....

must exceed that of constant volume, since under the former condition the gas can expand and do work on (transfer energy to) the agent supplying the pressure, and energy has to be supplied to it for this purpose as well as to increase the internal energy of the given mass of gas.

There is nothing to be gained from a derivation of the relation between the principal heat capacities of an ideal gas, since there would be no opportunity for teacher and pupils to make use of this result in a South African school science course.

(xii) It is a simple exercise to use the ideas of elementary kinetic theory to derive an expression for the heat capacity of one mole or kilomole of an ideal monatomic gas. This should be done, and the result compared with that obtained from measurements on real monatomic gases, as a simple illustration of the power of this elementary form of kinetic theory. To take the matter further to include diatomic or polyatomic gases would require the use of the principle of equipartition of energy and the concept of "degrees of freedom": the one could not be adequately justified at this level and the other is a shade too sophisticated for comfort. The syllabus says nothing on this point one way or the other; we think that advantage should be taken of this silence to recommend that no attempt should be made to develop expressions for the thermal capacities of diatomic or polyatomic gases, interesting though the topic undoubtedly is.

5.4.5 Non-ideal gases: introduction

The Common Basic Syllabuses also list "non-ideal gases", but do not specify which aspects of the behaviour of real gases are expected to be treated. Obviously it is important to discuss with pupils the departures of the behaviour of real gases from the predictions of the kinetic theory of ideal

/gases.

gases. Perhaps it would be best to initiate such a discussion with a consideration of what might be some of the weak points of the theory, notably the neglect of any attractive forces between the molecules and the assumption of negligible molecular size. In gases at high density, effects associated with the long-range attractive forces between molecules and with the finiteness of molecular volume might lead to significant departures from the predictions of the simple theory. This much is fairly obvious: what is less obvious is how the necessary modifications are to be made clear to pupils, if at all. The essence of our recommendation concerning this problem is that no attempt should be made to go into detail, and certainly not to state Van der Waals's equation of state and to account for the extra terms in it.

By way of furnishing the main argument for this recommendation, we shall consider what advanced kinetic theory and statistical mechanics have to say about non-ideal gases, and then go on to outline and discuss some of the elementary treatments of Van der Waals's equation that appear in standard texts.

5.4.6 Advanced treatments of non-ideal gases

The establishment of an (approximate) equation of state for an imperfect gas consists in the derivation of expressions for the so-called "virial coefficients", i.e. the coefficients of the successive powers of $\frac{1}{V}$ which appear in the "virial" expansion $\frac{PV}{RT} = 1 + \frac{B}{V} + \frac{C}{V^2} + \ldots$. Of these, it is the coefficient of $\frac{1}{V}$, i.e. "B" in the virial expansion (the "second virial coefficient") which is naturally the most important. There are various methods of obtaining the required expression for "B", all of which start from the assumption that the interaction between a pair of molecules can be described by a

/potential.....

potential function ϕ which depends only on the intermolecular separation r.

In what is perhaps the simplest of these methods in principle, the pressure in the interior of the gas is treated as being the sum of two contributions, viz. the kinetic pressure, which arises from the transfer of momentum by actual movement of molecules across a given plane in the gas, and the static pressure, which arises from the interaction between molecules on one side of this plane and those on the other. An expression is found for the static pressure in terms of the intermolecular potential function ϕ (cf. Fowler 1929, Present 1958). The static pressure p_s is found to bear a simple relation to "B", viz. $B = \frac{V p_s}{RT}$, if simultaneous interactions of a molecule with more than one other molecule are ignored. In a somewhat more sophisticated approach to the problem, Boltzmann's "e-theorem" (probability of a molecule having energy E is proportional to $e^{-E/kT}$) is used to derive a relation between the molecular concentration near the walls of the container and that prevailing in the interior of the gas (Sauter 1949; see also Fowler 1929). The measured pressure at the walls is the kinetic pressure, and is given by the product of the molecular concentration there (n_w) with kT (k is Boltzmann's constant), so that if n_w is found, it is possible to relate it to "B".

More sophisticated still is the approach which makes use of the Virial Theorem of Clausius (cf Fowler 1929, Wannier 1966). According to this theorem, the "pV" product for a gas may be set equal to RT minus a term equal to two-thirds of a quantity sometimes called the "internal virial", which is in turn given by the expression $-\frac{1}{2} \sum rF(r)$, where r is the intermolecular separation and F(r) the intermolecular force, and the summation is carried out over all pairs of

/molecules.

molecules. Substitution of $-\frac{\partial\phi}{\partial r}$ for F(r) leads to the full equation of state from which "B" may be extracted, if required. Yet another way of obtaining "B" is to find the assembly partition function of the gas, given by $Q = \sum_j \Omega_j e^{-Ej/kT}$, where Ω_j denotes the number of assembly eigenfunctions associated with the energy Ej, and to derive the equation of state from it by means of the relation $p = kT\left(\frac{\partial}{\partial V}(\ln Q)\right)_T$ (cf Fowler 1929, Band 1955).

All these approaches agree in yielding the following expression for "B", viz.

$$B = 2\pi N \int_0^\infty r^2 \left(1 - e^{-\phi(r)/kT}\right) dr$$

In order to relate this result to Van der Waals's equation of state, $\phi(r)$ is assumed to take the form appropriate to a gas composed of weakly-attracting rigid-sphere molecules. That is, $\phi(r)$ is set equal to infinity for $0 \leqslant r \leqslant d$, where d is the diameter of a rigid-sphere molecule, and is allowed to take on some specific functional form appropriate to an attractive force for $r > d$. When this substitution for $\phi(r)$ is made in the above expression for B, the result is an expression for $\frac{pV}{RT}$ of the form

$$\frac{pV}{RT} = 1 + \frac{b}{V} - \frac{a/RT}{V},$$

where $$a = 2\pi N^2 kT \int_d^\infty r^2 \left(e^{-\phi(r)/kT} - 1\right) dr$$

and $$b = \frac{2}{3}\pi N d^3,$$

or $$\left(p + \frac{a}{V^2}\right)\left(\frac{V}{1+\frac{b}{V}}\right) = RT$$

This reduces to Van der Waals's equation of state if b is small compared with V. Wannier's (1966) comments on this result are worth quoting: "... the nature of the coefficients a and b is just as Van der Waals predicted. The quantity a is a constant which takes into account the long-range attraction of the molecules, while b takes into account their hard-

/core.....

core repulsion; it is a low multiple of the close-packed volume of the molecules. Thus the Van der Waals equation is based on a considerable amount of imaginative insight which can partially be justified."

It will be noticed that the emergence of the term $\frac{"a"}{V^2}$ in the above form of the equation of state does not depend on the assumption of a particular form of $\phi(r)$ for $r > d$. However, the integral in which ϕ appears must converge and this means that if ϕ takes the form of a constant multiplied by r^n, n must be greater than 3. If ϕ is very much smaller than kT, "a" does not depend on the temperature; in general, however, "a" will depend on the temperature, the form of the temperature dependence being determined by the assumed form of $\phi(r)$.

5.4.7 Elementary treatments of non-ideal gases

We now turn to a survey of some of the elementary treatments of Van der Waals's equation that appear in elementary texts. For reasons which will be clear later, we shall be primerily concerned with the $\frac{"a"}{V^2}$ term in the equation, but we will include references to the term representing the finite size of the molecules where we think this to be appropriate.

(i) The standard, "classical" treatment (cf Jeans 1940, Roberts and Miller 1960). The attractive forces on any molecule in the interior of the gas have an average resultant equal to zero if averaged over a sufficient interval of time. Therefore the pressure of the gas at "internal points" is not affected by the presence of these forces. On the other hand, molecules near the walls of the container will effectively be in a permanent field of force exerting a steady inward force on the outermost layer of molecules of the gas. Therefore molecules reaching the wall have their momentum changed not only by their impacts with the wall but also by the effect of the steady inward force. To take this into

/account,.....

account, an extra term must be added to the measured pressure in the equation of state to yield a "total pressure" or "effective pressure." This extra term must be proportional to the number of molecules per unit area in the outermost molecular layer and to the intensity of the inward field of force. Both these factors are proportional to the density, or inversely proportional to the volume of the gas, so that the extra term can be written $\frac{"a"}{V2}$ and the "pressure term" in the equation of state becomes $(p + \frac{a}{V2})$.

(ii) Cowling (1950) gives basically the same account as does Jeans but, unlike him, does not assert that the "internal pressure" of the gas is unaffected by the intermolecular attractive forces. Instead he draws attention to the "static pressure" (see above) contributed by the mutual attraction of molecules lying on opposite sides of an imaginary plane in the gas, and shows that the effect of this is to reduce the internal pressure sufficiently to make it equal to the boundary pressure. In his discussion of the effect of the finite size of the gas molecules on the equation of state, he suggests that this would lead (other things being equal) to an increased boundary pressure, and that the internal pressure would correspondingly be increased as a result of momentum transferred by collisions between molecules lying on opposite sides of an internal surface.

(iii) (Rogers 1960, Nuffield Guide IV). The mechanism of the reduction of the pressure by intermolecular attractions below its "elementary kinetic theory" value is suggested to be a "weeding-out" of some of the slower molecules by the mutual molecular attractions, so that these slower molecules never hit the walls. "Thus there is a <u>less dense</u> layer near the walls, and that exerts a smaller pressure" (Nuffield Guide IV). (Rogers argues that the "weeding out" of slow molecules takes place "in collisions"; the frequency of

/molecular.....

molecular collisions is proportional to the square of the density, so that the correction term to the pressure must be proportional to the square of the density, or inversely proportional to the square of the volume.)

(iv) Born(1946): "If there are forces of cohesion acting between the molecules, one element of volume acts on another of equal size with a force which is proportional to n^2, where n is the number of molecules per cubic centimetre. The pressure exerted by the gas on an external body is correspondingly less than it would be if there were no forces of cohesion." This leads once again to the addition of a term $\frac{a}{v^2}$ to the measured pressure.

(v) (Halliday and Resnick 1966): This treatment starts off in much the same way as (iv), arguing that the effect of the attractive forces varies inversely as the square of the volume. These authors then go on to argue that, as a result, "the gas should, for a given external pressure, occupy a volume less than the volume it would occupy as an ideal gas, in which there are no such attractive forces. Or, equivalently, the gas acts as though it were subject to a pressure in excess of the externally applied pressure. This excess pressure is proportional to $\frac{1}{v^2}$, or equal to $\frac{a}{v^2}$, where a is a constant."

(vi) We have referred above to Sauter's (1949) derivation of Van der Waals's equation of state, which is concerned essentially with the calculation of the boundary pressure by finding the molecular concentration near the walls. In applying this method to the derivation of the correction term for the finite size of the molecules, Sauter makes use of the principle that in the absence of intermolecular attractions, the probability of finding a molecule within a given region of the container is proportional to the unobstructed volume in

/that.....

that region. The effect of the wall is to make the unobstructed volume (per unit volume) larger in its vicinity than it is in the interior of the gas. Thus on this account alone the molecular concentration near the walls should be greater than in the interior of the gas; Sauter shows that the effect of this is to introduce a correction term for the finite volume of the molecules (Van der Waals's "b") which is equal to four times the total molecular volume, i.e. the value predicted by more sophisticated treatments.

(vii) Sommerfeld (1964) has described a much-simplified account of Sauter's treatment of the derivation of the $\frac{"a"}{V^2}$ correction term to the pressure. Starting from the existence of an inwards-directed field of force near the walls of the container, arising from the intermolecular attractive forces, he suggests that the effect of this must be to cause "a kind of sedimentation....away from the wall" which must lead to a lower molecular concentration near the walls (cf (iii) above). By elaborating quantitatively on this idea, he obtains an expression for the molecular concentration near the walls in terms of that in the interior, which is essentially that given by more advanced treatments.

5.4.8 Critique of elementary treatments

It is apparent that it is not possible successfully to reconcile all these ~~three~~ different elementary or near-elementary ways of treating the problem with one another. In treatments (i) and (ii), it is implied that the inwards-directed field of force near the walls leads to a general lowering of the momentum of the molecules incident on the walls. (Cowling (op.cit.) says explicitly that "..as [a molecule] approaches the boundary it is slowed down by the backward pull of the others; it strikes the boundary so much the less violently, and the boundary pressure is thereby

/reduced".

reduced.") In treatments (iii) and (vii), it is suggested instead that the effect of this inward field of force is to reduce the molecular concentration near the walls. In version (i) of the standard treatment, the pressure is said to be unaffected at internal points and to be equal to the simple kinetic theory value, whereas in version (ii) it is claimed that the boundary and internal pressures are the same. On the other hand, (iv) and (v) are readily seen to be equivalent.

The reason for preferring the approach of (iii) and (vii) to that of (i) and (ii) is succintly stated in the Nuffield Guide (IV): "The usual explanation - that all bombarding molecules are slowed by that attraction [i.e. the inwards field of force] and therefore exert less pressure - is misleading because it would suggest a cooler layer of gas near the walls." The latter state of affairs is clearly impossible. Indeed, once we acknowledge that the gas as a whole must be characterized by a uniform temperature, it is plain that the existence of an inwards-directed field of force near the boundaries must be to produce the "sedimentation" referred to by Sommerfeld (op.cit). For the problem is essentially the same in principle as that of a gas at uniform temperature in a gravitational field; as is well known, the molecular concentration in such a gas decreases exponentially with increasing height. Calculation of the molecular concentration at the walls of the container by sophisticated methods (cf Fowler 1929, Sauter 1949) supports the prediction that the gas must be less dense near the walls.

If Jeans's view that "the pressure of the gas at internal points is unaffected" is taken literally, it implies that

/a.....

a layer of gas next to a wall would not be in equilibrium; the force on the side of the layer next to the wall would be less than the oppositely-directed force on the other side, and the layer would be accelerated outwards (cf Warren 1965). But perhaps Jeans simply meant that the so-called "kinetic pressure" - the pressure arising from the momentum transferred across an imaginary plane in the gas by actual molecular movement - is not affected by the intermolecular attractive forces; this is correct. In any event, rigorous calculation of the "internal pressure" (Fowler 1929, Present 1958) shows that it is indeed equal to the boundary pressure.

It will have been noticed that none of these treatments take into account any possible effects arising out of the attractive forces between molecules of the walls of the container and those of the gas. This aspect of the problem will be discussed in Section 5.4.10.

5.4.9 Non-ideal gases and thermodynamics

So far, our survey of advanced and elementary treatments of non-ideal gases suggests that by proceeding carefully enough, one can give elementary discussions of the behaviour of such gases which would be correct as far as they went; in particular, one can produce plausible arguments for the existence of an $\frac{"a"}{V^2}$ correction term to the pressure, which do not depend on the assumption of a particular law of attraction between molecules. According to Warren (1965), however, such approaches are misleading. Following certain physical chemists (e.g. Partington 1949 and Moelwyn-Hughes 1961a), he asserts that "....it can be shown thermodynamically that the power of V in the equation of state is in fact dependent upon the law of force". Now one might think it strange to talk about "the power of V" in reference to the pressure term in the equation of state. Without even considering Van der Waals's modifications to the equation of

/state,....

state, one might expect that the general expression for the pressure of a real gas kept at constant temperature will contain a dominating term of the form $\frac{\text{constant}}{V}$, with progressively smaller "correction terms" proportional to $\frac{1}{V^2}$, $\frac{1}{V^3}$, ... It would be surprising indeed if it turned out that the correction to the pressure should be characterized by a particular power of V^{-1}. At any rate, the argument quoted by Warren in support of his assertion has received eminent backing, and so we reproduce it below (note that the assumption is implicitly made in this treatment that Van der Waals's "a" and "b" are independent of temperature):-

Let U be the internal energy, S the entropy and V the volume of a certain mass of gas at temperature T.

(i) From the laws of thermodynamics, it follows that

$$\left(\frac{\partial U}{\partial V}\right)_T = T\left(\frac{\partial S}{\partial V}\right)_T - p$$

(ii) According to one of Maxwell's thermodynamic relations,

$$\left(\frac{\partial S}{\partial V}\right)_T = \left(\frac{\partial p}{\partial T}\right)_V ,$$

so that $$\left(\frac{\partial U}{\partial V}\right)_T = T\left(\frac{\partial p}{\partial T}\right)_V - p$$

(iii) For a gas conforming to Van der Waals's equation of state,

$$p = \frac{RT}{V-b} - \frac{a}{V^2}$$

therefore $$T\left(\frac{\partial p}{\partial T}\right)_V = \frac{RT}{V-b}$$ and, using the result in (ii) above, we have $$\left(\frac{\partial U}{\partial V}\right)_T = \frac{a}{V^2}$$

(This result also applies to a gas described by the equation of state which we encountered above in our consideration of advanced treatments of equations of state, viz.

$$pV = RT\left(1 + \frac{b - \frac{a}{RT}}{V}\right)$$

/=

$$= RT\left(1 + \frac{b}{V}\right) - \frac{a}{V} \quad ,$$

which reduces to Van der Waals's equation of state when "b" is very much less than V.

For in this case $T\left(\frac{\partial p}{\partial T}\right)_V = \frac{RT}{V}\left(1+\frac{b}{v}\right)$

$= p + \frac{a}{V^2}$ as before.)

(iv) "As the volume V must be proportional to the cube of the intermolecular distance, r, this equation gives

$$\left(\frac{\partial U}{\partial r}\right)_T \propto \frac{1}{r^4}$$ " (Warren 1965).

Hence the volume-dependent part of the internal energy of the gas must be inversely proportional to the cube of the intermolecular distance, from which it follows that $\phi(r)$, the intermolecular potential energy function, must take the form $\phi(r) = Cr^{-3}$

Now if this argument were valid, we should be presented with a disturbing contradiction. On the one hand, we have a supposed deduction from thermodynamics that if a gas obeys Van der Waals's equation of state or the equation $pV = RT\left(1 + \frac{b}{V}\right) - \frac{a}{V}$, the intermolecular potential energy must fall off inversely as the cube of the intermolecular distance. On the other hand, we have a variety of arguments from advanced statistical mechanics and kinetic theory in support of the conclusions that

(a) the equation of state for a nonideal gas that is not too dense should take the form $pV = RT\left(1 + \frac{b}{V}\right) - \frac{a}{V}$, where "a" can be evaluated from a knowledge of $\phi(r)$, and "b" is four times the total molecular volume;

(b) the above result does <u>not</u> depend on the form of the intermolecular potential energy function for r>d, the molecular diameter, provided that, if $\phi(r) = Cr^{-n}$, n must be <u>greater than</u> 3;

(c) if $\phi(r)$ is comparable with kT, "a" will depend on <u>temperature</u>

/in.....

in a manner which will be determined in part by the actual form of $\phi(r)$.

As Fowler (1922) puts it: "It is still frequently stated that if Van der Waals's 'a' is to be independent of the volume, the law of attraction [i.e. of force] must be the inverse 4th power. This is quite untrue...the law of variation with the distance affects and only affects the form of 'a' as a function of the temperature."

Partington (1949) and Moelwyn-Hughes (1961a) give both the above thermodynamical argument and the results of the statistical treatment of the problem; evidently they accept them both and are not aware of any contradiction. However, it is difficult to see how both views of the situation can be accepted without modification of at least one of them.

The statistical-mechanical arguments have no obvious fault, so one turns to the thermodynamical argument in an attempt to detect any flaws. The logic in the sequence of steps (i), (ii) and (iii) is unimpeachable, so that one must accept the prediction that $\left(\frac{\partial U}{\partial V}\right)_T = \frac{a}{V^2}$ for a Van der Waals gas, and hence that the volume-dependent part of the internal energy is inversely proportional to the volume at constant temperature. Indeed, direct calculation of the volume-dependent internal energy of a nonideal gas by the methods of statistical mechanics (cf Fowler 1929, Reif 1965) confirms this conclusion, and shows it to be independent of the actual form of the potential energy function, subject to the restriction that n must be > 3 if $\phi(r) = Cr^{-n}$. Hence it is to step (iv) that we must turn if we seek to find a fallacy in the thermodynamic argument.

In step (iv), the fallacy is not far to seek. It consists in the tacit assumption that if the <u>average</u> intermolecular potential energy per molecular pair is a certain function of the <u>average</u> intermolecular distance, then the true mutual potential energy of a pair of molecules must be given

/by.....

by the same function of their separation; i.e. if $\phi_{average} = f(r_{average})$, then $\phi = f(r)$. Now this kind of assumption is generally quite unwarranted in statistical computations. For example, no one would dare suggest that there should be the same functional relation between the average translational kinetic energy and the average velocity of a collection of gas molecules as exists between the translational kinetic energy and velocity of any given molecule. One can think of no good reason why the corresponding assumption should turn out to be justified for the mutual potential energy of gas molecules. In view of the firmness with which step (iv) has been asserted by the authorities cited above, one hesitates to accuse them of failure to detect so simple a fallacy as this; yet the conclusion that they have unwittingly absorbed this fallacy into their thinking seems inescapable.

There have been other treatments of the problem by physical chemists who have sought to show without reference to thermodynamics but by direct consideration of intermolecular forces that an $\frac{"a"}{V^2}$ correction term to the pressure depends on an inverse-cube law of variation of the potential energy with distance (e.g. Mellor 1902, Kleeman 1920, Porter 1928, Moelwyn-Hughes 1961b). Of these, Porter's may be taken as typical. He considers the attraction of one portion of the gas on the other, assuming a law of attractive force between molecules of the form $F = Cr^{-n}$. He asserts that if the volume V of the gas changes, "the force per unit area across the interface due to all pairs changes as $V^{-\frac{n+2}{3}}$". Therefore if the attractive force per unit area is found to change as V^{-2}, it follows that n must be equal to 4, and the potential energy must vary as the third power of the intermolecular distance. The key idea here is contained in the quoted statement, and it is readily seen that its validity

/depends.....

depends on the assumption that if the attractive force between a pair of molecules is given by $F = Cr^{-n}$, the resultant attractive force F_{total} between all possible pairs is given by $F_{total} = C' r^{-n}_{average}$; in other words, this is the same kind of fallacious assumption as was made in the thermodynamic "proof" that n must be 4 if the pressure correction term has the form $\frac{c}{V^2}$.

If this assumption is dropped, all difficulties cease in regard to the reconciliation between thermodynamics and statistical mechanics. In so far as the two overlap, they agree, and no more need be said.

5.4.10 The effect of the walls of the container on the behaviour of the enclosed gases

Before we pass on to recommendations for elementary teaching, there is a further problem to be considered in connection with the equation of state of a gas, viz. the absence of influence on it of the material of the walls of the container. The existence of attractive forces between the molecules of the walls and those of the gas evidently has no effect on the equation of state. It is customary to ascribe this fact to the operation of the Principle of Conservation of Momentum (cf Roberts and Miller 1960, Tabor 1965, 1969); it is pointed out that a given molecule entering the attractive field of force of the molecules of the wall with a given velocity would transfer as much momentum to the wall during the process of approaching and receding from the wall as it would in the absence of this attractive field. This is true, but we cannot conclude from this fact alone that the pressure of the gas will be unaffected by the field of force of the walls. The pressure of the gas is determined not only by the momentum transferred by individual

/molecules.....

molecules but also by the number of molecules per unit volume in the vicinity of the wall. The existence of a field of force near the walls must result in a concentration gradient near the walls, and this could conceivably affect the pressure.

The problem has been treated in detail by Sauter (1949). He bases his approach on the formula

$$p = \int_0^X n_x \left(- \frac{d\Psi}{dx} \right) dx$$

for the pressure of the gas, where n_x is the number of molecules per unit volume at a distance x from the wall and Ψ is the potential energy of a gas molecule at this distance in the force field of the wall. The formula expresses the fact that the pressure on the wall is equal to the net force exerted on the wall by all the molecules contained in a cylinder of unit cross-sectional area and height X mounted normal to the wall (X is the effective range of the field of force of the wall). Sauter first applies this formula to the calculation of the pressure of a gas between whose molecules there are no attractive forces. In this case n_x is given by the well-known "barometric" formula:

$$n_x = ne^{-\Psi/kT},$$

where n denotes the molecular concentration where Ψ is zero. Substitution of this expression in the formula for pressure yields

$$p = nkT \left[e^{-\Psi/kT} \right]_0^x$$

If Ψ is set equal to (plus) infinity where x is zero and equal to zero where $x = X$, this result reduces to the familiar expression for the pressure of an ideal gas, viz. $p = nkT$, where n is the concentration in the interior of the gas. This result is so simple that one wonders whether it could not be arrived at in some alternative, simpler way to that given by Sauter. In fact, this is the case. We can make

/use.....

use of the fact, previously remarked upon, that the "total pressure" ("kinetic" plus "static") should be the same everywhere within the gas. In the middle of the gas, out of range of the force field of the walls, the pressure is entirely "kinetic" and is given simply by nkT, where n is the molecular concentration in the interior. But this must be the same as the boundary pressure, so that the latter must also be equal to nkT.

Two points in particular are worth noting here. First, the assumption that Ψ tends to infinity as x tends to zero is in accordance with the general view that intermolecular forces become strongly repulsive at very short distances between molecules. Were it not for this assumption, we should obtain a net inwards force of the gas molecules on the wall, which would be an embarrassing state of affairs. Second, the equation $p = nkT$ can only be considered to be precisely equivalent to the equation of state for an ideal gas if n, the molecular concentration where Ψ vanishes, can be considered to be essentially the same as the mean molecular concentration; and this in turn can only be true if the force field of the molecules of the wall is not strong enough and does not extend far enough to produce appreciable changes in molecular concentration anywhere except in the immediate neighbourhood of the walls. Thus Warren (1965), whose own discussion of this point is brief for reasons of space, is evidently justified in claiming that "The reason why we do not have to consider the properties of the walls is that the range of the intermolecular forces is in every case very short compared with the dimensions of the container."

Sauter (op.cit.) also shows by appropriate substitution in his general formula for the boundary pressure of the gas, that the equation of state of a non-ideal gas is similarly unaffected by the force-field from the walls provided that

/the.....

the average molecular concentration in the gas is not sensibly affected by the presence of this force field.

(It is of some interest to compare this problem with one which is in some ways similar, viz. the case of a gas confined only by a gravitational field, such as the earth's atmosphere. Let us suppose that we may treat the atmosphere as an ideal gas and that any attractive forces exerted on its molecules by the walls of any manometer that may be placed in contact with it are entirely negligible. At any point in the atmosphere there will be both a "kinetic pressure", given as usual by $p = nkT$, and a "static pressure", contributed in this case by the gravitational attraction of the earth on the molecules of the gas. It is obvious that the static pressure is a stress or traction equal in magnitude and opposite in effect to the kinetic pressure, giving a "total pressure" in the sense in which we have been using the term equal to zero everywhere in the gas. This does not conflict with the fact that the measured pressure is obviously not zero everywhere; the measured pressure is the kinetic pressure in this case. For an ideal gas confined in a container whose walls exert non-negligible attractive forces on the gas molecules, the measured pressure is not just the kinetic pressure, since the force exerted on the walls by the molecules results not only from momentum transferred in molecular impacts, but also from attractive forces exerted by molecules some distance away from the walls. Here the measured pressure is in fact the total pressure. The essential difference between the two situations is that in one of them, the agent responsible for the "static pressure" lies outside the system whereas in the other it is part of it.)

/5.4.11 .....

5.4.11 Recommendations for an elementary treatment of non-ideal gases

The above discussion has shown that contradictory views have been expressed about the theory of non-ideal gases by some eminent authorities, and that some of the statements commonly made in elementary treatments of the subject are open to challenge. It has also shown that, generally speaking, the topic is far from elementary and should be approached with caution at the elementary level. We propose that the following points should be made in fulfilling the requirements of the syllabuses that non-ideal gases should be discussed:-

(i) Both the very-short-range repulsive intermolecular forces and the longer-range attractive ones can be expected to cause deviations from the simple predictions of the elementary kinetic theory of gases.

(ii) The existence of the repulsive forces has the effect of reducing the volume in which the molecules are free to move around. In effect, this is the correction for the finite size of molecules; the repulsive forces prevent the centre-to-centre distance between molecules from becoming infinitesimally small during a collision. The "volume" term in the equation of state should therefore be equal to the volume occupied by the gas minus a small correction term.

(iii) The existence of the longer-range attractive forces has the effect of bringing the molecules closer together in the same way as an applied pressure would. We should therefore expect that a positive correction must be added to the measured pressure in the equation of state.

Beyond this, it is neither necessary nor desirable to go in a school syllabus of the type at present under discussion (cf. Section 5.4.5).

/5.4.12.....

5.4.12 Dalton's law of partial pressures

This topic is also specified in the Common Basic Syllabuses. This Law is a straightforward deduction from elementary kinetic theory if it is assumed that in a mixture of gases which do not react chemically with one another, each species of molecule acts independently; i.e. each can be treated as though the other were not present (cf Sears and Zemansky 1964). It follows at once from this assumption that each species of molecule exerts a "partial pressure" given by an expression of the form $p_i = \frac{1}{3} \frac{N_i}{V} m \overline{v^2}$, ($N_i$ is the number of molecules of species i in the container), i.e. each exerts a partial pressure which would be the same at a given temperature as the actual pressure it would exert if it alone filled the container. It further follows that the total pressure of the mixture of gases will be the sum of the partial pressures - this is Dalton's Law. No immediate applications of this Law suggest themselves in the Common Basic Syllabus for Physics, which does not, for example, deal with the quantitative aspects of mixtures of gases and vapours, saturated or unsaturated. But the interested teacher might like to take the derivation of Dalton's Law a step or two further in order to show that in a gaseous mixture, the average translational kinetic energy of the molecules is the same for both species. (This, as Sears and Zemansky (1964) point out, is a simple illustration of the Principle of Equipartition of Energy. We do not, however, suggest that this point should be mentioned.) This means that molecules having larger mass would have smaller average speeds.

5.4.13 Conversion of "heat" to work

At first it may seem strange that we should consider this topic at this point. That we should consider it at all arises from its inclusion in the Common Basic Syllabus in

/Physics.....

Physics and for its intrinsic interest for twentieth-century pupils. We have placed it here because we believe that it would be a mistake to introduce pupils to the content of the Second Law of Thermodynamics via a purely macroscopic approach. Once some familiarity has been gained with the kinetic picture of matter, it is possible to give some account of the Second Law in microscopic terms. If it is thought unsatisfactory to defer consideration of this topic altogether until after quantitative kinetic theory has been dealt with, a compromise approach would be to turn to the Second Law immediately after the First Law (conversion of work to "heat"), with the promise that further clarification will follow later.

To introduce this section, one might point out that it is certainly possible to "convert heat to work"; much present-day engineering design and practice takes advantage of this possibility. However, it is a matter of experimental fact that it is <u>never</u> possible for any given heat engine to convert <u>all</u> the heat it takes in into work; some heat has always to be thrown away as part of the cycle of operations. Even the most efficient heat engine is subject to this limitation. Of course, no attempt should be made to introduce Carnot cycles into the discussion at this point; no surer way could be found of killing pupils' interest. Instead, pupils should be made to realize that this experimental fact concerning the behaviour of heat engines is not at all a consequence of the Principle of Conservation of Energy; they can be led from this idea to think about certain simple processes which never occur in nature but which would also not violate the Principle of Conservation of Energy. Examples of such processes are:-

(i) When two bodies at different temperatures are placed in contact, it is never observed that the cold body gets colder and the hot body gets hotter, i.e. heat never flows from the colder to the hotter body.

/(ii).....

(ii) A flywheel never spontaneously cools down and starts to rotate at the same time. That these things never happen, that nobody would for one moment expect them to happen must presumably be a consequence of a general principle or law not as yet made known to pupils. Clearly the law has to do with the limitations of the conversion from "heat" to work. Perhaps the teacher could go as far as to produce a simplified version of the Clausius statement of the Secord Law at this point, viz. "Heat does not of its own accord flow from a cold to a hot body." But of greater interest and urgency is the need to view the matter microscopically and to appreciate that processes of the kind mentioned above would correspond to spontaneous transitions from molecular chaos to (comparative) molecular order, and are hence extremely unlikely. On the other hand, spontaneous transitions in the opposite direction are "natural". This is the essence of the microscopic view of the Second Law.

Some pupils will have heard of the magic word "entropy" and may demand some kind of explanation of it. It may as well be accepted from the start that there is no simple way of explaining entropy to pupils as a macroscopic concept; as such, it defies any attempt at simple quantitative or "denotative" definition. (Theobald's (1966) "Entropy is... a measure of the extent to which a system lacks excess heat energy of a potential character" is both satisfactory and "simple", but not in the sense that it could easily be understood by beginners.) This is hardly surprising; the arguments by which one arrives at the concept of entropy in classical thermodynamics are both long and subtle, and one ought not reasonably to expect that such a concept could easily be wrapped up in a neat verbal package. From the microscopic or statistical point of view it is possible to say

/something.....

something about entropy that might be of value to the pupil. It can be represented as a measure of molecular disorder. The term "disorder" is used here in a particular and precise sense of which the teacher should be aware. To explain this, we consider the statistical-mechanical definition of entropy. In statistical mechanics, the entropy S of a system is given by the equation $S = k \log W$, where k is a proportionality constant (Boltzmann's constant, in fact) and W is a quantity sometimes called the "thermodynamic probability" of the system. It is this quantity W which corresponds directly with what is meant in statistical mechanics by "disorder" or "degree of randomness". To see what it means, we need to look briefly at some fundamental ideas in statistical mechanics.

For any system composed of a large number of microscopic "particles", there are many possible ways in which the particles can be distributed among the possible energy levels of the system, consistent with a given total internal energy. Any one such way of distributing the particles among the available energy levels is called a "complexion" or "microstate" of the system; it will correspond to a particular eigenfunction of the system as a whole. In general, a large number of complexions may be associated with a given "distribution-in-energy" of the system, i.e. to n_1 particles being "in" energy level 1, n_2 particles in energy level 2, and so on. (We follow here the standard "simple method" of presenting this aspect of the topic, in which it is assumed that effectively there are many eigenstates associated with a given energy level: see Band 1955.) It is a fundamental postulate of statistical mechanics that all complexions are equally probable, so that the most probable "distribution-in-energy" is that which has the largest number of complexions associated with it, i.e. the largest number of system eigenfunctions corresponding to it. It is the number of complexions or

/system.....

system eigenfunctions corresponding to the most probable distribution-in-energy of the system which is the quantity represented by W in the above statistical-mechanical expression for entropy.

The connection between W and "disorder" can easily be seen. The more "ordered" a system is, in the statistical-mechanical sense, the smaller will be the number of ways in which its most probable "distribution-in-energy" can be achieved. A homely example given by Feynman et al (1963) helps to illustrate this point. A collection of "white and black molecules" is to be distributed over a number of volume elements in space. A distribution in which "white is on one side" and "black on the other" can be achieved in fewer ways than one in which there is "no restriction on which goes where." Naively, we would regard the latter situation as being more "disordered", and this is seen to be consistent with the idea that "We measure 'disorder' by the number of ways that the insides can be arranged, so that from the outside it looks the same" (Feynman et al 1963). We would expect, too, that if the system of white and black molecules were started off in the condition of "white molecules on one side and black on the other" and were left to itself with the molecules free to move around, that they would distribute themselves in a manner which was increasingly "disorderly", thereby demonstrating the tendency for the entropy of an isolated system to increase.

It is easy to find other simple examples of correspondence between "disorderliness" and the value of W. There are much greater limitations on the motion of the molecules of a substance in the liquid phase than in the gaseous phase. Consequently, a particular "state" of a given mass of liquid can be realized in fewer ways than a particular state of a

/similar.....

similar mass of gas, so that from the point of view of statistical mechanics, the liquid phase is more ordered than the gaseous phase. Again, this checks with one's ordinary ideas about order and disorder. Also, two bodies represent a condition of greater molecular order when they are separately at different temperatures than when they are placed in contact and allowed to attain thermal equilibrium.

The examples given here are all consistent with one's naive, intuitive ideas of molecular order and disorder. The teacher might want to give an interested pupil an example of relative order and disorder as disclosed by actual numbers of "complexions" in a very simple situation. An instructive example would be a set of three distinguishable atoms (e.g. a "very simple solid consisting of three atoms" - Wilks 1961), each capable of taking on amounts of energy 0, E, 2E, etc. Suppose that this system has total energy 3E. There are six possible complexions associated with the distribution in which one atom has zero energy, one atom has E units of energy and the remaining atom 2E units; there are three complexions in which one atom has 3E units and the others have none; and there is one complexion in which each atom has E units. From the point of view of statistical mechanics, the last arrangement is the most ordered and the first the least: an interesting result, because it might at first be thought that a distribution-in-energy in which one atom was to be found on each energy "shelf" (0,E and 2E) was just as much "ordered" as one in which all atoms were on the same energy shelf. This type of example might make clear to teacher and pupil that naive ideas about order and disorder might not always be helpful in understanding entropy from the statistical point of view. (In this connection, cf. Glasser 1969.)

/5.4.14.....

5.4.14 The solid state: crystallinity

In view of the amount of time and money expended on research in solid state physics nowadays, it is certainly fitting that space should be devoted to the solid state in a modern physics syllabus. The real question to be decided is "How much?". The only realistic answer is, regrettably, "Very little"; very little, that is to say, in comparison with the range and complexity of the subject as it now is. The list of topics given under the heading of "Solids" in the Common Basic Syllabuses provides the basis for an introduction to the subject that is more than adequate. The list is: "Crystallinity, architecture of some important crystals, packing schemes, crystals in terms of bond types, polymers, physical properties, lattice energy, crystallization, melting." This general scheme is evidently intended to give the pupil some idea of how matter in the solid state is put together and of the processes of transition to and from this state. The entry "physical properties" is alarmingly vague; we shall discuss below how it might best be interpreted.

In the junior secondary school, pupils should already have acquired some experience of looking at crystals, growing crystals and making models of crystals, so that there should be little need for the teacher to emphasize that matter in the solid state is largely crystalline. The pupil is now ready to consider how and why the "molecules" of particular substances arrange themselves in particular crystalline forms. But before he studies this, he should have the following important question answered for him: "How does one gain knowledge about the crystalline structure of a material?" As he is about to hear it asserted that certain kinds of material have particular crystal structures, he should have some idea of where this knowledge comes from. From macroscopic

/measurements,....

measurements, two kinds of information can be obtained:
(i) The shape of the unit cell of the crystal can often be determined from measurements of the angles between the natural faces of the crystal (interfacial angles).
(ii) The symmetry of the arrangement of the "atoms" in a crystal can sometimes be inferred from the "habit" - characteristic shape - of the crystal and from the anisotropy of its physical properties (i.e. their dependence on direction).

To gain more detailed information, it is necessary to use the technique of X-ray (or other short wavelength) diffraction. This on its own will not usually be enough to pin down the structure to some definite form; one has to invoke other properties of or facts concerning the crystal in order to complete the picture. (For a more extended but still simple account of these ideas, cf. Holden and Singer 1961.)

Prior to the consideration of the "architecture of some important crystals", it is desirable for the pupil to study with the aid of models some of the ways in which "atoms" could be packed together in crystalline structures. This examination of packing schemes should start with close-packed structures, i.e. hexagonal and cubical close-packing (face-centred cubic) structures. It would be worthwhile pointing out that structures like these can be expected for crystals of <u>similar</u> atoms when the bonds between the atoms have no particular directionality. From here one can move on to body-centred cubic structures, which are no longer close-packed; it can be suggested that in the case of such crystals (again, if they are composed of <u>similar</u> atoms), the bonds between the atoms must to some extent be directional, or that the atoms are not effectively spherical (cf. Bunn 1964, Holden 1965). The simple cubic structures should obviously be discussed here also.

/It.....

It is usual in sophisticated treatments to describe a crystal in terms of a lattice and a basis (cf. Kittel 1966). The lattice is "a regular periodic arrangement of points in space" (Kittel, op. cit.); it is the abstract framework on which the true crystal structure is erected by attaching a group of suitably-chosen and- arranged atoms identically to each point of the lattice. Any such group of atoms is called the basis of the structure. For example, a sodium chloride crystal has a face-centred cubic lattice and a basis consisting of one sodium ion and one chlorine ion separated by one-half the body diagonal of a unit cube. This method of describing a crystal structure is particularly well suited to the analysis of the structure by X-ray or neutron diffraction, as Kittel has pointed out, but it is open to doubt whether there is much point in introducing it in an elementary course, in which there is likely to be very little time available for the discussion of this topic. Perhaps it would be better to follow Holden's (1965) suggestion of viewing certain complicated structures as consisting of interpenetrating simple substructures. For example, the sodium chloride structure can be thought of as being made up of two interpenetrating face-centred cubic centres, one of chlorine ions and the other of sodium ions. If Holden's suggestion is followed, then in an elementary treatment of crystal structure, the only basic types of crystal structure that need be discussed are the ones mentioned in the previous paragraph. To elaborate on these ideas to the extent of discussing "unit cells" for crystal types is not, we think, justified in the present context; this is a matter for the interested pupil to pursue on his own (he will find excellent guidance in Holden and Singer 1961).

/A.....

A variety of crystals suggest themselves as "important" and, as such, candidates for inclusion in any survey of the "architecture of some important crystals". Metals are certainly important and we suggest that pupils should know the names of some of the metals which exemplify the simple types of structure mentioned above. This information is summarized in convenient form in Table 6 of Chapter Eight of Holden's book "The Nature of Solids" (Holden 1965): here we may mention that beryllium, cobalt, magnesium and zinc exhibit hexagonal close-packed structures, aluminium, copper, gold and silver face-centred cubic structures, and chromium, iron, sodium and tungsten body-centred cubic structures. (It is significant that the last-mentioned metals do not show close packing.) Sodium chloride must be included, in view of the familiarity of the substance. The different crystalline forms of carbon - diamond and graphite - must be shown. The diamond structure (useful as an example of a covalent crystal) can be regarded as two interpenetrating face-centred cubic substructures displaced from each other along the body diagonal of a cubic unit cell. There is no point in extending this list here, but there is no reason why teachers should not add to it if they wish. Probably the best way of handling this topic is to set up a "museum" of important crystal structures, with each exhibit adequately captioned.

In surveying packing schemes and crystal structures, the teacher will probably already have given hints of the dependence of crystal structure on the nature of the bonds linking the "atoms" or "molecules". This type of discussion now needs to be extended into an account of the four idealized types of crystal, viz. ionic, molecular, covalent (or adamantine) and metallic. In a pure physics course, the emphasis in such an account should be on the relative strengths and directional properties of these bonds and the influence of

/these.....

these factors on the type of crystalline structure obtained. Even if the names of the different types of bond do not mean very much to the pupil as yet, a treatment along the lines indicated should be of some help to him. It can be mentioned that in "molecular" crystals, the electrically neutral "molecules" are held rather weakly together by Van der Waals forces which also have no marked directional properties. Such crystals are expected to show close-packed structures. Examples are the solid forms of the rare gases, hydrogen, methane and the pareffins. In ionic crystals, the bonds are much stronger and still essentially non-directional (although it is possible for positive and negative ions mutually to distort each other's charge distribution and so polarize one another, giving rise to a directional polarized ionic bond; this is especially likely for other than univalent ions). Thus it is still possible for them to form what are essentially close-packed structures: we have seen that sodium chloride can be regarded as consisting of two interpenetrating face-centred cubic structures. The relative sizes of the ions may determine the actual form of the structure: that the negative ions in sodium chloride are larger than the positive ions results in the structure taking on the form just described, while in caesium chloride the more nearly equal sizes of the positive and negative ions means that, say, a positive ion can have eight nearest negative-ion neighbours, instead of six as in sodium chloride, and the structure is effectively two interpenetrating simple cubic structures.

Covalent or adamantine crystals are characterized by strong bonds with very marked directionality (the number of nearest neighbours of any given atom is just equal to its valence). Consequently, this type of crystal must be expected to show some other form of structure than close-packed.

/The.....

The most celebrated example of a covalent crystal is diamond; other examples are germanium and boron nitride. In metallic crystals the bonds are moderately strong but are neither directional nor specific (in the sense that the metallic bond does not mind which metallic atoms it pulls togetner). Close-packing structures must be expected in "pure" metallic crystals, i.e. those characterized by metallic bonding only. Examples of such crystals have already been given above.

Pupils should not be left with the idea that in any given crystal, only one type of bonding is to be expected. It has already been noted that certain metallic crystals do have some tendency towards directionality. Covalent bonds between atoms of different species do have a slight flavour of ionic bonding; and Van der Waals forces, although their effect may be swamped by those of the other types of bond, are always present. In fact most actual crystals represent various combinations of the basic types of bonding.

To round out the picture the pupil has of the structure of solids, he should learn something of the character of partly crystalline or non-crystalline (amorphous) materials. The best-known example of an amorphous solid is glass, but the characteristic features of glass are best understood with reference to the properties of liquids, and if they are to be discussed at all, they are best left until the liquid state has been studied. Of partly crystalline materials the polymers are perhaps the most important; as has been pointed out (cf. Mark 1967), life depends fundamentally on organic polymers. The distinguishing feature of polymeric materials is the enormous size of their molecules, containing in some cases millions of atoms. In general, natural polymers are partly crystalline and partly amorphous, their physical

/properties.....

properties depending on whether crystallinity or amorphousness predominates. In synthesizing man-made polymers, chemists have sought to produce materials with desired characteristics. If the polymeric chains are suitably orientated, they can be persuaded to form crystalline-like structures by Van der Waals and electrostatic bonding between the chains, and so produce materials of great strength and resistance to heat. The qualities of strength and resistance can also be given to polymers by linking the polymeric chains chemically with one another.

5.4.15 The solid state: physical properties

The deduction of the physical properties of materials from their structure is one of the major tasks of the solid state physicist. Pupils should be given a little insight into this kind of activity by way of the qualitative deduction of certain selected physical properties. These last we think should be mechanical and thermal properties.

(i) Mechanical Properties. The behaviour of solids under tension and compression has already been used as a partial justification for pupils' benefit of the properties of the intermolecular force. The shape of the curve of force versus separation implies that for small strains, the behaviour of the material should be described by Hooke's Law. If the material is one in which the intermolecular force is comparatively large (i.e. its value for a given intermolecular separation is comparatively large), we should have to apply a large force to achieve a given elongation of the specimen (i.e. we should expect a high value for Young's modulus). The behaviour of a solid that is stretched beyond its elastic limit should be directly observed by pupils: the Nuffield Course makes provision in Year I for pupils to study this for

/a.....

a stretched wire. This "plastic flow" in metals could be explained to pupils in simple qualitative terms: suitable discussions are given by Mott (1967), Cottrell (1967) and Tabor (1965). The last-mentioned reference contains a useful general account of the relationship between the mechanical properties of various materials and their microscopic structure.

(ii) Thermal properties. The microscopic explanation of thermal expansion has been referred to in Section 3.12.1; that of thermal conduction was dealt with in Section 3.12.3. The possession of a thermal capacity by a solid is easily explained in terms of the kinetic theory of matter. If we suppose the "molecules" in a solid to be vibrating randomly about their mean positions, then the internal energy of the solid is simply the total kinetic energy of these molecules plus their mutual potential energy. So when we raise or lower the temperature of a solid we are simply adding to or taking away from this store of energy associated with the vibratory motion of the molecules.

Simple kinetic theory can be used to predict a value for the molar heat capacity of a solid if it is assumed that average kinetic and potential energies of a molecule in a solid are each equal to $\frac{3}{2}$ kT, where "k" is Boltzmann's constant. This leads to a value of 3R for the molar heat capacity of any solid at any temperature at which it stays solid (assuming that in metals, the electrons, for no apparent reason, make no contribution to the heat capacity). Experiment shows that the molar heat capacity of a solid is strongly dependent on temperature, tending to zero at absolute zero and to 3R at high temperatures. Classical kinetic theory cannot be modified in such a way as to bring it in line

/with.....

with experiment; use of the quantum theory does produce satisfactory agreement with experiment. This snippet of quantitative solid state physics is certainly something the teacher should know about and keep in mind, but one cannot convincingly argue that it can be included in a school syllabus.

5.4.16 The solid state: lattice energy

The existence of intermolecular forces entails the existence of mutual potential energy of the molecules composing any sample of matter in bulk. Let us choose as our zero for this mutual potential energy, the potential energy of the sample when its constituent molecules are separated by infinitely great distances. Then when the sample is in the solid state and its "molecules" are in their mean positions, their mutual potential energy will have a certain appreciable negative value. This is the binding energy of the solid or, as it is often called, the lattice energy; its magnitude is the amount of energy needed to tear the molecules apart from one another. If the condition of our sample of matter is such that the mean molecular separation is different from its minimum equilibrium value, then the mutual potential energy of the sample will be larger (less negative) than the lattice energy.

This account shows that it is comparatively easy to explain the concept of lattice energy. Certainly if pupils are able to follow the earlier discussions on intermolecular force and potential energy, the concept should give them no trouble. We are inclined to wonder, however, whether its inclusion in the Common Basic Syllabuses has much point. In an introductory course in solid state physics, it is customary to derive an expression for the lattice energy of an ionic crystal and to compare calculated and experimental values of lattice energy. There can be no question of doing any of

/this.....

this in a school course; without such development, there seems little need to add another new term to the pupil's vocabulary. What he does need to know is that the mutual potential energy of a system of molecules depends on their mean separation; this is particularly important in helping him to understand the energetic aspects of changes of phase.

5.4.17 Changes of phase

In connection with changes of phase, the pupil needs to have some idea of what is happening, both macroscopically and microscopically. He knows of the existence of "latent heats" of changes of phase, and he knows also that he can interpret the absorption of "heat" by a melting substance as an increase of internal energy of that substance. He needs to be able to interpret that increase in internal energy in terms of what happens to the molecules; and if he is particularly bright he may want to know why a change of phase should take place very abruptly at an extremely well-defined temperature.

The first point is readily explicable if it is assumed that the "molecules" of a substance are on the average, further apart in the liquid state than they are in the solid state. Then it follows from the above discussion concerning mutual potential energy of the molecules that in the liquid state, the contribution to the internal energy from the mutual potential energy will be greater in the liquid state than in the solid state. This simple account obviously runs into difficulty in the case of water, as the density of ice at 0°C is appreciably less than that of liquid water at 0°C. (As is well known, the anomalous behaviour of water in this respect comes about because of the hydrogen bonding between neighbouring water molecules, which favours an open structure in the solid state.) In this case one should proceed with

/caution.....

caution, taking refuge at this level in the statement that it remains generally true that in the liquid state the average molecular energy must be greater than in the solid state, and making reference to the structure of ice. If this seems a faint-hearted way to deal with the problem, one may take comfort from the fact that textbooks generally avoid discussion of it altogether; it is perhaps wiser not to rush in where the angels have shown a marked reluctance to tread.

The second point above - the well-defined temperature at which changes of phase occur, as in melting - is difficult to explain convincingly at an elementary level. There exists as yet no generally accepted theory of melting (cf. Zemansky 1968), so that any attempt to explain the process of the onset of melting in terms of molecular behaviour must be largely speculative. The following account by Reif (1967) might help one to form some kind of picture of melting, but it should not be taken too literally:-

If a solid is hot enough, its molecules will have enough kinetic plus potential energy to be able to oscillate about their mean positions with amplitudes comparable with their mean intermolecular separation. If a few adjacent molecules move <u>simultaneously</u> out of their mean positions in the lattice (there is always a finite probability that this will happen), it is then easier for neighbouring molecules slightly further away to move out of <u>their</u> mean positions, and in this way the process can spread throughout the entire solid, resulting in a collapse of the ordered arrangement of its molecules into the (comparatively) disordered condition of the liquid state. This sudden change of the degree of order amongst the molecules is characteristic of a change in phase and is worth pointing out, even in a school course.

/Perhaps.....

Perhaps the teacher would prefer to use the briefer and less specific statement by Kitaigorodskiy (1967): "At a given pressure, melting occurs at a strictly defined temperature corresponding to the point at which the motion of atoms or molecules becomes too extensive for long-range order to be maintained any longer."

A thermodynamic argument for the existence of a specific transition temperature can be given in terms of the Gibbs free energy. In reference to melting, the argument can be put as follows:- Above the melting point temperature, the Gibbs free energy per unit mass of the solid phase of the substance would be greater than that of the liquid phase, so that the latter corresponds to the stable phase at temperatures higher than the melting point (Pippard 1957, Morse 1964). To make this intelligible to a pupil requires some explanation of the concept of "free energy". About the best that can be done in this direction is to say that when one studies phenomena like melting or evaporation, it is advantageous to make use of a quantity representing the amount of work that can be withdrawn at constant pressure and temperature from the system and which is in some ways analogous to potential energy. Then it might mean something to the pupil to refer to the connection between stability and free energy. However, the pupil would still wish to know <u>why</u> the free energy in the solid phase was higher than that in the liquid phase above a certain temperature. Short of actually writing down the expression for the Gibbs free energy and working with it, there is little the teacher can do about this except to point out that the free energy of a substance depends on its internal energy, temperature and entropy (at constant pressure and provided that volume changes are of little account); that at constant temperature, increase of internal energy would raise the free energy and increase of entropy (increase of

/disorder).....

disorder) would lower it; that the solid phase of a given mass of substance has less internal energy and less entropy than the liquid phase at a given temperature; and that below a certain temperature, the decrease in internal energy produced by the freezing of a substance would more than compensate for the increase in free energy associated with the decrease in entropy, so that there would be a net decrease in free energy.

At the end of all this, some mystery would still remain. It is small wonder that Charles (1967), in writing a popular article, should have attempted to present what is essentially the above argument in terms even more fundamental. He uses the thermodynamic and the statistical-mechanical definitions of entropy, the Second Law of Thermodynamics and the assumption that "the geometric arrays produced by freezing are fairly independent of temperature" to show that there must be a specific freezing point for a substance: "at no other higher temperature can the geometric ordering required for crystal formation occur and leave the order of the universe undisturbed". The essential points of Charles's argument (not all of them explicitly stated by him) are:-

(i) The decrease in entropy of a freezing substance is determined by the "ordering" that takes place, i.e. essentially by the "geometric arrays produced by freezing";

(ii) the associated increase in entropy of the surroundings is determined by the heat given out by the freezing substance, divided by the absolute temperature;

(iii) the heat given out during freezing is also determined by "the amount of ordering that occurs on crystal formation";

(iv) from points (i) - (iii), it can be inferred that only if the temperature is equal to or less than a certain value can the increase in entropy of the surroundings be equal to or greater than the magnitude of the decrease in entropy of the

/freezing.....

freezing substance, and so satisfy the Second Law of Thermodynamics. If Charles's article is available, the teacher could refer the interested pupil to it and be prepared to elaborate on it, keeping the above points in mind.

From the above discussion one would expect that for a crystalline solid, the onset of melting as the temperature is raised would be sudden, as indeed we know it is. In trying to picture the reverse process of crystallization in these terms, one realizes that it does not follow that as the temperature is lowered, the onset of crystallization should necessarily be sudden - nor is it. It may be that crystallization may only occur at temperatures somewhat below the freezing point - this is the phenomenon of super-, sub- or undercooling (cf. Knight 1967). In certain cases solidification may take place without crystallization; this is what happens in the formation of glasses. From his own experience, the author knows that the phenomenon of supercooling leaves one with a nagging feeling of mystery if no attempt is made to explain it. To tell much of the story to pupils, even in a simplified version, would be difficult; the energetic aspects of the formation of crystal nuclei might puzzle the average class member. On the other hand, such a pupil should be able to follow a simple account of what the molecules of a freezing liquid have to do in order to take up their positions in the regular lattice of a crystalline solid. It is obvious that for the molecules to be able to do this, they must slide past one another and also alter their relative orientations. For most common liquids, the roughly spherical shape of the constituent molecules promotes easy sliding, and at temperatures only slightly below the freezing point it is quite easy for the molecules to arrange themselves in orderly fashion under the influence of their mutual attractions. They still would not do so, however, unless crystal nuclei

/were.....

were formed to start the process off; in Knight's words, we must have "the spontaneous appearance of a little ordered domain in the midst of randomness." But even if a crystal nucleus were to form spontaneously, it would not continue to grow unless its diameter exceeded some critical value (Turnbull 1965, Knight 1967). This is because energy is needed to create the interface between the crystal nucleus and the surrounding liquid. Only at a particular radius of the crystal nucleus will the decrease in the "free energy" associated with the change from the liquid to the crystalline state more than counterbalance the increase in the free energy associated with the formation of the interface between crystal and liquid. The lower the temperature to which the liquid is supercooled, the smaller will be this "critical radius" and the easier it will be for crystal nuclei to form. In liquids under normal conditions of purity (or impurity), the formation of these nuclei is aided by the presence of "motes" (impurity particles) in the liquid or on the walls of its container. Vibrating the liquid also aids the onset of crystallization.

By way of prefacing the above description of crystallization and supercooling, we remarked on the air of mystery that becomes attached to a phenomenon like supercooling in the absence of attempts to explain or describe it properly. We hope that the facts set out in the preceding paragraph would help a teacher to dispel such mystery; but we must also point out that there are aspects of supercooling and crystallization which are baffling even to present-day workers in the field (cf Turnbull 1965). Thus it is not understood exactly why "motes" in a liquid assist "nucleation", nor why vibrating a liquid should have a similar effect. There seems as yet to be no agreed explanation of why the energy needed to form unit area of interface between liquid and crystal is so large. We mention these points as we feel

/it.....

it would be as well for the teacher to be aware that this phenomenon, like many others associated with matter in bulk, is by no means fully understood.

5.4.18 The structure of liquids

In our observations on change of phase, we have already mentioned some of the points that should be made in describing the liquid state to pupils, viz. the "disorder" of the liquid state as compared with the solid state and the increased mutual potential energy of the liquid (per unit mass) as compared with that of the corresponding solid at the same temperature. The pupil needs to know a little more about liquids than this if he is to understand surface tension, for example; and his teacher should know still more by way of background, particularly about the "structure" of liquids. The following summary would be intended primarily for the teacher: we consider at the end of it how much of this should be known to the pupil and how it might be approached.

Let us remind ourselves of some of the significant (macroscopic) properties of liquids. They are highly incompressible, and their densities are much like those of the corresponding solids (within ten per cent near the melting point temperature). We usually say that they cannot transmit a shearing stress; that is, that we cannot twist a portion of the liquid. All this gives rise to our usual microscopic picture of a liquid: "The molecules...execute vibratory motion about the centres which are free to move but which remain at approximately the same distances from one another" (Sears and Zemansky 1964). From the fact that the densities are not significantly different in the liquid phase from those in the solid phase, one must conclude that the intermolecular distances cannot be much greater in the liquid phase than they are in the solid phase.

/Does.....

Does a liquid exhibit any kind of orderliness in its structure comparable to that displayed by crystalline solids? We must examine the evidence. First, there is the interesting fact that liquids do transmit shear waves provided that their frequency is high enough (i.e. ultrasonic). This is comparable to the behaviour of a solid and suggests that for short times and for localized regions the liquid possesses a rigid lattice structure. This conclusion is supported by experiments on the diffraction of neutron beams by liquids, the results of which imply that molecules in liquids have time to execute from 10 to 100 vibrations before the structure changes (cf Bernal 1960). But what kind of structure? Some have put forward the view that the instantaneous structure of the liquid is such that the immediate environment of any molecule is similar to that of the crystalline phase of the substance. But this view runs into the difficulty that some liquids can so easily be supercooled to temperatures well below their freezing point. If a liquid really did contain crystal-like structures, it is difficult to see why it would not simply form a fully crystalline structure on being cooled only slightly.

An alternative approach to the problem of liquid structure has been to start off with the requirement that the molecules of a liquid must be arranged in such a way that they are spaced at roughly the equilibrium spacing from one another, and to investigate the ways in which they could pack together. The most exhaustive investigations of this type have been carried out by Bernal (op. cit.) who, after trying out various "dense-random" structures, arrived at a picture of an "ideal" liquid. In this picture, a basic requirement is that the spacings between neighbouring molecules should be the same, at any rate to a first approximation. Any

/point.....

point in the liquid not at the location of a molecule will lie inside a polyhedral hole whose edges are formed by the lines joining the centres of the molecules in the vicinity of the point. In view of the requirement that the distances between neighbouring molecules should be the same, these polyhedral holes should have equal edges. The five smallest equal-edged polyhedra are the regular tetrahedron and octahedron and three semi-regular figures having twelve, fourteen and sixteen faces respectively. If attention is confined to "dense-packed" structures, in which the polyhedral holes are not large enough to contain any extra molecule inside them, only these five polyhedra need be considered. An "ideal liquid" structure would be formed by putting together various combinations of these polyhedra. A regularly repeating pattern of polyhedra can only be obtained if only two of the polyhedral types are used, viz. the tetrahedron and octahedron, and they must be in the proportion of two of the former to one of the latter. If a structure consisting of an arbitrary arrangement of the five polyhedral types is to fill the available space, then some of the faces of the polyhedral holes must be distorted, with consequent distortion of the edge lengths. It is extremely improbable that if a collection of similar molecules were to come together to form a structure in which neighbouring molecules were roughly at equal distances from one another, they would so group themselves that they would form a regularly repeating structure, since only two or three arrangements of the polyhedral holes could correspond to this, whereas there are very many ways in which arbitrary combinations of the slightly distorted polyhedral holes could produce a space-filling structure. Thus the characteristic "ideal liquid" dense-packed structure

/is.....

is one in which all of the five possible types of polyhedral hole are to be found and in which, therefore, the mean distance between neighbouring molecules is likely to be somewhat different from the equilibrium distance. In sum, Bernal's picture of a liquid is an irregular structure composed of myriads of polyhedra, with the molecules exchanging positions and migrating through the structure.

The above kind of model of liquid structure shows what is called "short-range order" in the arrangement of the molecules; that is, the immediate neighbours of a given molecule will not be grouped around it in a completely random manner but in an arrangement that is somewhat like that of the molecules in a crystalline solid (cf Bernal 1960, Pryde 1966). However, this orderliness does not extend beyond a few molecular diameters' distance from the given molecule. The existence of this short-range order in liquids has long been known from X-ray diffraction studies of liquids.

Of all the above, we suggest that the following should be known to pupils:-

(i) The macroscopic properties of liquids to which reference was made at the beginning of the above description.

(ii) The obvious inferences from these properties of the microscopic properties of the liquid state, viz. that the molecules are on the average further apart in the liquid state than in the solid state; that although they are still bound together almost as strongly as in the solid state, they are presumably able to slide over and exchange places with one another.

(iii) That the "life" of a liquid consists of a sequence of extremely short but significant periods during each of which it has a disordered, but quasi-rigid structure, so that during such periods the molecules of the liquid vibrate about mean positions which remain momentarily fixed.

/Points.....

Points (i) and (ii) depend only on simple observation and straightforward inference from it. Point (iii) has, regrettably, to be largely a matter of "telling", enlivened perhaps by illustrations taken from Bernal's "Scientific American" article and by use of the two-dimensional ball-bearing model described in Section 5.4.2.

Knowing something of the liquid state, of supercooling and crystallization, the pupil should be able to understand the nature of glass. (One would not suggest this as a syllabus topic, but glass is such a familiar substance that curiosity about its nature is bound to be felt, and the teacher should be ready to satisfy it.) Suppose we have a liquid of very high viscosity; that is, the molecules have great difficulty in sliding past one another. In such a liquid, the formation and growth of crystal nuclei may be impossible even when the liquid is cooled down very slowly, allowing the molecules a good deal of time in which to take up orderly positions relative to one another. The substance may then become rigid with a random structure frozen into it, characteristic of the liquid at a much higher temperature. Crystallographic studies confirm that glass has the short-range order characteristic of a liquid. It is evident from this description why glass is sometimes called a supercooled liquid of extremely high viscosity.

The lack of a definite melting point may be explained by the inequality of the lengths of similar bonds between atoms; varying amounts of energy are required to break these bonds, so that the higher the temperature, the more bonds can be broken. Alternatively, we can say that a cooperative collapse of a structure such as occurs in the melting of a crystalline solid cannot occur in glass, because in the latter case the structure is already in a state of collapse.

/5.4.19.....

5.4.19 Surface tension of liquids: introduction

This topic can fairly be said to demand inclusion in any elementary syllabus in physics or physical science. (The PSSC course does not deal with it explicitly, but the topic does hover in the background, so to speak, of the important experiment on the thickness of a monomolecular layer.) Every time we see a liquid drop or meniscus we are witnessing the effect of surface tension; the phenomenon is an important one in everyday life and it is the basis of many pretty demonstrations, some of which pupils may have known about since their kindergarten days; and it is a striking demonstration of the reality of intermolecular forces. Yet despite its familiarity, it is customarily taught in a highly unsatisfactory manner; and those who have tried to improve on this have not been unanimous about the way in which the improvement should be effected, since they evidently have divergent views of the nature of surface tension.

The traditional way of explaining surface tension in terms of intermolecular forces is to say that a molecule in the interior of the liquid is surrounded on all sides by neighbouring molecules, all exerting attractive forces on it, so that the net force on it is zero, whereas a molecule in the surface of the liquid, having virtually no liquid molecules above it, must be subject to a resultant downwards attractive force, so that the molecule will tend to be pulled inside the liquid. Hence the free surface of a liquid tends spontaneously to contract, and we have surface tension. More often than not it is stated that molecules in the liquid surface are subject to an unbalanced downward force (e.g. Gordon and Neser, no date, and Ference et al 1956; actually this statement and "explanation" turn up in so many textbooks that it is almost unfair to single out certain of them for mention). The absurdity of the latter idea is patent

(cf.....

(cf Warren 1965), since it leads to a picture of inwardly collapsing matter; also, the suggestion that a molecule in the interior of a liquid is acted upon by attractive forces only, leads to the idea that such a molecule will be in unstable equilibrium (cf our discussion of intermolecular forces in Section 5.4.1).

McKeehan (1948), who drew attention to these difficulties some time ago, charitably assumes that "an unwary student (or teacher) is likely to conclude that the unsymmetrical set of attractions [on a molecule in the surface] has to be balanced by a special - and unpictured - repulsive force, normal to the surface, to be associated in some mysterious way with a set of tensions in the surface - also not pictured". On the contrary, experience leads one to believe that "the unwary student (or teacher)" or textbook writer never gets round to thinking about repulsive forces at all. One can excuse the pupil or student in this; if neither teacher nor textbook trouble to mention repulsive forces between molecules, then the pupil or student must not be expected to conjure these up for himself; he is not yet sufficiently confident of his physical thinking to be able properly to appreciate the consequences of the absence of repulsive forces. There is obviously a real responsibility on teachers and textbook authors not to corrupt pupils in explaining this phenomenon to them; at the very least, care should be taken to introduce simple ideas about attractive and repulsive forces in a manner which implies no absurdities.

5.4.20 A survey of views of surface tension

As we have noted above, there are differing views concerning the way to teach surface tension; in particular, there is disagreement about the correct molecular explanation of surface tension. All authorities seem to agree that

/molecules.....

molecules within the liquid and in its surface must be in a condition of stable equilibrium; all agree that molecules within the liquid surface are in a different field of force from that in which molecules well inside the liquid find themselves, and that consequently the mutual potential energy of a group of neighbouring surface molecules will be greater than that of a corresponding group of molecules well inside the liquid. From this point onwards, the views begin to diverge.

Feather (1959) regards any molecule of the liquid, whether it be in the surface or in the interior, as being under the influence of "steady" attractive forces (the quotation marks are his). For him the repulsive forces are "transient", coming into play only during "collisions" (quotation marks ours), and being responsible for the random changes in kinetic energy of a given molecule. According to him, these "steady" attractive forces balance out on a molecule inside the liquid, and have a net downwards component on a molecule in the surface. Adam (1941) takes essentially the same view, attributing "surface tension" to the existence of a net downward attractive force on a surface molecule, although he does not mention repulsions. Rogers (1960) is evidently in agreement with these views, but he gives a more detailed account of surface tension in terms of molecular behaviour than do Feather or Adam. He says of a molecule inside the liquid that it is "jostled by frequent collisions with other molecules on all sides" and "also attracted to near neighbours ...with resultant pull zero." In accompanying diagrams he shows such a molecule experiencing zero resultant attraction (from "long range attractions of neighbours nearby") and zero resultant repulsion (from "short-range repulsions of very near neighbours during collisions"). On a molecule in the surface, there is a resultant downward

/attraction.....

attraction on the molecule, balanced by the repulsive forces operative during collisions with molecules from below. In other words, Rogers visualizes the equilibrium condition of a molecule in the surface as being like that of "an upside-down umbrella supported by upward-falling raindrops" (Rogers 1968b). He compares the behaviour of the molecules of a liquid drop with that of a crowd of people, the peripheral members of which are trying to push their way into the centre, and shows that the effect on the inner members is the same as would be produced by passing a belt around the perimeter and pulling it tight.

Chapin (1966) reproduces the familiar graph showing the variation of intermolecular force with distance, but in explaining surface tension he makes explicit mention only of the attractive forces. His diagram showing the attractive forces on molecules inside a liquid and in the liquid surface looks mildly unconventional at first, but this is only because it is printed upside down; it is, in fact, the usual diagram and the usual explanation accompanies it. He makes the additional observation that on account of the downward attractive force on a surface molecule "a shorter equilibrium distance then develops which results in there being more molecules per unit area in the [surface] film than in regions below it."

McKeehan (1948) prefers to approach the problem from the point of view of potential energy. He describes a molecule well inside the liquid as being in a potential-energy valley: the forces of attraction to its neighbours just balance the forces of attraction from its neighbours. If external pressure is applied to the liquid (e.g. by its own vapour pressure), the molecules will be pushed a little closer together and the interior of the liquid will acquire a definite

/potential.....

potential energy per unit volume with respect to its configuration at zero pressure. The value of the potential energy per unit volume near the surface must be higher than that at an interior point. In contradiction to Chapin, McKeehan says that the separation between neighbouring molecules must increase as the surface is approached. In general, he does not say very much about the behaviour of molecules in liquids, but he does seem to take a somewhat different view of the equilibrium condition of a liquid molecule from that of, say, Feather and Rogers. Whereas they see it as a dynamic process, with a "steady" attractive force being balanced by the averaged-out effect of the repulsive forces from transient collisions with other molecules, McKeehan's reference to molecules within the liquid being in potential-energy valleys suggests that he regards such molecules as being at the "equilibrium" distance from their neighbours, in essentially the same way as they would be in a solid.

Freier (1965) states this latter view explicitly: "Within a liquid", he says, "a molecule M is surrounded by nearest neighbours at an equilibrium distance" - "the forces are attractive if we try to pull the molecules apart and repulsive if we push them closer together." For him, the only difference between a molecule in the surface and one well inside the liquid is that the former molecule has fewer neighbours than the latter. But this difference is very important, because it means that to create a new portion of liquid surface, we have in effect to pull some of the neighbouring molecules away from the molecules which make up the new portion of surface. To do this, we need to do work. In terms of energy we can therefore explain the contractile tendency of liquid surfaces: for a small liquid drop, for example, the molecules would arrange themselves in such a

/way.....

way that the smallest possible number of molecules lay in the surface (i.e. the drop has the smallest possible surface area), because this corresponds to the minimum energy associated with the formation of the surface.

This explanation of surface tension implies that solids, too, must have a "surface tension", in the sense that energy is needed to create a solid surface. This expectation is confirmed by experiment: the "surface tensions" of several substances have been successfully measured at temperatures below their melting points (cf Cottrell 1964). The reason why liquid surfaces have an obvious contractile tendency whereas solids have not is, of course, to be found in the relative ease of mobility of the molecules of a liquid; it is easy for a given volume of liquid to adjust itself to a condition of minimum potential energy (see Section 5.4.21 for a fuller discussion of what is meant here by "potential energy").

Both the "Feather-Rogers" and the "Freier" treatments are capable of accounting in a general way for the phenomena attributable to "surface tension". Yet they depend on what seem to be essentially different views of the equilibrium condition of molecules in liquids, or indeed of the liquid state itself. Thus in the Feather-Rogers treatment the liquid is evidently treated as a very "dense gas," whereas in the Freier treatment it is treated as an "irregular solid" (these terms are borrowed from Pryde (1966)). Which, if either, is right? We can only look to experiment as a possible means of deciding between them. The relevant experimental evidence is that concerning the structure of liquids - the results of X-ray and neutron diffraction studies and of those of the propagation of ultrasonic shear waves. As we have seen, these point to the existence of <u>short-range</u> order amongst the

/molecules.....

molecules of the liquid, and to the possession by the liquid of an irregular but definite structure which undergoes very frequent metamorphoses. This would tend to support Freier's view - that we may picture the equilibrium positions of neighbouring molecules as being spaced apart by a distance roughly equal to the "equilibrium distance" corresponding to the equilibrium spacing of molecules in a solid.

Still, we should not be too hasty in preferring one description of the situation above the other. Each is to some extent an oversimplification of an exceedingly complex situation. The Feather-Rogers picture oversimplifies because it focuses attention on a supposedly static molecule which has other molecules dashing about in its vicinity. The Freier picture is an oversimplification in that, in effect, it ignores the molecular motion altogether and looks only at an idealized situation within the liquid in which the molecules can be considered to be momentarily frozen in their equilibrium positions. In actual fact, the molecules are vibrating randomly with, on average, considerable amplitudes about their mean positions. It is not surprising that two different kinds of oversimplification of the resulting molecular chaos should both prove to be capable of "explaining" the phenomenon of surface tension. Moelwyn-Hughes (1961a) remarked not so long ago that "[the liquid state] still remains the least understood of the states of matter." In view of our lack of understanding of it, we should be ready to accept that there is, as yet, no uniquely satisfactory way of accounting in molecular terms for various aspects of liquid behaviour.

5.4.21 Surface energy and surface tension

We have referred above to the fact that a liquid drop is able to assume the configuration appropriate to "minimum potential energy". The facts are somewhat loosely expressed

/by.....

by this statement. We have failed to specify that the changes of surface area spoken of in this connection must be isothermal, and that by "potential energy associated with the surface" we would mean what is properly called the (Helmholtz) "free energy" of the surface. It is shown in standard texts on thermodynamics (cf Pippard 1957, Zemansky 1957) that a system maintained at constant temperature and volume is in equilibrium when its (Helmholtz) free energy is a minimum. When work is done to increase the area of a liquid surface, and this expansion is carried out at constant temperature, the increase in internal energy per unit area of the surface is greater than the work done per unit area of new surface. This is because heat has to be added to keep the temperature constant, and the increase in internal energy is made up of contributions from both the work done and the heat added. It is the contribution from the work done which, if the process is reversible, becomes added "free energy" of the surface and which is recoverable if the surface contracts.

It is well known that the definitions of surface tension as "force per unit length of a line drawn in the surface of a liquid" and "free energy per unit surface" are equivalent. Does this mean that there is a real force parallel to the liquid surface, so that, as is often said, part of a liquid surface on one side of an imaginary line drawn in the surface will exert a pull on the liquid on the other side? This is sometimes categorically denied. (Cf. Cottrell (1964): "We do not regard surface tension as a tensile force between particles in the skin".) If there is such a force, how is it related to the force exerted by a liquid surface on a body with which it is in contact, and by what "mechanism" does such a force operate?

/Freier's.....

Freier's own discussion of surface tension stops short of the consideration of this problem. It is easiest to treat it by working backwards from the answers to the second pair of questions to the first. It is easily shown that "a liquid surface exerts a mechanical force on unit length of any solid body with which it is in contact, numerically equal to the free surface energy per unit area of the surface" (the quotation is from Yarnold 1960, but the derivation of this result may be found in many elementary textbooks, e.g. Feather 1959). This is the same as the force per unit length which one part of the liquid surface is supposed to exert on another. How does this surface-solid-body force arise? We can easily understand that when we stretch a liquid surface (e.g. a soap film on a wire frame with one movable arm), we need to apply a force, because we are in effect tearing away some of the neighbouring molecules from the molecules which form part of the new surface we are creating, and we have to pull against the long-range attractive forces between molecules to accomplish this. (There will, of course, be adhesive forces between the arms of the wire frame and the liquid which enable us to apply this pull in the first place.) Conversely, if we allow the surface to contract, we can imagine the following picture: an energetic surface molecule "jumps" into one of the "potential energy valleys" inside the liquid, leaving a gap between its original neighbours in the surface, which the long-range attractive forces between them will tend to close up.

What if the movable arm of the frame is kept stationary? In terms of the Freier picture of a liquid surface, the existence of a steady force exerted by the surface on the frame must mean that on average, neighbouring surface molecules are more widely separated than the equilibrium distance (cf.

/Colbeck.....

Colbeck 1968, Warren 1968). This in turn implies that there are fewer molecules in the surface layer than in corresponding layers within the liquid. The reason for this is the above-mentioned tendency for molecules to leap into potential energy valleys: as many as possible will go into the interior, where the potential energy is lower, leaving a depleted surface layer with an average intermolecular separation greater than the equilibrium value. It follows also from this picture that there must be a "tensile force between particles in the surface."

5.4.22 The elementary teaching of surface tension

It is generally accepted that in elementary syllabuses, such as those at present under discussion, the teaching of surface tension should not go as far as the introduction of surface tension as a quantitative concept. We do not wish to contradict this view, since the concept of surface tension does present some difficulty to beginners; also, avoidance of direct use of it enables one to bypass the controversy about whether this concept is best taught in terms of "force per unit length" or "free energy per unit area" (cf Chapin 1959, Yarnold 1960, Satterly 1961 and R.C. Brown 1967).

We recommend that in an elementary syllabus, the treatment of surface tension should concern itself mainly with the explanation of the experimental fact that the surface of a liquid has a contractile tendency. This aspect of surface tension should be explained in terms of molecular interaction. The discussion in Sections 5.4.19-20 has shown that errors are all too readily committed in such an explanation, and that the complexity of the phenomenon makes possible different, but perhaps equally permissible simplified accounts of the phenomenon. Concerning the first point, we note that while these errors are readily committed, they are as readily

/avoided,.....

avoided, provided the teacher does not insist on relying on treatments to be found in older textbooks. Concerning the second point, we make the following comments:-

(i) As a background to his understanding of surface tension, the pupil must have some picture of the liquid state in mind. This he should have gained from what he has already learnt about the liquid state, but it would be apposite for the teacher to stress here that it is meaningful to talk about a liquid structure, and that it is apparently reasonable to regard the liquid as passing through a succession of irregular structures of very fleeting duration. This will enable the pupil to relate his picture of a liquid quite simply to that of a solid, with which he should by now be thoroughly familiar.

(ii) The teacher should point out that where one has such a structure in which the molecules are vibrating randomly with amplitudes comparable with the separation between adjacent mean positions, any description of what is happening to any one molecule in terms of its interaction with neighbouring molecules will have to be an oversimplification of whatever is the actual situation, What is to follow, therefore, can only be a rough description capable of yielding some insight into the phenomenon but not very much more.

Of possible elementary approaches to the contractile tendency of liquid surfaces, that used by Freier is undeniably more elegant than that of Rogers, since it does not deal directly with intermolecular forces and sums up the situation with very simple statements which can readily be accepted as applying even to the complex actual situation of a liquid as well as to the idealized, "frozen" structure from which this approach is derived. The essence of this approach is that a surface molecule has fewer neighbouring molecules than an

/interior.....

interior molecule, so that when a new portion of liquid surface is created, work must be done to pull some of the neighbouring molecules away from the molecules which are to form the new surface. By the same token, energy will be released if the liquid surface contracts; if the principle is accepted that a system will, if given the chance, go over to the condition in which its potential energy is a minimum, then it will readily be understood that a small drop of liquid would so arrange itself that the potential energy associated with its surface would tend to a minimum. The molecules are continually in motion, so that the liquid certainly has the chance to assume the configuration appropriate to minimum potential energy.

The teacher wishing to plug all (or at any rate most) loopholes can point out that the above reasoning about the behaviour of surfaces applies to isothermal processes only (cf Section 5.4.21); that it is in such circumstances that the work done in expanding a surface becomes "extra potential energy". Since the cases of interest to pupils effectively concern isothermal processes only, the point can be seen to have practical relevance.

Besides its simplicity and elegance, this approach has the advantage that it does not run the risk of confusing the pupil about a supposed tension in the surface resulting from an attractive force normal to the surface: the surface tends to contract because it tends to the condition of minimum free ("potential") energy, and that is that. A seeming disadvantage is that it appears to offer no obvious mechanism by which the contraction is accomplished, whereas the traditional descriptions in terms of forces do provide such a mechanism. Certainly the explanation which says that the molecules in the surface are subject to unbalanced inward attractive forces

/gives.....

gives us a picture of surface molecules being pulled into the liquid; but since it also leads us to a picture of a liquid as being in a short-lived state of inward collapse, we cannot take it seriously. Rogers's picture of the surface molecules as pressing in upon the others does suggest a contracting surface without at the same time being physically unrealistic; as we have seen, it lends itself to a simple analogy of people trying to push their way into a crowd. However, the "energy" approach does not suffer as much by comparison as might have been supposed. In the "frozen" approximation considered by Freier, the molecules in the surface of the liquid as well as those in the interior are in potential energy "valleys", since they are in a condition of stable equilibrium. However, the valleys containing the surface molecules are not as deep as those for the interior molecules. If a surface molecule acquired enough kinetic energy to move out of its valley, it might find itself "trapped" in a deeper valley inside the liquid. (Compare its behaviour in this respect with a stone that has been ejected from a depression in a corrugated iron roof, only to land in a hole in the ground.) Since the molecules of a liquid have, on average, a great deal of kinetic energy, they are able to carry out this process of energy transition and sort themselves out into an arrangement of minimum free energy.

Because of the above-mentioned advantages, we are strongly inclined towards the use of the "energy" approach in introducing surface tension in the senior secondary school. We would not go so far as to suggest that all talk of intermolecular forces must be suppressed when discussing matters related to surface tension. On the contrary, pupils should be fully aware of the relevance of intermolecular forces to the phenomenon. If they are interested in the problem of the

/forces.....

forces exerted by or in surfaces, they can be given an outline of the explanation given in Section 5.4.21 of the existence of these forces. Pupils could also be invited to consider Rogers's "dynamic", "dense-gas" picture of molecules in liquids as an alternative simplification of the "true" picture.

One thing that should, if possible, be avoided is the description of surface tension as "the tendency for the surface of a liquid to behave as if it were a stretched elastic membrane." Gordon and Neser print this statement in bold type as if it were a definition of surface tension, whereas at best it is an analogy of strictly limited usefulness. The contractile tendency of a liquid is a simple fact, deducible from many simple demonstrations and experiments. It is this that should be made the basis of a discussion of "surface tension", bolstered as little as possible by the use of the "stretched-elastic-film" analogy.

5.4.23 Surface tension and the floating of objects on liquid surfaces

Under this heading, we consider briefly the floating of objects denser than water (say) on a water surface whose "skin" they have failed to break. Very often this situation is described as though "surface tension" forces are sufficient to support the floating object. (Cf. Freier (1965): "Razor blades and screen boats float on water because of surface tension forces".) As Chapin (1966) has recently pointed out, the major contribution to the supporting force often comes from buoyancy forces. He describes an experiment in which an aluminium disc of diameter 100 mm and thickness 1.57 mm was floated on distilled water at a temperature of 23°C. The maximum possible upwards force which surface tension could

/contribute.....

contribute in this case was equal to the circumference of the disc multiplied by the surface tension of water, i.e. about 0.020 N. The buoyant force on the disc was found by multiplying the area of the disc by the difference in hydrostatic pressure between its lower and upper surfaces. Chapin found this to be 0.320 N. (This meant that the depth of the lower surface below the level of the surrounding water was some two-and-a-half times the thickness of the disc.) The weight of the disc was 0.336 N, so that these figures support Chapin's claim. In his words, "The principal effect of the film is to dam back the water from around the disc and thus permit it to displace a greater weight of liquid".

CHAPTER 6: VIBRATIONS AND WAVES.

6.1 Vibrations

6.1.1 Introduction

"Vibrations" may be regarded at first as something of a "fringe" topic in a school syllabus in physics or physical science. Certainly a detailed treatment of it would be hard to justify at this level, since it would require the use of mathematics beyond both the knowledge and the capabilities of the average (South African) school pupil. But the idea of vibration is inevitably used in many places even in an elementary physics course; we talk about vibrations when discussing the principles of time measurement, we use vibratory devices for timing in fundamental mechanical experiments, and we attribute the internal energy of solids and liquids to the kinetic and potential energy of randomly vibrating "molecules". Above all, we study periodic waves, many of which have their origin in vibrating sources of one kind or another. It is therefore fitting that some time should be devoted to vibrations as such. In the comments accompanying the Common Basic Syllabuses, it is stated that "Vibrations are included because it is felt that they offer a simple introduction to waves". By this it is presumably meant that certain concepts used to characterize periodic waves can be introduced by way of the parallel concepts used for vibrations, viz. displacement, amplitude, period, frequency and phase. The study of vibrations is indeed useful from this point of view; but its intrinsic interest is considerable, and the teaching of vibration should be regarded as being important both for its own sake and for its links with periodic waves.

/6.1.2.....

6.1.2 Examples of vibrating systems

To arouse interest in vibration, reference could be made to its importance in everyday life and in various branches of physics. Attention could be drawn to both the pleasurable and the annoying aspects of vibration in our everyday lives; on the one hand, we have vibrations as the sources of musical sounds (including the vibrations of our own vocal cords), on the other, we have annoying "sympathetic" vibrations produced, for example, by a lorry rumbling past one's house and causing parts of it to vibrate, or by the engine of a motor car which, when "turning over" at a certain number of revolutions per minute, may set certain parts of the car in vibration.

To illustrate the interest of the physicist in vibration, one could mention that any molecule consisting of more than one atom is a system capable of vibration, and that a study of the energy associated with this type of vibration can tell us a great deal about how the molecule is put together. Also, molecules as a whole carry out vibrations about their mean positions in solids and liquids, and a study of these vibrations is important for an understanding of the thermal properties of solids and liquids.

Only two or three simple examples of vibrating systems need be discussed in detail. The simple pendulum, an object hung from a spring and a hacksaw blade or a metre stick clamped at one end will do very well for this purpose. The first two of these are particularly convenient, since the repetition period of the motion can be made long so as to facilitate the observation of the characteristics of simple vibrations. In the case of the third, for which the period of the motion could be very short, a stroboscope could be used to "slow down" the motion and show that the characteristics of rapid vibrations are the same as those of slow ones. This might seem

/a.....

a trivial point to the sophisticated; it need not necessarily seem so to the beginner.

6.1.3 Physical characteristics of vibration

From a study of simple vibrating systems such as the ones mentioned above, the pupil should learn that vibratory motion is periodic, and he should appreciate in a general way the manner in which the velocity and acceleration of a (simple-harmonically) vibrating system vary with time during a cycle of the motion. He should realize from Newton's Second Law that the motion of a simple vibrator is subject to a restoring force whose magnitude might well be proportional to the displacement from the mean position. The terms amplitude and frequency should be introduced in reference to simple harmonic motion. This material should give no trouble in presentation; elsewhere (Helm 1963) the present author has given a more detailed account of how it might be presented.

It would be advantageous to introduce the concept of the phase of the motion, even if only in a semi-quantitative way. All that is needed is that pupils should be able to think of "phase" in this context as corresponding to the stage of the motion of the system; if, that is to say, two similar vibrating systems are simultaneously in the same stage of their vibratory motion (e.g. both passing through their mean positions in the same direction) they are said to be "in phase". If they are not in step, there is said to be a "phase difference" between their motions, meaning that at the same instant, the phases of their motions are different (i.e. they have reached different stages of their motion), or that they attain the same stage of their motion at different instants. If they are exactly out of step, i.e. they pass through their mean positions in opposite directions at the same instant,

/they.....

they are said to be exactly out of phase. These points can easily be illustrated with a pair of similar pendulums or springs and should give no trouble in presentation.

6.1.4 The energy of vibrating systems

The energy relations in a harmonic vibration are worth closer attention, provided that the pupil knows something about the potential energy associated with a system undergoing elastic vibration. It is very easy to show graphically that the work done in stretching a spring by an amount x is proportional to x^2. Using this result, we are able to say at once that the energy associated with a simple harmonic vibration is proportional to the square of the amplitude. In displacing the system initially from its equilibrium position in order to set it in motion, we give it an amount of potential energy that is proportional to the square of this displacement; in the absence of frictional losses, this initial displacement remains the amplitude of vibration of the system, and the energy initially supplied as potential energy is the energy of the vibration. (Those who would prefer to discuss this situation with reference to the simple pendulum will find it easy to show that its energy of vibration must be proportional to the square of the amplitude; one need simply use the well-known geometrical result that when two chords of a circle, AB and CD, intersect in E, AE.EB = CE.ED, to relate the height through which the pendulum swings to the square of its amplitude.)

It is also possible to show by means of a simple argument that the energy of a simple harmonic motion must be proportional to the square of the frequency. Suppose we have a particle vibrating with amplitude A and frequency f. The average speed of this particle during one complete vibration is given by the distance travelled (4A) divided by the time

/taken....

taken $(\frac{1}{T})$; i.e. the average speed of the particle must be 4Af. This shows that the average speed is proportional to the frequency. We can reasonably conclude that the maximum speed must be proportional to the frequency. The maximum kinetic energy, which is equal to the total energy of the vibration, is proportional to the square of the maximum speed; hence the total energy is proportional to the square of the frequency. From this result and the previously-obtained one concerning the dependence on amplitude, we conclude that the energy of a simple harmonic vibration is proportional to the square of the amplitude and the square of the frequency.

To complete this discussion of the energy associated with simple harmonic vibration, one would have to say something about the damping of the vibrations of any real macroscopic vibrating system, such as those shown to the class. By now pupils would readily enough ascribe the diminution of amplitude of the vibrations as time goes on to the effect of friction, and account for the mechanical energy lost by the vibrating system as "heat".

6.1.5 Forced vibration and resonance

The essential characteristics of forced vibration and resonance that are likely to be of immediate interest to pupils are easily demonstrated with the aid of a simple pendulum, or an object hanging from a spring or rubber band. If the point of support is held in the hand and subjected to vibration at suitable frequencies, the main features of forced vibration and resonance are easily observed.

6.2 Waves

6.2.1 Waves in the senior secondary school

In the senior secondary school it is possible to treat waves in a more quantitative manner than was possible in the junior school. This higher-level treatment should begin with a revision of the ideas acquired in earlier years.

/In.....

In particular, the concepts of wavelength, frequency and wave velocity should be revised ("wavelength" can now be more formally defined; cf. the definition quoted in Section 3.17. 3), as also the graphical representation of transverse and longitudinal waves. The amplitude and frequency of a periodic wave should be related to the amplitude and frequency of the vibrating source generating it, e.g. by consideration of transverse waves sent down a long spring or string from a vibrating source at one end. In such a case it would be obvious that in the absence of the damping, the amplitude of the wave must be the same as the amplitude of vibration of the source. It will also be obvious that one complete vibration of the source will produce one complete "wave" (i.e. a kink to one side followed by a kink to the other), and that the number of complete waves passing a given reference point in the medium per second must be equal to the number of complete waves produced per second, i.e. that the frequency of vibration of the source is the same as the frequency of the waves.

The principal new ideas to be introduced in this section are the polarization of transverse waves, the Principle of Superposition and its consequences, and dispersion.

6.2.2 Polarization

This topic is not always to be found even in modern physical school science syllabuses; it is omitted from both the P.S.S.C. and the Nuffield Physics courses, for example. Yet the idea of polarization of a transverse wave is not difficult to grasp, and since elementary discussions of transverse waves make use of waves polarized in a given plane, one might as well introduce the term that describes this state of affairs and follow this introduction with an elementary discussion of polarization and extinction. To illustrate these ideas, one can make use of the well-known

/arrangement.....

arrangement using a stretched string passing through a pair of cardboard or hardboard sheets with long slots cut in them (cf Prinsloo et al 1968). Randomly polarized waves on the string incident on the first slot will be "polarized" by the first slot and "analysed" by the second; if the two slots are parallel, the polarized waves incident on the second slot pass through unhindered, whereas if the two slots are at right angles to each other, "extinction" takes place. There are certain difficulties about the use of the slot-and-stretched-string set-up as an analogy for the polarizing action of a sheet of Polaroid on light waves; these will be discussed in Section 8.10.

6.2.3 Principle of superposition

As applied to waves, the Principle of Superposition states that when two or more waves pass over a point simultaneously, the resultant disturbance at that point is simply the algebraic (vector) sum of the disturbances that would have resulted from each of the individual waves. This Principle applies only if the medium is "linear" in its behaviour, i.e. (in the case of mechanical waves) the "restoring force" exerted on any part of it is proportional to the magnitude of the displacement of that part of it from its mean position. If the waves passing through the medium give rise to very large displacements, the behaviour of the medium may no longer be linear and the resultant disturbance at a point may not be accurately given by the Principle of Superposition. These are points for the teacher rather than the pupil to note; for the pupil, the Principle should remain essentially a fact of experience, although it will do him no harm to know that there are limitations on its range of applicability.

/We.....

We think it a good plan for pupils to begin their study of the superposition of waves with observation of the superposition of transverse pulses on a stretched spring (cf PSSC 1960). Such experiments are best done with a "Slinky" in view of the relatively slow speed of waves on it, but will only really be successful if the "Slinky" is suspended in the manner described earlier. Failing this, the experiments can be done with a round-wire spring, in which case it will be more difficult to see what happens when the pulses momentarily overlap. Pupils should see what happens when pulses are sent down the spring simultaneously from opposite ends, both when the pulses are on the same side and when they are on opposite sides of the spring.

If these experiments begin with the sending of similar pulses down the same side of the spring, it is possible that pupils may interpret what they see as being the result of reflection - the two pulses apparently collide and bounce apart, the reflection apparently taking place without change of phase. Pupils can then be asked to check this hypothesis; it might occur to them that a way to do this would be to send down similar but opposite pulses from opposite ends of the spring, or that they could send down pulses of different shape from opposite ends, or both. At the end they should be convinced that the only consistent interpretation they can give to their results is that the pulses pass through each other when they meet, and that in the region of overlap the pulses "add" according to the Principle of Superposition. (The latter Principle should not be formally stated as yet.) The famous sequence of PSSC photographs (PSSC 1960) showing successive stages of the process of superposition can be used to lend support to these ideas.

/The.....

The logical next step is the study of the superposition of periodic waves instead of pulses, and it is also logical to carry out this study first for periodic waves on springs. When the pupil was first shown the propagation of periodic waves on a "Slinky" spring, he must have noticed that interesting things started to happen when the reflected waves returned from the far end while incident waves were still travelling down. This is the obvious way in which to introduce the idea of an "interference" pattern, with its alternating points of maximum and minimum disturbance.. There is no need to become involved at this stage with a discussion of "standing waves" and so lay oneself open to doubts as to whether "standing waves" should be regarded as the superposition of two progressive waves travelling in opposite directions or simply as the natural modes of vibration of the system concerned (cf Rekveld (1958a), who takes issue with the "superposition of waves" approach). Before there is time for a complete "standing wave" to build up - the frequency may not be right for this anyway - the pupil sees an interference pattern as the reflected waves start to overlap with the incident waves, and this is all that he need note for the moment. The existence of points of zero and maximum disturbance should not surprise him overmuch; if he now accepts the validity of the Principle of Superposition, he can account for this phenomenon in terms of overlapping "crests" or "troughs" or overlapping "crests" and "troughs". What is less obvious is that these points should be stationary. Rather than pursue this point further at this stage, the pupil should now study the corresponding phenomenon with waves on the surface of water in a ripple tank. A P.S.S.C. "wave generator" with two sources attached to it produces very successful interference patterns. The most significant feature of these

/patterns.....

patterns is the existence of nodal lines which maintain constant positions on the surface of the water. The pupil should investigate the manner in which the number and spacing of these lines depends on the spacing of the sources and on the wavelength of the waves.

The analysis of the appearance of these interference patterns can be done by the standard technique of drawing sets of alternating crests and troughs forming circles centred on each source, and showing that the lines that pass through points of overlap between crests and troughs look just like the nodal lines of an actual interference pattern. From this emerges the most important quantitative fact for the pupil to appreciate about such interference patterns - that the difference in the distances from any point on a nodal line to the two sources must be equal to an odd-half-integral number of wavelengths. In the last resort, it is much more important for the pupil to grasp this fact than for him to be able to go through the motions of deriving the familiar approximate formula relating the separation of the nodal lines to the wavelength of the waves and the separation of the sources. At some stage he should be taken through the derivation of this formula, but he should not be expected to reproduce it in examinations.

So far the study of interference patterns has been confined to those produced by two sources <u>in phase</u> with each other. It is most important for the later study of the interference of light waves for the effect of difference in phase between the sources on the interference pattern to be demonstrated. A teacher with a mechanical bent would find the task of devising and constructing an "interference pattern generator" using a pair of out-of-phase sources an interesting challenge, well worth the effort. The interference

/pattern.....

pattern produced by this device will have its nodal lines displaced relative to those for the "in-phase" pattern.

It is possible to show the effect of sudden changes in phase difference between the sources by using a pair of vibrating wires attached to beads as the sources. An experiment of this kind is described in the P.S.S.C. Laboratory Guide.

It is generally acknowledged that the word "interference", whose use in this context we seem to owe to Thomas Young (cf Magie 1935), is an unfortunate name for the phenomenon under discussion above, since the waves do not "interfere" with one another in the literal sense of the word. On the other hand, experience in teaching "interference" to university students and of being taught the subject himself leads the author to believe that there is little danger of confusion resulting from the use of the word "interference" as a name for the phenomenon and as an adjectival noun to describe the patterns obtained. One comes to accept that the word is not meant to be taken literally, and all is well. Perhaps it would be unwise to use the term freely in the early stages of teaching the topic, but it will have to be introduced in the end, as it is the accepted scientific term.

6.2.4 Diffraction

The main features of the diffraction of waves at an aperture or by an obstacle are easily demonstrated with a ripple tank. The P.S.S.C. recommend the use of blocks of paraffin wax as obstacles; a pair of these suitably spaced from each other would make up an aperture of the desired size. Pupils should observe what happens when periodic straight waves are incident on apertures of various widths. They should try the effect both of keeping the wavelength constant and varying the width of the aperture, and of keeping the

/width.....

width of the aperture constant and varying the wavelength. Their investigations should establish the following points:-

(i) In general, straight wave-fronts become curved on emergence from an aperture, i.e. diffraction occurs in these circumstances.

(ii) The bending of the wave-fronts is most marked when the dimensions of the aperture are of the same order of magnitude as or less than the wavelength.

(iii) Diffraction effects are small when the dimensions of the aperture are large compared to the wavelength.

It is to be hoped that neither teachers nor examiners will require of pupils that they should be able to "define" diffraction. It is certainly possible to describe diffraction in a single sentence, but the results as a rule are only illuminating to those who already understand the subject. Here are some randomly-selected quotations to illustrate this point: "Diffraction is the behaviour of waves that pass through apertures or by edges of screens " (Andrews 1960).

"The general term for the phenomena...in which one is concerned with the resultant effect of a large number of waves from different parts of a source, is diffraction" (Sears and Zemansky 1964).

"... after passing through the opening not all of the wave propagates in its original direction. Part of it is bent. This phenomenon is called diffraction." (PSSC 1960, 1965).

Of these the second is unsatisfactory in any case because it is based on a particular theory of diffraction (see below). The first is unexceptionable as a "definition", but will be found unsatisfying by those who would feel that it ought to go on to give some indication of what the "behaviour" is. The third more nearly represents the kind of thing one might expect from a pupil asked to describe diffraction, viz. a

/statement.....

statement of what he has seen happening to waves in certain circumstances.

It is regrettably necessary that the phenomenon of diffraction should remain largely unexplained from the pupil's point of view. Some might be surprised at this, pointing to the PSSC example of the formulation of a simple theory of diffraction by a slit (PSSC 1960, 1965) In the PSSC text-book the experiment is described of sending out periodic surface waves from a line of in-phase sources spaced across a "slit"; photographs are shown of the resulting wave-pattern and of the diffraction pattern of a single slit, from which it is apparent that the two patterns are substantially identical. Or rather, one half of the multiple-source pattern is identical with that of the single-slit, for the multiple-source pattern is double-sided (i.e. symmetrical about the line of sources), whereas that of the single slit is not. (Oddly enough, this fact is not mentioned in the PSSC text). So the multiple-source theory, while successful in predicting the shape of the single-slit pattern, also predicts a duplicate version of this pattern in a region where it does not exist. How is this situation to be handled by the teacher? One way is to discuss the whole problem "from the textbook", using the PSSC photograph in which the "embarrassing" portion of the multiple-source pattern is conveniently suppressed, and hope against hope that brighter pupils in the class will not ask awkward questions about why there is no backward-spreading pattern from the single slit. This is surely neither safe nor wise. Another way would be to do the experiment - in which case the realities of the situation would at once be apparent to pupils. There is, unfortunately, no simple way of explaining away the difficulty that is encountered here; the best that can be done is to say that the

/absence.....

absence of the backward waves can be explained by a more sophisticated theory.

The "multiple-source" theory of diffraction by a single slit is none other than an application of what has been called the Huygens-Fresnel principle (cf Towne 1967), according to which each point of a wavefront acts as a source of secondary wavelets, giving rise to a resultant disturbance at any point in advance of the wavefront, compounded from the superposition of contributions from all these secondary sources. In the nineteenth century this principle was successfully applied by Fresnel to the solution of many fundamental problems concerned with the diffraction of light; its use for the same purpose in intermediate optics teaching is now taken for granted, and it will no doubt continue to be used in this way for some time to come. Yet there have always been difficulties attached to its use in diffraction theory, of which the need to suppress the "back wave" (see above) is only one. For example, what is the nature of the supposed secondary sources? Do they and the secondary wavelets to which they are supposed to give rise have any real existence? A detailed analysis of this aspect of the matter shows that the secondary sources have to have physically implausible characteristics, and this presumably rules out the possibility both of their existence and that of the waves alleged to originate from them (cf Mendoza (1953) for a simple discussion of this point). The truth regarding the application of the Huygens-Fresnel principle to diffraction is the following: that for this purpose it is "a convenient mathematical device which gives the right answer (for no very obvious reason) under conditions that are often satisfied in optical [and other] experiments" (Mendoza 1965). The true mechanism of diffraction is concerned with the interaction between the incident waves

/and.....

and the obstacle on which they are incident; putting it another way, we can only give a complete description of diffraction in any given case if we set about solving the appropriate wave equation putting in the boundary conditions. Commenting on the present position in diffraction theory, Andrews (1960) hopefully remarks "We may expect the subject of diffraction to be in a more satisfactory state within the next generation". Whatever may happen in the next generation, we feel that the present-day teacher will not help his pupils very much by giving them the Huygens-Fresnel theory of diffraction by a slit, unless he does this in the spirit of "let's try this theory and see how far it can take us."

6.2.5 Reflection and refraction of surface waves in a ripple tank

It is well known that Huygens's principle applied to the reflection and refraction of plane wavefronts very easily yields the results that the angle of incidence is equal to the angle of reflection and that the ratio of the sines of the angles of incidence and refraction is equal to the ratio of the velocities of the incident and refracted waves. Clearly the use of the principle here is as lacking in physical justification as it is in the case of diffraction. However, for mechanical waves incident on a boundary between two media, we can evolve a physical picture for the generation of reflected and refracted wavefronts which, when treated mathematically, gives us a formal description closely similar to that obtained from Huygens's principle.

As background for this treatment the pupil needs to know a little more about the reflection and transmission of waves than he learned in the junior years. In his earlier work on the reflection of pulses sent down stretched springs, he only studied reflection at a "fixed" boundary. This study should

/now.....

now be extended to take into account reflection from a "free" end of the spring (simulated by attaching one end of the spring to a length of string), and reflection and transmission from a boundary between two different springs. For the latter purpose a long round-wire spring and a "Slinky" could be used; the PSSC Teachers' Resource Book and Guide (Part Two) discusses this demonstration in some detail. The mechanism by which these reflected and transmitted disturbances are initiated is not difficult to understand if the forces acting at the boundary when the incident disturbance arrives there are considered.

Let us now turn to the analogous problem for surface waves in a ripple tank. The pupil knows already that when straight or circular waves are incident on the boundary between two regions of different depths, reflected and refracted wavefronts appear. In view of his observations on the corresponding phenomenon on a pair of coupled springs, it should not surprise him that reflected and transmitted disturbances should start out from the boundary between the two regions. A little further thought on the matter might convince him that as the incident wavefront sweeps across the boundary, it will set off a sequence of semicircular ripples in each region, starting out from successive points touched by the incident wavefront, and that the overall effect is simply to produce the observed reflected and transmitted wavefronts. "But", it may be argued, "this is just the application of Huygens's principle all over again". So it would be if we failed to specify a mechanism for the production of the reflected and transmitted ripples. Fortunately a mechanism of sorts can be specified (cf Arons 1965). As a matter of experimental fact, an object moving across the surface of water with a speed greater than the speed with which waves

/will.....

will travel on it will generate a bow wave. The details of the mechanism are complicated, but the phenomenon is familiar enough; moreover, it is easy to show that the angle between either segment of the bow wave and the line of travel of the source is given by the equation $\frac{v_i}{v_s} = \sin\theta$, where θ is the angle in question and v_i, v_s are respectively the speed of propagation of the waves and the speed of the "source". Now, returning to the problem of reflection and transmission of water waves, it is also easy to show that the speed with which the incident wavefront sweeps across the boundary (i.e. the speed of travel of the point of intersection of wavefront and boundary) is in excess of the wave speed in the region on either side of the boundary. Therefore we can interpret the reflected and transmitted wavefronts as bow waves generated by the advance of the incident wavefront, and use of the above-quoted equation for these wavefronts readily yields the results that the angle of reflection is equal to the angle of incidence, and that the sines of the angle of incidence and refraction are in the ratio of the wave velocities in the two regions.

The use of bow waves to explain reflection and refraction of surface waves may seem alarming at first. But as Mendoza (1965) has pointed out, there is a good case for including a very brief treatment of bow waves and shock waves in a school syllabus; only those pupils who have never been near a large river, a lake or the sea will never have seen a bow wave, and in future years an increasingly large number of pupils will encounter the corresponding acoustic phenomenon in the form of "supersonic bangs". It is realistid to recognize the existence of the phenomenon in planning a syllabus, and far from unreasonable to expect pupils to understand the very simple physics, geometry and trigonometry required to prove that $\frac{v_i}{v_s} = \sin\theta$. At any rate, the "bow wave" approach is one way in which a teacher can tackle the problem of

/elucidating.....

elucidating the behaviour of water waves. If he prefers not to use it, he can fall back upon the simple PSSC (1960, 1965) proof, applicable to periodic waves, that $\frac{\sin i}{\sin r} = \frac{v_1}{v_2}$. This makes use of the appearance of the waves when periodic "straight" waves are incident on a boundary. In each region the wavefronts are parallel to one another and separated from one another by the appropriate wavelength; at the boundary corresponding wavefronts meet. From these facts it follows at once that $\frac{\sin i}{\sin r} = \frac{\lambda_1}{\lambda_2}$ (where λ_1, λ_2 are the respective wavelengths in the two regions) and, since the frequency is the same, that $\frac{\sin i}{\sin r} = \frac{v_1}{v_2}$.

6.2.6 Energy transmission by waves

Since one of the reasons why waves are studied is because of their ability to transmit energy, pupils should have some idea of how energy is transmitted by waves and on what factors the rate of energy transmission depends. The mechanism of energy transfer is best explained with reference to a specific example, such as waves propagated down a stretched string or spring. If the waves are generated by shaking one end of the spring up and down, energy is fed in at that end because a force has to be applied (in opposition to the restoring force supplied by the tension) to displace the spring, and the point of application of the force is moved. So work is done by the hand shaking the spring, and energy is constantly being transferred to the spring. As the waves travel down, one element of the spring exerts a force on the next and does work on it; and so the energy taken up from the source is transferred along the spring.

In general, the energy transferred to a given element of spring will, at a given instant, be part kinetic and part potential. To discover the factors on which the rate at which this energy is transferred past a given point depends, is not as

/straightforward.....

straightforward as might first appear. Suppose we focus attention on a particular vibrating element of the spring. We have seen earlier that the energy associated with this vibrating element is proportional to the square of its amplitude of vibration and of its frequency. We can therefore say that in one complete wavelength of the wave, there is a total amount of energy, potential plus kinetic, which is proportional to the total mass of this length of the spring multiplied by the square of the amplitude and the square of the frequency. But the total mass is proportional to the wavelength, which is inversely proportional to the frequency (assuming that the properties of the spring are such that the wave velocity does not depend on the frequency); hence the total mass of one wavelength is inversely proportional to the frequency, and the total energy associated with one wavelength of the spring is proportional to the first power of the frequency and the square of the amplitude. Before the waves started to travel down the spring, no part of it possessed this energy, which is supplied by the waves alone. The time taken for this energy to be supplied to one wavelength of the spring is simply equal to the reciprocal of the frequency (i.e. the time taken for the wave to travel one wavelength).

We can now group these facts together to find the factors on which the rate of energy transfer by waves depends. We have established above that as the waves travel down the spring, the amount of energy delivered to any wavelength's-worth of the spring is proportional to the product of the frequency and the square of the amplitude; and we have noted that it takes a time equal to the reciprocal of the frequency for this energy to be delivered. Hence the <u>rate</u> at which is energy is delivered to any part of the spring is proportional to the <u>square</u> of the frequency and the <u>square</u> of the amplitude.

/From.....

From the pupil's point of view, it is the dependence of the rate of energy supply on the square of the amplitude that is most important.

For waves propagated in three dimensions (e.g. sound or light waves), the quantity that is used to characterize the rate at which energy is propagated is the intensity, defined as the rate at which energy is propagated per unit area of an imaginary surface drawn normal to the direction of propagation. This concept is so useful in discussing matters related to waves that it should certainly be explained to pupils, who ought also to be able to appreciate that the intensity of waves from a small source emitting waves equally in all directions must fall off inversely as the square of the distance. It should not be difficult for them to accept the generalization from the arguments given above that the intensity is proportional to the square of the amplitude.

6.2.7 Sound waves

Under this heading we take up certain topics concerned with sound that are best dealt with in the senior school. To add to pupils' conviction that sound is indeed a form of wave motion, they should have demonstrated to them that sound waves may be superposed to produce a detectable interference pattern. A pair of similar loudspeakers placed one at each end of the demonstration bench and connected to an audio-frequency oscillator giving out a signal of around 800 Hz will do admirably for this purpose.

The superposition of sound waves of slightly differing frequencies to form beats is very easy to demonstrate, and if the equipment used for the above-described demonstration is available, plus a microphone, extra oscillator and an oscilloscope, it will be possible to show the characteristic "waveform" of the combination of sounds on the oscilloscope. Provided that the oscilloscope has a low frequency response

/that.....

that will enable it to cope with signals of frequency 20-30 Hz, no difficulty should be experienced in obtaining the desired picture. The two oscillators should be set to produce sounds of frequency of about 1000 Hz, differing by an amount of some 20-30 Hz, and the oscilloscope should be synchronized to the beat frequency. No attempt need be made to account in detail for the fact that the beat frequency is equal to the difference in frequencies. The keen pupil can amuse himself by proving this result graphically for a particular case.

(The following points about beats would be worth noting by the interested teacher. First, it is the intensity of the sound which fluctuates at the beat frequency; the amplitude goes through a complete cycle of variation, from maximum positive to maximum negative to maximum positive again, at a repetition rate equal to half the beat frequency. The mathematics of superposition in this case can easily be used to prove this. Second, the production of beats is in no way dependent on any non-linearity in the propagating medium, although the ability of the ear to hear the sound as having a single pitch and varying loudness may well depend on some non-linearity in its behaviour. It is one of the curiosities of the behaviour of the ear that while it is able to a remarkable extent to analyse the sounds it receives into their pure-tone components, it does not do so for a pair of sounds of neighbouring frequency.)

Another application of the superposition of sound waves is the formation of "standing waves" in air columns. This can be introduced by analogy with standing waves on stretched strings or springs, which the pupil ought to see first.

/The.....

The Common Basic Syllabus for Physics specifies "relation between frequency, length of string, tension and mass per unit length" as one of the topics to be considered under the heading of "sound". The motivation for this inclusion is presumably that it is of interest to a pupil to know why the frequency of vibration of a stretched string rises when its tension is increased or its length reduced. The latter point is easy enough to establish from what the pupil would already know about standing waves on strings, but the former requires knowledge of the relationship between the wave velocity and the tension and mass per unit length of the string. A simple derivation of this can be given (cf Rogers 1960), but its inclusion would not be justified in an already crowded course. Thus we have the unsatisfactory situation that if the appropriate formula is to be used, it will have simply to be given. We feel that in place of this it would be better not to introduce this formula, but simply to suggest to pupils that it is plausible that to increase the tension in the string would tend to increase the wave velocity, while to increase the inertia of the string would tend to reduce it. Sacrifice of the formula would mean sacrifice of experiments based on use of the sonometer, but perhaps this loss will be regretted much less today than it once might have been.

If it is sought to include other topics in the syllabus under the heading of "sound", some consideration might be given to the inclusion of the Doppler Effect, which applies to all kinds of wave motion and has a number of technological and scientific applications, ranging from the measurement of speed of motor-cars to the estimation of the speed with which distant stars are receding. The treatment need make no reference to formulae, but simply explain the effect of the motion of the source on the wavelength of the waves emitted,

/and.....

and the effect of the motion of the observer on the frequency of the waves received. The former point can be beautifully illustrated on the ripple tank; if the required auxiliary apparatus is difficult to set up, an excellent film-loop (Ealing A80-2371/1) can be used instead.

CHAPTER 7: ELECTRICITY AND MAGNETISM IN THE SENIOR SECONDARY SCHOOL

7.1 Some problems in the teaching of electricity and magnetism

When in an earlier section we discussed an introductory approach to electricity in the junior secondary school, we were in an area relatively free from controversy. When we come to consider the teaching of electricity and magnetism in the senior secondary school, we are concerned with what is perhaps the most controversial aspect of present-day teaching of elementary physics. Opinions differ widely over the question of what method of presentation of these topics should be followed. At the centre of this controversy used to stand the fundamental question of what system or systems of units to adopt - MKS - ampere (or, as we should now call it, SI) or the traditional triumvirate of electrostatic, electromagnetic and "practical" units, or perhaps even some quite different system of units with its own particular advantages. (The PSSC course uses "nature's own" unit of charge, the electronic charge, in its earlier stages and later works its way round to SI units.) With SI units now achieving "official" status all over the world, one may reasonably hope that the controversy over units will gradually die out. But there remains much disagreement about major and minor details of presentation. Should a senior school course in electricity begin with electrostatics or with current electricity? Should quantitative (point-pole) magnetostatics form part of the course? How, in treating current electricity, ought one to introduce the definition and measurement of current - treat the ammeter as a "black box" whose method of working would be explained later (cf. our treatment in Section 3.15.4), or attempt a quantitative account of the working of this instrument, based on a definition of current

/in....

in terms of magnetic effects? Is the ammeter to be calibrated by means of electrochemical or magnetic effects? Must the ampere and the ohm be treated as fundamental units from which the volt can be derived, or should the ohm be derived from the ampere and the volt? To what extent and in what manner should modern ideas be incorporated into the presentation of the subject? How are static and current electricity to be bridged? These are only some of the matters on which the teacher will find little or no unanimity of opinion if he consults syllabuses, textbooks or journals devoted to the teaching of the subject.

7.2 Units

Now that SI units are the officially recommended units for scientific and industrial use in South Africa, there may seem to be little point in devoting a separate section to a brief discussion of electrical units. But since this matter once aroused bitter controversy, particularly in Great Britain, it is appropriate to consider whether the use of SI units is likely to create any difficulties in the teaching of electricity and magnetism in South African schools.

The vehement opposition in Great Britain to the use of SI units in the teaching of electricity was seen at its most extreme in the notorious remark by C.J. Smith in the preface to his book, "Electricity and Magnetism", that "It will be found that I have made no use of a certain system of units which finds favour with some" (C.J. Smith 1959). Evidently Dr. Smith could not even bring himself to mention by name the system of units which has since received official approval from the Conférence Général des Poids et Mesures (cf Kaye and Laby 1966). The publication in 1954 of the report by the Science Masters' Association on "The Teaching of Electricity

/with.....

with Special Reference to the Use of M.K.S. Units" (Science Masters' Association 1954) sparked off a heated debate in the pages of various British journals, echoes of which continued to be heard for some time (cf Macfadyen 1954, 1958, Lipson 1957a,b for representative statements of the "opposition's" point of view). Unfortunately the S.M.A. Committee responsible for drafting this Report committed itself to a teaching syllabus which was too radical to be acceptable to many who were otherwise in sympathy with its aims. One regrettable result of this was that the controversy about the choice of a system of units became inextricably entangled with a controversy on the relative merits of different teaching approaches. In fact, it is perfectly possible to teach via a traditional approach and still use SI units, while Purcell (1965) has brilliantly demonstrated that it is possible to present electricity and magnetism in a stimulating "modern" way while using "Gaussian" CGS units. In Britain, the situation has now changed a good deal. The successor to the Science Masters' Association, the Association for Science Education, has published a new report (Association for Science Education 1966) more moderate in tone and approach than the old and likely to win much wider support. British journals devoted to science teaching reflect a much greater degree of acceptance of the once-controversial system of units than was evident even a few years ago, and the official adoption of SI units seems to have been calmly received. (Cf Spurgin 1969 for an example of a favourable reaction from an experienced British teacher.)

In the context of South African science teaching, the choice of a suitable system of units was less agonizing than it otherwise might have been. This is because the level of the South African Matriculation examination falls considerably short of the British A-level examination, so that the

/teaching.....

teaching of electricity in South African schools must needs treat the subject in far less detail than is customary in British A-level teaching. For this reason, it avoids just those complications which constitute the most controversial aspects of the Gaussian -CGS-versus- SI controversy, viz. the introduction of "rationalization", the explicit use of the concepts of the "permittivity" and "permeability" of free space, and the relationships between the $\vec{E}$ and $\vec{D}$ field vectors in electricity and between the $\vec{B}$ and $\vec{H}$ field vectors in (electro-)magnetism. Thus even before the official adoption of SI, it was possible for South African examining bodies to recommend the use of SI units without fear of rousing a storm of protest. Both the new Joint Matriculation Board Syllabuses of 1965 and the more recently published Common Basic Syllabuses specified the use of SI units without, so far as is known, provoking any adverse comment on this score.

The advantages of the adoption of a single system of units in the teaching of any subject need no stressing, and the limited scope of the syllabuses under discussion makes it possible for these advantages to be fully exploited without there being any serious disadvantages. In current electricity, the SI units of current, charge, potential difference, resistance and power are identical with the so-called "practical" units, i.e. the units in which these quantities are normally measured. In electrostatics, use of the coulomb as the unit of charge means that the right-hand side of the equation stating Coulomb's Law incorporates a proportionality constant whose value is very nearly 9×10^9 newton-metre^2/coulomb^2; this should not lead to any serious practical difficulties, as it simply requires the use of one extra number in inverse-square-law and certain electric-field

/calculations.

calculations. This constant need not and should not be written as $\frac{1}{4\pi\varepsilon_o}$ in the context of the present syllabus, as it would bring with it no advantages whatever; if it is desired to replace the numerical constant in the Coulomb's-Law equation with a symbol, this can simply be "k" or, if preferred, "k_e". (Our recommendation here (cf Helm 1963) is essentially the same as that of Sears (1963b).) Calculations of magnetic induction ($\vec{B}$) require the use of formulae of the same form as are encountered in the electromagnetic system of units, with the inclusion of a multiplying constant of value 10^{-7} newton/ampere^2. Again, within the limits of the syllabus, no advantage would be gained by writing this constant as $\frac{\mu_o}{4\pi}$; it can be called "k_M" if it is desired to use a symbol for it.

7.3 Some features of the proposed treatment

One of the major objectives of the more formal treatment of electricity and magnetism appropriate to the senior secondary school is the introduction and use of vector quantities to characterize the electric and magnetic fields. The language of modern physics is, to a large extent, the language of fields, and the context of electricity and magnetism provides pupils with their most extended opportunity of learning the fundamentals of this language. The electric field vector is most conveniently introduced with reference to charges at rest, so our treatment below (and the Common Basic Syllabuses) begins with a study of electrostatics. For reasons we shall set out below (Section 7.15.2) we believe that the magnetic field vector is best introduced in terms of the forces acting on current-carrying conductors or moving charges.

The treatment of electricity and magnetism laid down in the Common Basic Syllabuses and considered below deals only

/with.....

with charges, conductors and magnets in vacuo. This means that the pupil will not be in a position to design capacitors, electromagnets or transformers, or give a quantitative account of what happens, electrically and magnetically speaking, inside dielectrics or magnetic materials. On the other hand, he will be able to understand the fundamentals of the behaviour of charged particles in molecules, atoms, cathode-ray tubes and particle accelerators, so that he is allowed entry into some of the most interesting areas of modern physics. Also, as has already been pointed out, exclusion of dielectric and magnetic materials from consideration does permit a much simpler presentation of the subject than would otherwise be possible.

7.4 Coulomb's Law

7.4.1 Should quantitative electrostatics start from Coulomb's Law?

Traditionally, quantitative electrostatics begins with Coulomb's Law. From time to time, objections have been raised against this approach; since Rekveld (1958b) has taken the trouble to state his objections in some detail, we shall discuss them briefly before stating our own reasons for adhering to tradition in this respect.

First, Rekveld points out that the accuracy with which Coulomb's Law can be established in a direct experiment (e.g. Coulomb's own) is not impressive: "it does not seem very likely that we could suggest any convincing evidence for this law by means of a classroom performance". This, he feels, makes the use of Coulomb's Law as an introductory law questionable. His argument seems to be that if we are going to treat some physical law as fundamental, in the sense that all other results in that part of the subject are to be derived from it, then that law should be susceptible of accurate,

/direct.....

direct experimental verification by pupils or students. Ideally, one might wish this to be so; but a glance at other areas of physics shows that this principle is not heeded when they are taught, and at the very least, the consequences have not yet proved to be alarming. Gravitation is customarily taught on the basis of Newton's Law of Universal Gravitation, for example, but no teacher or pupil has ever "verified" this Law in the laboratory. Is it not both realistic and sensible to approach Coulomb's Law in electrostatics in the spirit of the following statement:- "Crude experiments suggest that the law of force between two charged particles is of the form $F \propto \frac{q_1 q_2}{r^2}$. It seems a reasonable guess that this is the law of force between two charges at rest; after all, we already have a law of gravitational attraction which has the same mathematical form. At any rate, we can do worse than assume that electrical forces are described by this law of force and see how far this assumption takes us. If our assumption is correct, the predictions made from our theory will be confirmed by experiment."

Surely we teach a great deal of physics in just this spirit; why, then, should we frown on the teaching of electrostatics along these lines? Of course, if we can find some other law or principle in electrostatics which can be established with greater accuracy than Coulomb's Law and which can serve as the basis of an equally simple approach to the subject, we should undoubtedly consider very seriously its claims as an alternative to Coulomb's Law. Unfortunately there does not seem to be any result which meets <u>both</u> these conditions. Rekveld himself offers no satisfactory alternative; the best he can do is to describe an experiment which can lead to the equation relating the induced charge density on a pair of parallel plates in a uniform field to the intensity of the

/field.

field. He does not say how the electric intensity is to be measured; he suggests that the induced charge on the plates should be measured with an electroscope or a ballistic galvanometer. This does not sound like an experiment which can be performed with superior accuracy to the verification of Coulomb's Law; and since it is intended to lead to Gauss's Law, it is in any case not worth considering for South African syllabuses.

Another objection of Rekveld's is that the use of Coulomb's Law is a barrier to the understanding and acceptance of the electric field concept, since it places emphasis on the action-at-a-distance viewpoint. Rekveld would have one use the field concept right from the start. In our view, beginning with Coulomb's Law would present no real difficulty to a beginner who has inevitably grown used to thinking of electrical forces as exerted by one charged object on another;

with charges, conductors and magnets in vacuo. This means that the pupil will not be in a position to design capacitors, electromagnets or transformers, or give a quantitative account of what happens, electrically and magnetically speaking, inside dielectrics or magnetic materials. On the other hand, he will be able to understand the fundamentals of the behaviour of charged particles in molecules, atoms, cathode-ray tubes and particle accelerators, so that he is allowed entry into some of the most interesting areas of modern physics. Also, as has already been pointed out, exclusion of dielectric and magnetic materials from consideration does permit a much simpler presentation of the subject than would otherwise be possible.

7.4 Coulomb's Law

7.4.1 Should quantitative electrostatics start from Coulomb's Law?

Traditionally, quantitative electrostatics begins with Coulomb's Law. From time to time, objections have been raised against this approach; since Rekveld (1958b) has

study electromagnetic induction, the much greater ease of description made possible by the use of "field" language should become apparent to them.

Rekveld's other objection to the introductory use of Coulomb's Law concerns the difficulties it presents for the introduction of "rationalization". This objection has some force, but since "rationalization" is not introduced in the present syllabus, it does not apply here.

In short, we do not think the objections brought by Rekveld and others against the introductory use of Coulomb's Law can be sustained - in any event, not when the Law is intended to form a cornerstone of a syllabus of the present kind. We think that it is advantageous to begin with Coulomb's Law because it can be given <u>some</u> experimental justification, and it offers a simple and direct means of bringing the pupil into contact with the quantitative aspects of electrostatics. This does not rule out the possibility that there are other and perhaps better introductory approaches to electrostatics at this level. As Gledhill (1969) has suggested, there is room for further research here.

7.4.2 Introducing Coulomb's Law: (i) the inverse-square law

One can pave the way for the introduction of Coulomb's Law by pointing out the need for <u>measurement</u> if electrostatics is to be established as a quantitative branch of physics. The forces between static charges are measurable and provide a natural starting point for such an establishment. The dependence of these forces on the separation between the charges is the first point to be investigated. The P.S.S.C. Laboratory Guide describes a simple way of doing this, using pith balls coated with conducting material as the charged "particles" between which the force of repulsion is to be measured. Ideally such experiments should be done on dry

/days.....

days, but this may not be easy to arrange. The P.S.S.C. recommends that the apparatus should be partially enclosed in a box; in this way the apparatus can be protected against air currents, and the box can be heated to reduce the relative humidity before the experiment is done. (It has been reported that this experiment "worked admirably on a humid day" (cf Lord and Ritchie 1962).) With care, this experiment "establishes" the inverse square law in electrostatics to an accuracy of a few per cent. At very short distances between the balls (e.g. less than a centimetre or two) one can expect bigger departures from Coulomb's Law behaviour, on account of the mutual perturbation of the charge distributions on the balls brought about by their proximity to each other. This is a most important point: pupils must certainly be made to realize that the inverse square law (and Coulomb's Law as a whole) can be expected to apply accurately only to the electric force between point charges, or between charged bodies whose separation is very large compared to their linear dimensions.

7.4.3 Introducing Coulomb's Law: (ii) dependence of the force on the magnitude of the charges

The inverse-square law is not the whole story of the law of force between charged particles; a knowledge of the dependence of the force on the magnitude of the charges of the particles is also needed. One has first to find a method of obtaining a set of charges of known relative magnitudes. (Many textbooks simply ignore this far from unimportant point.) There are two main ways in which this can be done. One is to define what is meant by equal charges as follows: "The charges on two bodies shall be said to be equal in magnitude if they exert equal forces on a third charge equidistant from them, the distance being such that all three can be treated

/as.....

as point charges" (Coulomb's Law Committee 1950a), and then to use the Principle of Superposition of electrical forces (treated as an experimentally established result) to show that "if n equal point charges of the same sign are brought together at the point P, the force that the combination exerts on any other point charge X will be n times the force which any one would exert if acting by itself." (Coulomb's Law Committee 1950a). These facts can then be combined to yield the conclusion that "the ratio of the numerical values of any two point charges is to be identified with the ratio of the forces which they individually exert on a third point charge at the same distance r." The other method is that apparently followed by Coulomb himself (cf Arons 1965 on this point); it consists in the following set of operations:

(a) Select a number of conducting spheres, as nearly identical as possible, as the objects to be charged.

(b) Charge any two of these balls, A and B.

(c) Place B in contact with an uncharged ball C; by symmetry, the charge on C must equal the charge now on B and half B's original charge (the equality of the charges on B and C can be tested by placing each in turn at the same arbitrarily chosen distance from A and comparing the forces between A and B and between A and C).

(d) Obtain other simple fractions of the original charge of B by appropriate sharing of charge between balls.

It will be noticed that in steps (c) and (d) appeal is implicitly made to the Principle of Conservation of Charge; it is assumed that when the charge is shared between two conducting spheres, the total charge on the spheres is the same at the end as it was in the beginning. (We are not concerned here with charge lost to or gained from the surroundings). Thus in either method of establishing a scale of charge, use

/has.....

has to be made of a fundamental principle in electrostatics other than Coulomb's Law.

In the Common Basic Syllabus in Physics, one of the points listed is: "If different charges are placed at a fixed distance from a fixed charge, the former charges are proportional to the forces which they experience." This looks like a recommendation to use the first of the approaches described above; if this recommendation is to be adopted, we hope that teachers will draw attention to the part played by the Principle of Superposition. Perhaps, on the whole, the other approach is simpler to use in an introductory treatment (it is used in the P.S.S.C. treatment, for example), since it uses somewhat simpler conceptual operations. On the other hand, the first approach does have the advantage that it does not necessitate the description or performance of an experiment to show that the electric force on a charged particle is proportional to the charge on the particle.

If the second approach is used, then one can argue as follows:- "Now that we have a method by which, in principle, a scale of charge can be set up, we can, also in principle, proceed to measure the force between some standard charged particle and other charged particles placed in turn at the same distance from the standard, the charges on the latter particles bearing a known relationship to one another. Such an experiment should show that the force is proportional to the amount of charge on the 'second' particle, from which it is easy to conclude that the force is proportional to the product of the amounts of charge on each particle." Let us admit that it would be difficult to carry out such an experiment convincingly, and that it is best to let it remain a "thought experiment". (The same reservation applies to the experimental establishment of the Principle of Superposition.) In other words, our assertion that the force is proportional

/to.....

to the product of the charges is to be treated essentially as a postulate, to be justified by the success of the predictions based on it.

7.4.4 Introducing Coulomb's Law:(iii) the proportionality constant and the coulomb

By means of the foregoing arguments and experiments it has now been established that the electric force between two charged particles is proportional to the product of the charges and inversely proportional to the square of the distance between them or, in symbols,

$$F = k \frac{q_1 \, q_2}{r^2}$$

where k is a proportionality constant whose value will be determined by the choice of the unit of charge. It will be noted that in our development of this equation we have not laid down a definition of the concept of charge. This is because we have taken the view that charge as the essentially new fundamental concept in electricity, is basically indefinable; or, if it is to be defined, such a definition must be in terms of some other fundamental concept treated as indefinable. True, we have described procedures by which we can set up a scale of numerical values of charge, and we can, if we insist, state a definition of the ratio of numerical values of any two point charges in terms of one or other of these procedures; but this, it will now be recognized, takes us no nearer to the "logical identity" we would require as a definition of charge itself. (Cf. Bunge's (1966) criticism of Mach's definition of mass, considered in Section 4.7.4.) In the junior secondary school the pupil will have met a "denotative definition" (cf Section 3.2.6) of charge in some such form as "the property by virtue of which a body is able to attract light objects of any description" (cf our discussion

/in.....

in Section 3.14.3 of the introductory treatment of charge). This will be an adequate background to the present quantitative treatment.

Since pupils will already have made the acquaintance of the coulomb in the junior secondary school, it may seem that one should proceed to use it as one's unit of charge in electrostatics without any hesitation. But there is a difficulty about this; at this stage the pupil does not know, even in principle, how to set about measuring the charge on a particle or an isolated conductor and obtain an answer in coulombs, if the coulomb is supposed to be defined in terms of the ampere and the second. Thus it would not be very helpful to tell a pupil that we already have a unit of charge, the coulomb, so that the proportionality constant k in Coulomb's Law must be found by experiment; the pupil might well retort "How can you determine 'k' from measurements on the force between charges if you don't know the value of the charges in 'coulombs'?" What is needed to solve this problem is a temporary definition of the coulomb, which we can obtain simply by using Coulomb's Law and making an appropriate choice for the value of the constant "k".

In view of the air of controversy which still hangs over the matter of electrical units, we should like to dilate a little on the question of the choice of "k". It is obvious that if the unit of charge is as yet undefined, one way of defining it is to assign some arbitrary value to the constant "k". The simplest choice is to put k = 1. This is what is done in defining the electrostatic unit of charge (sometimes called the "statcoulomb"); with the CGS units of force (the dyne) and length (the centimetre) already available, k is put equal to unity and one can define the e.s.u. of charge as "that point charge which, when placed one centimetre away from an equal and like point charge, repels it with a force

/of.....

of one dyne." There is nothing wrong with this; it just turns out that it is an inconvenient choice of unit for many purposes. On the other hand, there is nothing divinely ordained about putting k = 1; Lipson's (1957a) view that "a unit is that which produces unit effect in unit circumstances" (he claims that this is "the physicist's point of view") looks like an attempt to elevate a useful convention into a papal decree. One snag attached to putting k = 1 is that it tends to distract attention from the fact that k is not a pure number, but is a physical quantity having dimensions, and that by the same token it tends to create the impression that the unit of charge defined in this way is a derived unit, i.e. one which is expressible in terms of the units of mass, length and time. Now it is contrary to our intuition that the new electrical unit should be a derived unit of this type. As Feather (1966) puts it, "Through familiarity with the phenomena we come to the conclusion, intuitively, that [electric and magnetic phenomena] cannot be described convincingly in terms of the 'massiness' of bodies, or of the elementary constituents of bodies, acting in space and time - as gravitational phenomena can be described". Therefore if we put k = 1 or any other value, we should stress that the quantity concerned has units; in the case of the CGS-electrostatic system of units, they would be dynes-(centimetres)^2 per (e.s.u. of charge)^2.

We now come to the choice of "k" for our temporary definition of the coulomb. We suggest that this be approached in the following way:-

"Putting 'k' = 1/obviously the simplest choice we can make. However, using SI units for force and length, this would give us a unit of charge which, although of a convenient size for electrostatic work, is too small for practical convenience in current electricity. Putting 'k' equal to

/9.....

9×10^9 gives a unit of charge that is of suitable magnitude and agrees very nearly (i.e. to within 0.1%) with the internationally accepted unit of charge. We shall call the unit of charge defined by this choice of 'k' the coulomb. It follows that the units of the proportionality constant must be newton-metre^2/coulomb^2 (N-m^2/C^2)". (Some teachers may have been taught electrostatics in such a way that they may now be surprised at our failure to state that "k" is only equal to 9×10^9 N-m^2/C^2 when the charges in question are in vacuo. What about the influence of the medium? Should we not allow for this by letting the value of "k" in Coulomb's Law depend upon the medium in which the charges find themselves? This matter has been exhaustively dealt with by the Coulomb's Law Committee of the American Association of Physics Teachers (1950a); here we shall only indicate the main points of the argument. The force exerted by one charged particle on another is always to be computed, in SI-units, by setting $k = 9 \times 10^9$ units, very nearly. If both particles are immersed in some dielectric medium, the resultant force experienced by either one of the charges will certainly be less than what it is in a vacuum; but this is because the charges on the particles induce polarization charges on the atoms of the medium, and each particle "feels" the resultant of the forces exerted by the other charged particle and by the polarization charges. We feel it is unwise to introduce any complications into the picture by references to "in vacuo"; the electrostatics in the syllabus is primarily concerned with charges in air or in vacuo, and any pupil who is worried by the "influence of the medium" can always be given the above explanation.)

We do not think it worthwhile to write out or state a formal version of this temporary definition of the coulomb.

/It.....

It is only temporary, after all; and if stated in the form that "1 coulomb is that point charge which, when placed 1 metre away from a similar point charge, repels it with a force of 9×10^9 N", it is scarcely realistic! What pupils should appreciate about the choice of "k" is that it does, in principle, enable one to find the value of the charges on a pair of equally charged particles by measuring the electric force between them at a known separation. Also, if one has a pair of charged particles, the charge on one of which is unknown, the charge on the other can be found by measurement of the force between them at a known separation; and finally, that the force between a pair of charged particles carrying known charges can be calculated if their separation is known. The use of plus and minus signs when substituting for the charges in the Coulomb's Law equation can be explained as leading to a simple convention regarding the direction of the electrical force on either particle; when the charges are opposite in sign, the value obtained for the force is negative, and this means that each particle is attracted towards the other; when the charges are like, the positive value of the force means that it is repulsive. (There is no need at this level to turn the Coulomb's Law equation into a proper vector equation by the use of an appropriate unit vector.)

Pupils should certainly come to realize that a coulomb is a very large amount of charge to have on an isolated charged conductor. They should be familiar with the microcoulomb (μC) as a submultiple of the coulomb; it would be a good thing for them to know that a microcoulomb is the order of magnitude of the charge on the dome of a "school model" Van de Graaff generator when it is some distance away from its auxiliary dome and the air around it is on the point of "breaking down". They should be duly impressed with the magnitude of the effects this amount of static charge can produce.

/It.....

It would also be of interest to them to know that the amount of charge associated with a lightning discharge is of the order of tens of coulombs.

Pupils would also be interested to know of the accuracy with which Coulomb's Law is known to hold: does the force between a pair of charged particles fall off exactly as the square of the distance between them, or is it possible that the power of the distance on which force depends may differ slightly from -2? At the time of writing, the latest information on this subject is that obtained by Plimpton and Lawton (1936), who have established the accuracy of the inverse-square law to within a few parts in 10^9. Also of interest is the range of distances over which it might be expected to hold (cf Purcell 1965). This is enormous - from 10^{-15} m to many kilometres. It is suggested that these facts can be mentioned to pupils simply for the purpose of arousing interest: there can be no question of giving a description of the experimental methods by which this information has been discovered. (The brighter pupil can be referred to the relevant passage in Purcell's (1965) book, which gives a brief non-mathematical discussion.)

7.4.5 Coulomb's Law and conservation of charge: some relevant experiments

Prior to further quantitative development of the subject, pupils should carry out some important experiments with an electroscope additional to those which they carried out in the junior secondary school. These will help to give some further insight into the working of the electroscope and shed new light on certain aspects of electrical behaviour. First, Faraday's Ice-Pail Experiment, with which one can show that if a large, hollow conductor with a suitable opening is attached to the central system of the electroscope, any charged

/conducting.....

conducting object touched to the inside of the hollow conductor will give up virtually all of its charge to the latter, with practically none of this charge remaining on the inside of the hollow conductor. This result is of practical importance as it shows that we can arrange for all the charge on a charged conductor to be transferred to an electroscope - something that would not happen if the conductor were simply touched to the knob of the electroscope. Lewis (1964) has pointed out that the absence of charge on the inside of a hollow charged conductor can be shown better by using a "proof plane" (conducting disc mounted in an insulating handle) to sample the charge on different parts of a charged conductor and an electroscope to "detect" the charge on the proof plane; he suggests that Faraday's Ice-Pail is chiefly valuable in showing that a charged object inside a hollow conductor induces charges on the inside and outside of the conductor which are equal in magnitude to the inducing charge. We recommend that as time is likely to be short, the Ice-Pail experiment should be done as the means of showing both of these results. In view of the familiarity of the experiment, we shall not trouble to describe it here; adequate descriptions and discussions of it are given by Scott (1959), Sears and Zemansky (1964), Lewis (1964) and Bennet (1965, 1968), among others.

The electroscope-plus-ice-pail arrangement can be used to illustrate the Principle of Conservation of Charge (cf Lewis 1964, Jardine 1967). Simple experiments can be done to show that when we rub two "neutral" things together and give them opposite charges, no net charge is created. With a perspex rod and a polythene rod as the originally neutral objects (any initial charge on them can be removed by passing them rapidly through a bunsen flame), the sequence of steps

/would.....

would be as follows:-

(i) Show with the aid of the electroscope that the rods are uncharged.

(ii) Rub the rods together and place the rubbed ends in turn inside the "ice-pail".

(iii) Hold the charged ends of the rods simultaneously in the "ice-pail", making sure that they do not touch.

Step (ii) should show that the charges produced are equal and opposite, a result confirmed by (iii) which should show that the net charge on the two rods is still zero.

If a duster is rolled up around a polythene rod and the two are placed inside the "ice-pail", the rod can then be slowly withdrawn. The duster will acquire a charge, as shown by the electroscope, and after the duster has been removed, the rod can be placed inside the can to show that it has acquired an equal and opposite charge.

It is also useful to show that conservation of charge holds when charge is shared between conductors. One can do this with a pair of metal spheres of unequal diameter, each mounted on insulating rods. Initially one conductor can be charged (e.g. by placing it in contact with the dome of a model Van de Graaff generator) and lowered into the ice-pail without touching it; the deflection of the electroscope leaf should be noted. The other conductor should now be lowered into the "ice-pail" and placed in contact with the first. This should produce no change in the deflection of the leaf. Then the first conductor can be removed, to show that the second conductor is charged. The deflections of the electroscope leaf corresponding to the charge on each conductor after sharing should be noted; they should indicate that the larger conductor has obtained the bigger share of the charge. The deflection of the electroscope is not directly proportional to the charge on it, so that one should not expect

/that.....

that the sum of the latter two deflections should be equal to the original deflection; it would be as well to warn pupils of this. It might be instructive for them actually to measure these deflections so as to investigate this point for themselves.

Some pupils may be interested in the mechanism whereby the charge distributes itself between the two conductors so that the larger conductor takes the larger share. It will not be possible for the teacher to give a detailed answer at this stage, but he can say that the mutual repulsion between like charges will lead to the excess charge on a pair of conductors distributing itself so that most of it will be found where there is more "room" for it to be. He can add the rider that the equilibrium condition is one in which the mutual potential energy of the [excess] charge will be a minimum.

An interesting point arises in connection with the sharing of charge between a pair of conductors. Suppose that these are two spheres of unequal radius, separated by a long, fine wire. It is a well-known fact that the concentration of charge on the smaller sphere will be bigger than that on the larger sphere, giving rise to a stronger electric field in the vicinity of the smaller sphere. This is actually not in conflict with the general principle that most of the charge goes "where there is more room for it". There will be more charge on the bigger sphere, since the total charge will divide itself between the spheres in the ratio of the capacitances, which are proportional to the radii of the spheres. But the concentrations of charge on the spheres are inversely proportional to their surface areas, which are in turn proportional to the squares of their radii, so that the concentration of charge will be greater on the smaller sphere (cf Feynman et al 1964 and Halliday and Resnick 1966).

/7.4.6.....

7.4.6 Conservation of charge and the "earthing" of a charged conductor

It is customary for the process of "earthing" a charged conductor to be explained as a process of sharing of charge between a very small conductor and a very large one (the earth), so that the earth takes virtually all the charge (cf Rogers 1960, PSSC 1960, 1965 and Jardine 1967, to name only a few sources in which this explanation is given). In principle, there is nothing wrong with this; a simple calculation shows that the capacitance of the earth is some 700 μF, which is enormously greater than that of any ordinary isolated conductor that we might expect to find on earth, so that if there were direct contact between charged conductor and earth, we might reasonably expect that virtually all the charge on the former would transfer itself to earth in the process of charge "sharing". However, Morgan (1953) has pointed out that for experiments ordinarily carried out in the laboratory, there is no need to bring in the (real) earth in order to explain "earthing". The walls of the laboratory are sufficiently conducting for the laboratory to be regarded as a hollow conductor, and therefore if we place a charged conductor in contact with the inside of the laboratory, all of its charge will be transferred to the outside of the laboratory.

Morgan has also shown that the usual description of the charging of a conductor permanently by induction is slightly erroneous. Suppose that one gives a permanent charge to an isolated conductor by the following sequence of operations: (i) charge a perspex rod positively by rubbing it with a handkerchief; (ii) hold it near the conductor; (iii) touch the conductor with a finger (or other earthing connection) while still keeping the charged rod nearby; (iv) remove the finger; (v) remove the charged rod. It is

/usual.....

usual to say that when the rod is held near the conductor and the earth, causing the induction of some negative charge on one side of the conductor, and on the portion of the earth nearest the rod, an appropriate amount of positive charge will be induced on the opposite side of the earth. But this ignores the presence of the handkerchief which was used to charge up the rod. This has a negative charge on it, which will induce an appropriate amount of positive charge on the portion of the earth nearest to it. A detailed application of the Principle of Conservation of Charge to the problem, taking into account the charge on the handkerchief, shows that there will not be any charge induced on the opposite charge of the earth. It would be most instructive for brighter pupils to analyse these different stages of charging by induction with the aid of Conservation of Charge.

Conservation of charge may seem almost self-evident to pupils, at any rate as applied to everyday matter which they are accustomed to thinking of in terms of a simple model of positively-charged, negatively-charged and neutral particles. It would interest them to know that this Principle has profound implications in the exciting world of elementary-particle physics. In high-energy physics laboratories, in the upper atmosphere and all over the universe, extremely high energy nuclear reactions are constantly taking place in which new particles are continually being formed. These particles lead a transitory existence, some of them decaying into other particles which may in turn disintegrate yet further, while others suffer annihilation on encountering their "antiparticles". In these processes of creation and annihilation of matter, it turns out that the Principle of Conservation of Charge is never violated - the net amount of electrical charge in the universe remains constant.

/7.4.7.....

7.4.7 The Principle of Superposition

The Principle of Superposition has been mentioned earlier in these pages (Section 7.4.3); as we have already suggested, it is important for pupils to know of its existence and to appreciate that, like the Principle of Conservation of Charge, it is not self-evident. It need not have turned out that "the force with which two [point] charges interact is not changed by the presence of a third charge" (Purcell's (1965) statement of the Principle), but experimentally we find that it is so. The P.S.S.C. even include an experiment in their laboratory course intended to verify that electric forces add vectorially, but they admit that difficulties with leakage of charge limit the accuracy to no better than 10%. We do not think it essential that such an experiment should be performed by pupils or even demonstrated to them; like certain other principles they have encountered, the Principle of Superposition can be treated as being justified by the success of the predictions based upon it.

7.5 The electric field

7.5.1 Introduction

The concept of the electric field is the most important single topic in the electrostatics section of the course. In the junior years, the pupil has encountered the magnetic field, treated in purely qualitative fashion; he has learned to think of the magnetic field as a "region of influence" surrounding magnets or current-carrying circuits, and he knows that with any point in the field he can associate a characteristic direction. The task now is to build up the pupil's picture of the electric field in a similar kind of way, but to extend it to take into account the association of a characteristic magnitude as well as a direction with any point in

/an.....

an electric field. This idea is so important that all possible methods of attack should be used to develop it viz. analogies with other physical fields, qualitative and quantitative demonstrations and experiments, and simple illustrative calculations and problems.

7.5.2 Teaching the electric field concept; analogy and preliminary ideas

The analogy with the magnetic field has already been mentioned. It is useful up to a point since, like the electric field, the magnetic field is a vector field, but it cannot aid the quantitative development at all. The gravitational field is more helpful in this respect; as a result of the similarity in form between the fundamental laws of force of the gravitational and electric fields, there are many correspondences between the two kinds of field. There is much to be said for developing the concepts of the gravitational and electric fields "in parallel", as it were. Other physical fields should also be mentioned: the new Teachers' Guide for Physical Science, Part I (Prinsloo et al 1968) specifically mentions the nuclear force field (i.e. the "strong interaction") in order "to suppress the idea that all fields of force obey the inverse square law." The nuclear force cannot be summed up in a neat formula, but the teacher could sketch a graph of the variation of force with inter-nucleon distance (cf Atkins 1965).

Teachers who search the textbooks will find various statements of what is meant by the term "field", be it gravitational or electric, ranging from the simple and colloquial to the formal and sophisticated. On the one hand, we have Rogers's account of the gravitational field (Rogers 1960): "....we may express the experimental connection between weight and... mass by imagining that there is a 'field of force'

/extending.....

extending out from the Earth like a set of clutching tentacles waiting to pull down on any mass placed there. The field is not itself a force but a state of waiting-to-pull-on-mass". On the other, we have Holton's (1952) description of the electric field: "The totality of electric intensity vectors in a given region we call the 'electric field' in that region." Somewhere in between the two is the statement in the P.S.S.C. Teachers' Resource Book and Guide, Part Four: ".... This pattern of forces is called the electric force field or even just the electric field." The pupil's understanding of the concept of a force field ought eventually to take in all these ideas. That embodied in Rogers's statement forms an excellent starting-point. Approaches which tackle the problem in this way are to be found in a number of textbooks in common use at university level (cf Sears and Zemansky 1960, 1964, Halliday and Resnick 1966). As applied to the electric field, the argument runs as follows:-

Let two charged particles A and B be placed in a region of otherwise empty space. Each charge experiences a force owing to the presence of the other; in particular, B experiences a force owing to the presence of A elsewhere in the region. One way of describing this state of affairs is to say that the presence of A has modified the surrounding space in such a way that a force is now exerted on another charge in the space; A has set up an electric field in the surrounding space, whose distinguishing characteristic is that forces of electrostatic origin act upon charges placed within it.

7.5.3 Can the electric field vector be defined?

It is standard practice when introducing the concept of the electric field to "define" the electric field vector at a point as the force per unit charge which would act on a "small"

/positive.....

positive test charge placed at the point, or as the limiting value of the force per unit charge as the value of the test charge is allowed to tend to zero. That is, it is tacitly assumed that this concept is capable of genuine definition; however, we have seen in other branches of physics that it is unwise to assume without question that any given fundamental concept can be given a logically satisfactory definition. In view of the fundamental importance of the electric field concept, we should at least pay some attention to the question of its definition. We begin by subjecting the above-quoted "definition" to critical examination. It is at once evident that it is an "operational definition" and, as such, open to certain objections. An obvious objection to the practice of defining the field vector in terms of the force on a test charge is that we conceive of the field independently of the presence of test charges. The whole point of the introduction of the concept is that we can hypothesize it independently in this way. The theory of electromagnetic waves, according to which changes in electric and magnetic fields are propagated through space with a finite speed, rests on the hypothesis of electric and magnetic fields existing independently of any test charges which may be used to detect their presence.

Purcell (1965) objects to the "force-per-unit-charge-on-a-test-charge" definition of the electric field vector, but not on the above grounds. He points out that the process of letting the magnitude of the test charge tend continuously to zero cannot correspond with reality (charge comes in "packages" equal to integral multiples of the electronic charge), so that one cannot in this way overcome the objection that the presence of a test charge will upset the previously-existing electric field. He therefore adopts as his definition of the electric field vector the equation

$$\vec{E} = \sum_{j=1}^{N} \frac{q_j \hat{r}_{oj}}{r_{oj}^2}$$

where $\vec{E}$ represents the field set up at the point in question by N charges $q_1, q_2, \ldots q_N$, r_{oj} is the distance from the j^{th} charge to the point, and $\hat{r}_{oj}$ is a unit vector drawn from the j^{th} charge to the point. At first, this does seem a convenient escape route. Not only does it remove the difficulty associated with letting the test charge tend to zero; it also furnishes a constitutive definition of the electric field vector in terms of the charges setting up the field and the distances from the charges to the point in question. However, it is at once obvious that its usefulness as a constitutive definition is confined to static electric fields. The moment we have a field set up by accelerating charges, it is no longer possible to obtain the electric field at a point by using Purcell's formula. In principle, one can still manufacture a revised constitutive definition of the electric field vector in terms of the field-creating charges and their motions, but this can be recognized at once as a highly artificial procedure, requiring the use of the field equations of electromagnetic theory and the Lorentz force equation as essential steps in the production of the required "definition". It is significant that Purcell shifts his ground by the time he comes to the sixth chapter of his book, concerned with the magnetic field. Here he refers to "... having defined the electric field vector $\vec{E}$ as the force on unit test charge at rest"!

The best way out of these troubles is to acknowledge that one cannot find a simply-stated, genuine <u>definition</u> of $\vec{E}$ that will meet all logical objections and that, in the last resort, $\vec{E}$ is "defined" by the set of fundamental axioms of electromagnetic theory.

7.5.4 <u>The quantitative treatment of the electric field</u>

The equation $\vec{F} = q\vec{E}$ states the fundamental relation between $\vec{E}$ and something we can measure, $\vec{F}$ being the electrostatic

/force.....

force on a positive charge of magnitude q. This point must naturally be established early in any quantitative treatment of the electric field. This could be done by appropriately continuing the story from the point at which the qualitative idea of the electric field was introduced (cf. end of Section 7.5.2).

That introduction indicated a method of testing for the presence of an electric field in a given region of space, viz. place a charged particle in the region and note whether it experiences an electric force or not. But a force on a charged particle is something that is measurable, at least in principle, so that we now have a means of associating a magnitude aswell as a direction with a point in the electric field - we can measure the magnitude and note the direction of the electric force on a "test charge" placed at the point. It is sensible not to have the characteristic magnitude of the electric field dependent on the amount of charge on the particle used as a test charge. Therefore we can take as our measure of the "electric field vector", "electric field strength" or "electric intensity" at a point the electric force per unit charge on a small particle carrying a small positive charge placed at the point. (We have reproduced this "definition" as it was worded by the present writer in the Teacher's Guide for Physics (Helm 1963). In the new Teachers' Guide for Physical Science, Part One (Prinsloo et al 1968), another author has chosen to re-word the "definition" as follows: "[The electric intensity] is the electric force per unit charge on a small particle carrying a unit positive charge and placed at the point." The practical absurdity inherent in the use of a test charge of 1 coulomb was just what the author sought to avoid in his original statement. On this point, cf Worrell 1960). To call this a "definition" is not strictly correct, as we have indicated, and there is no need to claim that it is a true definition.

/If.....

If the pupil is happy to accept an undefined concept of mass, having "force divided by acceleration" as its measure, he should be similarly happy to accept an "undefined" concept of electric field vector, having as its measure the force per unit charge on a test charge. (If magnetic fields were present, it would be necessary to specify that the test charge must be stationary; but this is a complication with which teacher and pupils need not be saddled at this stage.)

For the sake of standardization, one of the above-mentioned names for the electric field vector should be agreed upon as the "official" name. We suggest that it is safest to use the term "electric intensity", if only because this avoids confusion between the field vector and the broader concept of the electric field. The parallel between the vector characterizing the electric field and the corresponding vector for the gravitational field should be drawn by the teacher.

7.5.5 Experimental support

There are a number of simple ways in which the ideas set out above can be illustrated experimentally (cf P.S.S.C. Teachers' Resource Book and Guide, Part Four, Lewis 1964, Jardine 1967). Just as is the case with magnetic fields, it is possible to show the directional aspect of electric fields by the use both of "compass needles" and patterns of lines of force. As a "compass needle", one can use a 2" long piece of stiff paper or aluminium, balanced, and able to rotate on a fine pin mounted on a long insulating handle (cf Lewis 1964). With this device one can explore the electric field in the vicinity of various charged objects, such as the dome of a Van de Graaff generator, insulated charged conductors of various shapes (including a hollow cylinder), and a pair of large, oppositely-charged parallel plates.

/We.....

We have referred in Section 3.14.3, part (vii) to the demonstration of electric field patterns with the aid of a paper "hula skirt" placed on the dome of a Van de Graaff generator and with grass or lettuce seeds sprinkled on an insulating liquid. If these demonstrations were not done in the junior years, they should certainly be done now; and even if they were done earlier, they would be well worth repeating at this stage.

To demonstrate the quantitative aspect of the electric field, the force on a charged ball suspended in the uniform field between a large pair of oppositely charged parallel plates could be measured. In one version of this experiment, the plates are mounted vertically and the ball is suspended from a pair of long nylon threads. The plates are charged from an E.H.T. supply. As in the inverse-square law experiment mentioned in Section 7.4.2, the electric force on the ball will be proportional to the tangent of the angle of deflection of the threads. Alternatively, the plates can be mounted horizontally (Lewis recommends a separation of 10 cm), with a ball suspended from a helical glass spring passing through a hole in the upper plate. In either version of the experiment, the electric force can be shown to depend on the amount of charge on the ball.

7.5.6 Calculations

Use of the "definition" of electric intensity can be illustrated by the calculation of the electric intensity set up at a point by a single point charge and by two equal or unequal or unlike point charges. Pupils should not be burdened with awkward calculations requiring much use of vector addition. Teachers should simply discuss with pupils the principle of the method of calculating the electric intensity in such instances and confine numerical problems to electric intensities in the fields set up by not more than two point

/charges.

charges. Examples of some practical interest are the calculation of the electric intensity at points on the axis and on the perpendicular bisector of an electric dipole (i.e. equal and opposite charges placed a short distance apart).

7.5.7 Uniform electric fields

From the point of view of an elementary course in physics, the most important type of electric field is the uniform field, which is the kind of field one obtains between and away from the edges of a pair of equal and oppositely charged parallel plates whose separation is small compared with their linear dimensions. Pupils are not in a position to understand the standard proof of the uniformity of the field between a pair of parallel plates, which makes use of Gauss's Law. Brighter pupils would be helped by the argument given, among others, by the P.S.S.C. (1960,1965) and Rogers (1960) which makes use of a hypothetical observer looking at a charged sheet at different distances away from it. If his "cone of vision" remains the same (i.e. subtends the same solid angle), the charge it "intersects" on the sheet increases as the square of the distance from the sheet, while the force from any element of charge falls off inversely as the square of the distance. The net effect of this is for the force felt by a test charge held by the observer to remain the same while the distance changes. For the rank and file, the uniformity of the field will have to be shown experimentally. This can be done with the experimental arrangements described above, in which the force on a charged ball in the field between parallel plates is measured; the force on the ball can be shown to be independent of its position between the plates. (For its uniformity of direction, the demonstration of the lines of force provides an independent proof.)

/7.6. .....

7.6 Potential difference

7.6.1 Potential energy in gravitational and electric fields

Potential energy in an electric field is closely analogous to potential energy in a gravitational field and pupils should have little difficulty in accepting and understanding electrical potential energy, potential and potential difference if they have understood the kind of discussion given in section 4.15.3 of gravitational potential energy.

Work done in transporting a small positive test charge in an electric field can be compared to the work done in transporting a particle in a gravitational field. If the particle is lifted against the gravitational field, then work is done in pushing particle and earth apart, and we can regard the energy transferred in this operation, which we call gravitational potential energy, as being given either to the particle-earth system or to the gravitational field. If we adopt the former view, then we note that the process of "pushing particle and earth apart" has negligible effect on the earth, so that we can regard the potential energy as being to all intents and purposes associated with the particle alone. This view is consistent with what happens when the particle falls freely in vacuo; we obtain a consistent description of this situation if we assume that we are simply concerned with the interchange between the gravitational potential energy of the particle and its kinetic energy.

An analogous situation arises if we push a "small positive test charge" against an electric field without giving it kinetic energy. The energy transferred is presumably "electrical potential energy"; we would be justified in speaking of it in this way if the work done in transporting the charge was independent of the path followed. This "electrical potential energy" can be regarded as having been given

/either.....

either to the system "test-charge-plus-charges-setting-up-the-field" or to the electric field itself. It is often the case in problems of interest in electrostatics that the transport of a "test charge" (or any charged particle of interest) in the field has negligible effect on the charges setting up the original field, either because the latter charges exert far bigger forces on one another than any forces that might be exerted on them by the test charge, or because the field-creating charges are "nailed down" in some way. In such cases we can regard the "electrical potential energy" as belonging essentially to the test charge itself. (For further comment on the "location" of electrical potential energy, see Sections 7.6.7 and 7.10)

The analogy between gravitational and electric fields in respect of potential energy should be discussed with pupils in some detail. It is important for the teacher to emphasize the path-independence of the work done in transporting a charge in an electric field. This result can easily be established for a uniform electric field; it is suggested that this should be done for the pupil, or that he should do it himself, using the derivation he will have met when studying the gravitational field. In the course of proving this result, it will emerge that the work done in transporting a positive charge q a distance d against an electric field of intensity E is equal to qEd - an important result which can be used again.

7.6.2 Introduction of potential and potential difference

It is easy now to pass to the concepts of potential and potential difference. One may put the matter in this way: "It has already been seen that to every point in an electric field we may assign a vector quantity called the electric

/intensity,....

intensity, whose magnitude and direction will depend on the position of the point. Physicists have found it extremely useful also to assign to every point in the field a quantity connected with potential energy in the field. Now potential energy is a quantity associated with a charge as well as a point; somehow we must remove the dependence on the magnitude of the charge if we are to define this new energy-associated quantity conveniently. This we do by once more employing the idea of a 'small positive test charge': we define this new field quantity, called the potential at a point in the field, to be the potential energy per unit charge of a small positive test charge placed at the point". This yields the formula $V = \frac{W}{q}$ for electrical potential, where q is the charge on the test particle and W is the work done in transporting it to the point in question from a point arbitrarily chosen as the place where the potential energy of the charge is zero. The discussion and the definition of potential also make it clear that the units of potential must be joules per coulomb. All this should remind the pupil of his meeting the concept of "potential difference" in the junior secondary school, defined in terms of energy given up per unit charge. There the concept was developed exclusively with reference to electrical circuits; here it is defined in terms of change of position of charge in an electric field. At this stage, the identity of the viewpoints adopted in both cases will not strike the pupil. There is, however, no hurry; the better he understands "potential" in this context, the better he will be able to grasp the full significance of potential difference in electrical circuits.

With reference to electric fields, it is particularly easy to "define" potential difference; the potential difference between two points in an electric field is simply equal

/to.....

to the difference in potential between the points or, in terms of work per unit charge, it is the work done per unit charge against the field in transporting a small positive test charge at constant velocity from the one point to the other. These definitions of potential and potential difference are evidently operational, meaningful only in terms of the operations of transporting a small test charge and measuring the appropriate kind of work done on it. In fact, these concepts can be given satisfactory constitutive definitions in terms of the electric field vector; the potential at a point in an electric field is defined as the "line integral" of the electric field vector along any path from the point in question to infinity, and the potential difference between two points is the line integral of the electric field vector along any path from the one point to the other. It is readily seen that the previously-quoted operational definitions are operational interpretations of these line-integral definitions. The line integral of a force along a given path is the work done by the force when its point of application is displaced along that path (cf Sears and Zemansky 1964, Halliday and Resnick 1966), so that when the electric field vector is interpreted as "force per unit charge", it follows that the line integral of this vector must be equal to the work done by the electric field per unit charge. It would obviously be unreasonable to try explicitly to introduce the concept of a line integral into an elementary course, so there is nothing for it but to try to render the essence of this concept in terms intelligible to pupils. The best that can be done in this respect is to use the above operational interpretations of the line integrals, i.e. to "define" potential and potential difference in terms of work done per unit charge when a test charge is transported in the field. The "operational" overtones of such statements are regrettable, no doubt,

/but.....

but will just have to be accepted as a necessary part of the compromise solution to the present problem. We shall encounter a similar difficulty later when discussing E M F. (cf Section 7.17.4).

7.6.3 Potential difference and the electric field between charged parallel plates

The expression quoted in Section 7.6.1 for the work done in transporting a test change in a uniform electric field can be used to show that the electric intensity between a pair of parallel plates is equal to the potential difference between the plates divided by the distance between them. This fact can be made the subject of experiments with parallel-plate electric fields. Using the parallel-plate arrangement described earlier, in which the force on a suspended charged ball was measured, one can show the dependence of electric intensity on potential difference and on the separation between the plates. The "logic" of such an experiment requires a little thought. The experiment makes use of an E.H.T. supply to charge the parallel plates, with a voltmeter to monitor the potential difference between the plates. The pupil knows voltmeters from the junior course, in which he met them as black boxes which could measure energy transferred per unit charge between a pair of points in an electric circuit - "potential difference", in fact. In this experiment he can be shown that as the reading on the meter goes up, the force on the charged ball increases. This result can be regarded as showing that there is a connection between "potential difference" as defined for the electric field and "potential difference" as defined for electric circuits. With the potential difference between the plates kept constant and the separation of the plates increased, the force on the ball is reduced - something that should also be shown to the pupil.

/7.6.4.....

7.6.4 The surface of an isolated charged conductor is equipotential

In the above experiment use is implicitly made of the fact that the potential is the same at all points of an isolated charged conductor. At some stage this fact will have to be discussed with pupils; it might just as well be done here. The usual simple argument by which this result is established is based on the fact that the excess charge on an isolated conductor rapidly attains an equilibrium condition in which there is no mass motion of the charge. This means that there can be no electric field inside the conductor, so that there can be no potential difference between any pair of points inside the conductor. That there is no potential difference between any two points on the surface of the conductor is a result which can be rigorously proved with the aid of the fact that the work done by the electric field on a charged particle is zero when the particle is taken round a closed path. (For the purposes of the proof, the hypothetical path chosen has one section in the surface of the conductor and the other just beneath it.) But this proof need not be given to pupils, who can be given the simple argument above to establish that the interior of a charged conductor is an equipotential volume, and who can be reminded that the electric field patterns that they have seen demonstrated show that the field lines always leave the surface of an isolated charged conductor at right angles to it. From this experimental fact it is easy to argue that the surface of the conductor must have the same potential all over.

The essential point to make is that a charged particle moved along the surface of the conductor would "feel" no electrical force, as the electric field next to the surface is perpendicular to it. Therefore the electric field would

/do.:...

do no work on a charged particle moved along its surface, and no work would be done against the electric field by the "agent" moving the particle. From the definition of potential difference it then follows that the potential difference between any two points on the surface must be zero.

The comprehension of this result depends on a reasonably secure comprehension of the concepts of force as a vector, the electric field, work and potential difference. Thus although the argument is "easy", it will not be understood unless the ground has been carefully prepared. This is typical of the kind of challenge that electricity presents to the teacher.

7.6.5 The electroscope as a voltmeter

The fundamental role of the electroscope as a device which, in effect, measures the potential difference between its central system and its case, should be emphasized in the teaching of this concept. Using a simple diagram of the electric field inside a charged electroscope, it should not be difficult to convince pupils that there must be a potential difference between the central conductor and the case, and that this potential difference must depend on the amount of charge on the central conductor. With this fact established, the electroscope can also be used to investigate the potential difference between a pair of large, oppositely charged parallel plates. Provided the plates are sufficiently large, so that their capacitance even when widely separated is still appreciable compared with that of the electroscope, the well-known demonstration in which it is shown that the potential difference between the plates increases with increasing plate separation will work very well. The plates, when close together, are given their charges by induction; their separation is then increased, and it is shown that the deflection

/of.....

of the electroscope leaf increases. The difference between this situation and what happens when the plates are connected to the terminals of an E.H.T. supply should be made clear to the pupil. In the former case, the charges on the plates remain effectively constant, and the potential difference between the plates increases as the separation is increased. In the latter case, the power supply withdraws charge from the plates as the separation is increased so as to keep the potential difference constant.

7.6.6 Calculations on potential difference; the choice of a potential "zero"

Calculations on potential difference should include examples on the kinetic energy acquired by charged particles allowed to move freely in a uniform electric field. This idea is of great importance in other parts of the course besides those of purely electrical interest. The "electron-volt", the energy unit used so much in modern atomic, nuclear and particle physics, can conveniently be introduced at this point, although some teachers may prefer to defer this introduction until after Millikan's experiment has been dealt with (see Section 7.7).

The presentation of potential difference outlined above will have made it clear that the choice of zero for potential is arbitrary, since only potential differences are physically significant. However, pupils should know of the choices of zero of potential that are commonly made, viz. in fundamental electrical theory it is customary to specify that a point very far away (at "infinity", in fact) has zero potential, while in laboratory work it is sometimes convenient to take the earth as having zero potential.

/7.6.7.....

7.6.7 Further comment on the association of electrical potential energy with the electric field

Gledhill (1969) has suggested that there is an aspect of the association of electrical potential energy with the electric field that may need clarification. According to the conventional treatment of energy in the electric field (cf Feynman et al 1964), potential energy is already present in the electric field before one puts any test charge in it. (For macroscopic electric fields, the origin of this energy is readily attributable to the work done in setting up the system of field-creating charges.) How, if at all, is this relevant to the type of situation discussed in Section 7.6.1, in which one put a charge into the field and thereby, according to the field-energy viewpoint, gave energy to the field? The point is best explained with reference to a specific example, such as the field between a pair of equally and oppositely charged parallel plates. This field by itself can be regarded as having a certain energy which is not normally of any interest unless the field is somehow changed, e.g. by allowing the "capacitor" formed by the plates to discharge, when the "lost" stored energy can appear as "heat", or by pushing the plates further apart, in which case the work done on the plates transfers more stored energy to the field. If a positively-charged particle (e.g. a proton) is introduced into the space between the plates near the negatively-charged plate and then pushed towards the positive plate, additional energy is given to the field. Typically, this additional energy would be many orders of magnitude smaller than the energy originally present in the field. If the particle is now released, it accelerates towards the negative plate; using the "field" language, we would say that it acquires kinetic energy at the expense of the electrical potential energy stored in the field. (When it strikes

/the.....

the negative plate one is left with the original field, together with its original potential energy.)

We may sum up by saying that in terms of the field-energy picture, the introduction of a positive charge into an electric field alters the potential energy stored in the field. Displacement of the charge in the field will give rise to a further change in the field energy; for the right kind of displacement, this change will in fact be an increase. If the particle is then allowed to move under the action of the field alone, it will acquire kinetic energy at the expense of the electrical potential energy stored in the field.

If we prefer to discuss potential energy with reference to the field-creating and -responding charges only, then we must maintain that we can only have mutual potential energy of the system "field-creating charges plus field-responding charge" when that system has actually been set up, i.e. when the test charge has been placed in the field. Any further displacement of the charge in the field will alter this mutual potential energy; e.g. pushing the (positive) test charge from near the negative to near the positive plate will increase the mutual potential energy of the system. If the test charge is allowed to accelerate under the forces exerted on it by the other charges, then it acquires kinetic energy at the expense of this mutual potential energy. Usually it is both convenient and permissible to regard this mutual potential energy as being effectively associated with only the test charge (cf Section 7.6.1).

From this account it should be clear that a consistent description of energy changes associated with a charged particle in the field can be obtained with whichever viewpoint one chooses to adopt, "energy stored in the field" or "energy associated with the interacting charges". The former

/viewpoint.....

viewpoint requires us to accept that there is already potential energy in the field before we place a charge in it, but this energy subtracts out of any calculations we may do on energy changes suffered by the particle when it is displaced in the field. If we adopt the latter viewpoint, then there is simply no "initial potential energy" to subtract out, as there can be no mutual potential energy of the relevant system of interacting charges until all the charges are present. Either way, we arrive at the same result for the magnitudes of the energy changes undergone by the system.

It is worth stressing that acceptance of the fact that potential is a concept associated with a point in the field does not commit one to the view that electrical potential energy must be regarded as being stored in the field. With the aid of potential, a constitutively-defined concept for the field, we can compute the potential energy associated with the placing of a small test charge at a particular point in the field. Whether we regard that energy as associated with the field or with the interacting charges is a matter of taste. At least, this is the case in electrostatics; when we come to consider the radiation of energy from accelerated charges, we find that by far the simplest description of the processes of electromagnetic radiation is obtained with the aid of the field concepts, and that the radiation of energy finds a natural explanation in terms of the transport of energy through space by means of disturbances in the fields.

When the teacher introduces potential difference in the context of electrostatics, it is unnecessary for him to make much of the idea that potential energy can be thought of as being stored in the electric field. At this stage, the existence of electrical potential energy and the association of

/"potential".....

"potential" and "potential difference" with it are the most important points to establish. Once the pupil has acquired some understanding of these concepts, he can be led to consider the possible association of the potential energy with the field itself.

7.7 The "graininess" of charge

7.7.1 Introduction

The "graininess" of electric charge is not a concept which is fundamental to the development of the main ideas of this part of the course, as will have been evident from our own outline above. On the other hand, proof that electric charge really is "grainy" - that it always comes in "packages" of whole-number multiples of a natural unit of charge - lends considerable strength to the reality of the microscopic picture of electrical behaviour that pupils will have been encouraged to keep in mind. Therefore it is worth discussing as soon as the pupil has acquired sufficient understanding of basic electrical concepts to be able to appreciate the reasoning underlying the relevant experiment.

In the P.S.S.C. course the demonstration of the graininess of electric charge is actually placed prior to the explicit introduction of electric intensity and potential, in accordance with the P.S.S.C. "philosophy" of treating the electronic charge as a (temporary) unit of charge. We are of the opinion that the South African matriculation pupil will find it easier to follow introductory quantitative electrostatics if this topic is presented along the lines sketched out earlier in this chapter, and therefore recommend the placing of Millikan's experiment at this point. (It could be done even earlier, before potential and potential difference are introduced, but we think it will be advantageous for the teacher to have the concept of potential difference at hand when discussing the experiment.)

/There.....

There are various ways in which the need to do Millikan's experiment can be justified. It can be presented as an attempt to measure the very small charges that might be found on very small particles, with the "graininess" emerging as a surprise result. Or it can be presented as an attempt to investigate the hypothesis that matter contains very tiny particles, some of which carry electric charge. This was the reason why Millikan carried it out; he had been stimulated by powerful hints from electrochemistry and the work of J.J. Thomson and others to believe that there might be such a thing as an elementary charge. We shall not do our pupils any harm if we omit reference to the historical background, though there is no reason why teachers should not mention it if they wish. The experiment can be very happily accommodated in the course in its present position; indeed it is much better to deal with it here than in a segregated area styled "modern physics".

7.7.2 Films of Millikan's experiment and the associated theory

For reasons which are fully set out in the Nuffield Teachers' Guide (IV), it is in the last resort more satisfactory to show a film of Millikan's experiment to pupils than to have them do it individually. Anybody with personal experience of the experiment, even with one of the modern "simplified" versions that are now commercially available, knows that it is time-consuming and difficult to carry out and, as such, likely to put off all but the keenest pupils. Any "excitement" it generates is confined to the emergence of the desired result from the data; for some reason those who perform the experiment nowadays perversely refuse to be excited by the experience of looking at illuminated oil drops or plastic spheres through a microscope.

Two varieties of Millikan experiment are currently available on film. Both use plastic spheres instead of the classic "oil drops"; from the point of view of calculations

/which.....

which one might choose to carry out later, this is more convenient, as the masses of these plastic spheres have been previously determined (by a method which does not make use of this type of experiment; cf Nuffield Guide IV). In the form of Millikan experiment shown in the Nuffield film, the charged sphere is held stationary between the plates; when its charge is changed, the intensity of the electric field is changed until the sphere is stationary again. The theory of the method is as follows:-

Let E be the intensity of the applied electric field, V the potential difference between the parallel plates between which the field is set up, d the separation of the plates, m the mass of the sphere, q the charge on the sphere and g the acceleration due to gravity. Then when the sphere is balanced,

$$mg = qE = q\frac{V}{d}$$

so that $$q = \frac{mgd}{V} = \frac{(\text{Constant})}{V}$$

Hence observation of the changes in $(\frac{1}{V})$ is sufficient to disclose whether or not there is any "graininess" in the charges on the sphere. This method has the advantage that its theory is very simple to follow. However its use does depend on the use of a voltmeter to measure V and, with that, the acceptance by the pupil that the "charge" that flows around circuits is the same sort of stuff that one finds on isolated charged objects.

In the P.S.S.C. film of the Millikan experiment, the intensity of the electric field is maintained constant throughout the experiment, and the terminal velocity of the charged sphere is measured for different charges on it. As we show below, the changes in charge on the sphere are proportional to the changes in its terminal velocity under the action of the field. This removes the dependence on the voltmeter reading, but it does require of the pupil an understanding of

/the.....

the motion of bodies in a resisting medium which exerts a resistive force proportional to the speed. Such an understanding is not difficult to acquire if the basic principles are presented as experimental facts, e.g. by demonstration with ball-bearings falling through oil or glycerine. It should be shown that when a small sphere is acted upon by a constant driving force like its weight or an electrical force and moves through a fluid, it very soon attains a terminal velocity, and that this terminal velocity is proportional to the driving force (i.e. the resisting force is proportional to the speed). Then the treatment of the observations could be based on the following outline:-

Let m be the mass of the plastic sphere, q its charge, v the magnitude of its terminal velocity under gravity, u the magnitude of its upward terminal velocity when it is under the additional influence of an upward electric field of intensity E between the plates. Then $mg = Kv$

and $qE - mg = Ku$,

where K is the proportionality constant relating the force on the sphere to its terminal velocity. Combining these equations, one obtains $qE = K(u+v)$

which, for a constant intensity of the electric field, implies that the charge on the sphere is proportional to (u+v). (Note that this refers to the sum of the magnitudes of the terminal velocities). For any given sphere, v is a constant; consequently any changes of charge on the drop must be proportional to the changes in its upward terminal velocity.

In analysing the results of the experiments, one has to show that the changes in $\frac{1}{V}$ (in the first experiment) or in u (in the second) are integral multiples of some constant value. When explaining the deduction from this of the existence of a fundamental unit of charge, teachers may find it

/helpful.....

helpful to make use of the "marbles-in-paper-bags" analogy, described in various sources (Rogers 1960, Nuffield Guide IV, Jardine 1967; see also P.S.S.C. 1960, 1965).

7.7.3 Analogues of Millikan's experiment

If electricity is taught according to the sequence given in these pages, then teachers would probably find it better to use the second version of Millikan's experiment, as shown in the P.S.S.C. film, to demonstrate the existence of the natural unit of charge. To help prepare their pupils for the film, they could show them a macroscopic model of Millikan's experiment. The apparatus used for investigating the field between a pair of horizontal parallel plates would do for this purpose; though it shows a static version of the experiment, corresponding to that shown in the Nuffield film, it nevertheless helps to emphasize the presence of more than one force acting on the ball. In a more direct model, which could help to demonstrate the second version of the experiment, the parallel plates setting up the field are brass plates forming the lid and base of a transparent box with plastic walls (cf Jardine 1967). The box is filled with liquid paraffin, a few grains of sand are introduced into the liquid and a high potential difference is applied to the plates. Some of the grains will become charged by contact with the plates, and their motion through the liquid under the action of the electric and gravitational driving forces can then be studied.

7.7.4 Conclusions from Millikan's experiment

In interpreting the results of Millikan's experiment, it is important to state precisely what the experiment does show. It furnishes evidence for the existence of an irreducible minimum amount of charge. It does not by itself show that this charge is to be identified with the amount of

/charge.....

charge on the electron, though it certainly prompts one to think that <u>if</u> there are electrons, <u>then</u> they must surely carry the "natural unit" of charge. Pupils might be interested to know of the speculations among theoretical physicists concerning the existence of "particles", for which the name "quark" has been suggested, carrying amounts of charge equal to $\frac{1}{3}$ and $\frac{2}{3}$ the "natural unit" of charge (cf Weisskopf 1968).

7.8 <u>The distribution of charge on isolated conductors</u>

This topic has some interest, principally in the special case of the distribution of charge on a conductor having a pointed end or ends. The matter can only be dealt with qualitatively, partly by experiment and partly by the use of the concept of equipotential surfaces in an electric field. The pupil has already met this concept in noting the equality of potential over the surface and through the interior of an isolated charged conductor; its extension to the electric field surrounding such a conductor or to any static electric field should present no problems. To show that the charge distribution on a conductor is greater where the surface has the greatest curvature, one can show a diagram in which a set of equipotentials is drawn in the field surrounding a non-spherical charged conductor (Sears 1946). Near the surface of the conductor the equipotentials will have nearly the same shape as the conductor, while further away the equipotentials will be more nearly spherical in shape. The result of this is the crowding together of the equipotentials near the more strongly curved portions of the conductor, the most marked crowding occurring where the curvature is greatest. Crowding together of equipotentials implies a strong electric field, and this in turn results from a large concentration of charge. We conclude that the charge spreads itself over the surface of a conductor in such a way that the greatest concentration of charge occurs where the surface is most sharply curved.

/(cf.....

(cf. our remarks at the end of Section 7.4.5).

The manner in which the distribution of charge varies over the surface of a nonspherical conductor can be shown by sampling the charge at different parts of the surface with a proof plane and using an electroscope to indicate the amount of charge on the proof plane. The pupils can now understand the phenomenon of point discharge, which was demonstrated to them in the junior years and can now be demonstrated to them again (cf our discussion in Section 3.14.3, part (vi)).

7.9 Lightning

The incidence of thunderstorms is high in South Africa, and it is thus not surprising that there should be a distinguished tradition in South African physics of research into the origin and nature of the lightning discharge. It is thus fitting that this topic should have been included in the new Common Basic Syllabuses in Physics and Physical Science; the following summary indicates the points that need to be made in its presentation.

The most spectacular electric sparks known to man are those we call "lightning discharges". Many of the details of the processes responsible for the formation of lightning are still obscure, but it is possible to give some account of the matter in broad outline. An electric spark passes between the domes of a Van de Graaff generator when sufficient amounts of opposite charge have been built up on the domes. With these large amounts of opposite charge is associated a very strong electric field and a correspondingly large potential difference between the domes. The "breakdown" of the air between the domes can be explained along the lines of the discussion in Section 3.14.3, part (vi).

Since a lightning stroke is a gigantic electric spark, we may conclude that a very large potential difference is

/responsible.....

responsible for it. This is indeed the case; potential differences of the order of 20 - 100 million volts may be generated between the base of a thundercloud and the earth. In the formation of a thundercloud large-scale separation of positive and negative charge takes place by a mechanism as yet incompletely understood, although it is in some way dependent on the behaviour of the air currents, water drops and ice particles in the cloud, and it is the very large amount of negative charge at the base of the thundercloud that makes possible the "breakdown" of the air in its vicinity. (A recent summary of modern ideas in this connection is to be found in Feynman et al (1964).)

The success of the "point discharge" phenomenon in preventing catastrophic electric discharges is easily demonstrated with a Van de Graaff generator and a pointed metal object held in the hand. This demonstration should already have been done in the junior years; it should be repeated here, with emphasis on its practical importance.

7.10 Energy in the electric field: capacitors

The capacitance of an isolated conductor or pair of conductors (capacitors) is not a concept which need be introduced into an elementary physics syllabus, since such a syllabus need make use of no quantitative results which must be derived explicitly with the aid of this concept. On the other hand, the capacitor is a useful device for a pupil to know about, and for this reason it is a pity that the new Common Basic Syllabuses make no mention of it. In the context of these syllabuses, the capacitor would be useful in demonstration experiments on account of its ability to store charge and energy. It can be introduced as a device consisting of a pair of conductors similar in shape and placed close togetner but separated from each other by some insulating

material. As examples of capacitors the pupil can first be shown a pair of parallel plates (with air between them!) and that old stand-by, the Leyden jar. He should see the latter charged up from a Van de Graaff generator and then discharged with discharging tongs, and the spark he sees in the latter case should be discussed in terms of the energy released in it as well as "flow of charge". To expand on the idea that the energy for the spark was stored in the electric field between the capacitor "plates", reference can be made to the demonstration in which the potential difference between a pair of oppositely-charged parallel plates was shown to depend on their separation. It can be pointed out that work has to be done to separate the plates, that this represents electrical potential energy apparently transferred to the charges on the plates, and that the volume occupied by the electric field between the plates increases as they are separated. It is easy to argue that the view that the potential energy transferred to the plates is stored in the electric field between them is consistent with the above facts; if the energy in this uniform field is assumed to be uniformly distributed through its volume, then we would expect increased stored energy to be associated with a larger volume of field.

7.11 The principle of the Van de Graaff generator

The (model) Van de Graaff generator is now recognized to be an indispensable piece of apparatus in the elementary teaching of electricity, and we have recommended its use in a number of important demonstrations. Up to this point, this device will have remained for pupils a "black box" known to be capable of accumulating appreciable amounts of (opposite) charge on its domes. Its design is of such elegant simplicity that there is no need for it to remain wholly "black"; much can be said concerning its mode of operation

/that.....

that would be intelligible to pupils. In the typical school model of the Van de Graaff generator (e.g. Griffin and George L81 - 280), the belt passes over two rollers made of different insulating materials, e.g. perspex and polythene. Contact between the belt and the rollers produces opposite charges on the two rollers, positive on the perspex and negative on the polythene. The positive charge on the perspex roller induces negative charge on the points of the metal comb nearest it; similarly the negative charge on the other roller induces positive charge on the points of the other metal comb. As a result of point discharge, negative charge is sprayed on to the outside of the belt from the comb nearest the perspex roller and positive charge is sprayed on to the belt from the other comb. If the perspex roller is the lower of the two, the net effect of this is that negative charge is transported from the lower to the upper comb.

A certain amount of controversy surrounds the details of the transfer of charge from the belt to the dome and, in conjunction with this, the fundamental theory of the operation of the instrument. Thus Simon (1954) has objected to the statements often made in textbooks that the charged belt discharges completely to the outside of the dome, that all the charge transferred to the upper comb must find its way on to the outside of the dome and, were it not for insulation difficulties, the potential of the latter could be increased without limit. He points out that for charge to pass from the belt to the upper comb by point discharge, there must be a potential difference between the two, and that for this reason alone it can be concluded that not all of the charge can be transferred from the belt to the dome. He stresses that the charge on the belt acquires increased potential energy because it is transported to the upper comb against an

/adverse.....

adverse electric field, and he develops a theory for the action of a generator which predicts a limiting voltage attainable for its upper dome, determined by the geometry of the machine. Subudhi and Tiwari (1962) report experimental results in agreement with Simon's theory. Fortunately the details of this theory need not concern the teacher; nor need he make any serious modifications to the standard account of the operation of the generator. Certainly it would be wise for him to make the very important point that as charge is transported towards the upper comb, its potential energy must increase, so that there must on this account be a potential difference between the upper region of the belt and the lower (this will be negative if the transported charge is negative); and it would be unwise for him to suggest that all the charge on the belt finds its way on to the dome. But the limit set by Simon's theory on the potential attained by the dome is never attained under laboratory conditions, since "insulation difficulties" always cause breakdown beforehand; accordingly the teacher need not worry particularly about the existence of this theoretical limit. (According to Simon, it arises because of the charge induced on the lower comb by the charge on the dome.) For all practical purposes, it is true that all of the charge transferred to the upper comb finds its way on the outside of the dome (cf Faraday's Ice-pail experiment).

7.12 The bridge between static and current electricity

7.12.1 The pupil's background

The pupil has studied current electricity to some extent in the junior years. There his understanding of the properties of electric circuits was built up on the assumption that there was some kind of electrical "stuff" circulating around the circuit. This hypothetical "stuff" was called "charge";

/its.....

its rate of flow past a given point in the circuit was called the "current". The current was measured in terms of the magnetic effect it produced. The unit of current was called the ampere, and it was established in terms of measurements with a "current-balance", which was used to calibrate more convenient current-measuring instruments called ammeters. The unit charge was defined in terms of the ampere and the second and called the coulomb. The concept of potential difference between two points in a circuit was defined as the energy given up per unit charge by charge flowing between those points; its unit, the joule per coulomb, was given the special name "volt".

In the senior years, the pupil begins his electrical studies with the force between stationary charges, where by "charge" is meant the property by virtue of which a body is able to attract light objects of any description. As a unit of charge, the coulomb is introduced, this time defined (temporarily) in terms of the force between charged particles at rest relative to each other. These ideas lead on to the concepts of the electric field and its characteristic vector; consideration of the energy of a charged particle in an electric field leads to the emergence of the concepts of potential and potential difference, defined in terms of potential energy per unit charge - and thus having as their unit the joule per coulomb, or volt.

7.12.2 An experimental approach

Before the pupil goes on to pursue a higher-level study of electric currents and circuits and the magnetic effects associated with them, he needs to develop a unified outlook on electrical phenomena. To be able to manage this, he needs to attain the conviction that the "charge" he supposes to be flowing around circuits is the same kind of "charge" as

/produces.....

produces readily observable effects of another kind when it is at rest on isolated conductors or on insulators. The ideal way of establishing the identity of the two kinds of charge is by means of Rowland's experiment (cf Shamos 1959, Purcell 1965). Rowland placed electrostatic charge near the rim of a hard-rubber disc and showed that when the wheel was rotated rapidly, a magnetic field was obtained in its vicinity similar to that which would result from a very small current flowing in a circular loop. Unfortunately, it is not a practical proposition to carry out this experiment in the school laboratory (cf Nuffield Guide IV). Nevertheless there are other experiments which can readily be carried out and which furnish evidence which, when collected together, points to the same conclusion as that deduced from Rowland's experiment.

First, pupils can be shown that batteries, electronic power supplies and the A.C. mains can charge up electroscopes. (The more batteries there are and the higher the voltage they have, the better - four or five 90-volt batteries connected in series will produce a substantial deflection in an ordinary leaf electroscope.) This experiment is valuable in showing that the black boxes which can maintain a current in a closed circuit can push "electrostatic" charge on to electroscopes and presumably on to other conductors to which they may be connected. If, as was suggested above, an E.H.T. supply is used in demonstrations to create a manifestly electrostatic field between a pair of parallel plates, pupils should not need much convincing on this point.

Next it can be shown that batteries, power supplies and the A.C. mains can charge up capacitors of the commercial type, e.g. paper and electrolytic capacitors. All that is needed here is to connect up the terminals of the capacitor to the terminals of the charging device, and then to show that discharge of the capacitor through an air-gap produces

/a.....

a spark. By itself this does not tell us a great deal, but it is an essential preliminary step to the next and much more important experiment, in which the flow of charge from a slowly discharging or charging capacitor is used to produce a marked magnetic effect. It is essential for the success of this experiment that the "time-constant" of the charging or discharging process should be long, so that the effects can be long-lasting rather than momentary. It is also essential that the charging or discharging current should be large enough to produce a detectable magnetic effect. Both these requirements can be met if the following conditions are satisfied:

(i) The capacitor should have a very large capacitance - e.g. 1000 μF.

(ii) There should be a large resistance in the circuit - e.g. 20 000 ohms. With a 1000 μF-capacitor, this would give a time constant of 20 seconds, which is quite long enough.

(iii) A capacitor of the above capacitance would typically have a voltage rating of 30 volts or less. If it were to be charged up through a resistance of 20 000 ohms from a D.C. supply of 20 volts, or charged up to a potential difference of 20 volts and then allowed to discharge through a resistance of 20 000 ohms, the maximum current in either case would be 1 milliamp. This current is large enough to produce an appreciable deflection in demonstration ammeters of the right kind, and the experiment can be done with such meters if they are available. But an even more direct way of showing the magnetic effect of the charging or discharging current is to let it pass through the coil of a tangent galvanometer. If these instruments are to be found in the school laboratory, they are likely to be provided with several sets of coils, one of which will have several hundred turns.

/A.....

A current of 1 milliamp in a coil like this will readily cause a deflection of the order of nearly 90° in a pivoted magnet mounted at the centre of the coil. In effect, this would then be Oersted's experiment carried out with the charging or discharging current of a capacitor. This "poor man's Rowland's experiment" is both very easy to carry out and very convincing in effect.

To reinforce the conclusion of the previous experiment, one can perform a simple experiment with a Van de Graaff generator (preferably one having a variable-speed motor), and a microammeter. In the simplest form of this experiment, one can connect the leads of the microammeter to the upper and lower combs of the Van de Graaff generator, turn the motor on and direct attention to the fact that there is a reading on the microammeter when the belt is moving, and that the reading grows bigger as the motor speed is made audibly faster. This makes the point that an increase in the rate of flow of electrostatic charge from the lower to the upper combs is accompanied by an increase in the "current" registered by the ammeter. To make the experiment more spectacular and to help convince pupils that there is electrostatic charge flowing right round the "circuit" that includes the combs and belt of the generator, the ammeter and its leads, one can make the above arrangement more elaborate. The main dome of the generator can be put back in its rightful place, and instead of connecting the appropriate lead of the microammeter directly to this dome, one can connect it to a sharp-pointed conductor, the point of which aims at, but does not touch the main dome. If a neon discharge tube is available, it can be connected between the sharp-pointed conductor and the microammeter lead. Then when the generator is running, there will be a continuous flow of charge from the dome to the point by "point discharge", the neon discharge tube will

/glow.....

glow and the microammeter will give a reading. It will certainly be difficult for pupils to resist the conclusion that charge is indeed streaming continuously all the way from the main dome through the air, the neon tube and the meter down to "earth".

Once the pupil has accepted that "electrostatic charge" and "charge flowing in circuits" are one and the same thing, he is ready to study the application of the concepts appropriate to the electric field to the study of the flow of charge in various kinds of conductor and in complete circuits.

7.13 Elementary consideration of electrical conduction as a movement of charges

7.13. 1 An introductory experiment

It is now time for the process of flow or transport of charge to be studied in some detail. A fundamental question to ask is "How does a cell (or any other "black box" that achieves the same purpose) maintain a steady current in a conductor?" In reply, one need not say anything about chemical processes within the cell. Instead, a simple experiment can be performed that will help to orient pupils' thinking suitably. For this experiment, the apparatus needed consists of a battery (preferably a "battery of batteries", such as several 90-volt batteries connected in series), a 10-megohm resistor and an electroscope. (The reason for choosing a value of resistance as high as 10 megohms is that this is high compared with the internal resistance of the "battery", which can be appreciable if the batteries are old.) First it should be demonstrated that the resistor is a conductor, by showing that when the electroscope is charged, it can be discharged by placing the leads of the resistor simultaneously in contact with the knob and the case of the

/electroscope.

electroscope. Then the leads of the battery should be connected to the knob and the case of the electroscope, and, while the battery is still connected, the leads of the resistor should also be secured to the knob and the case. The leaf of the electroscope will continue to deflect, showing that the battery is able to maintain a steady potential difference across the resistor.

This much can readily be seen and understood. To complete the picture of what is happening in this demonstration, certain additional points should be made. First, the steady deflection of the leaf of the electroscope while the resistor and the battery are connected across it shows that the battery is able to maintain opposite charges on the electroscope knob and case, despite the presence of a conducting path between them. It is reasonable to deduce that what is happening in this case is that charge is continuously flowing from knob to case (or vice versa) through the resistor, and continuously being "pumped around" by the battery to the knob and case (positive in the one case, negative in the other), to replace what is lost through "leakage" down the resistor. Of course these ideas are implicit in the acceptance of the view that an electric current is a literal flow of charge, but it is just as well for them to be made explicit in reference to an experiment like the one above, particularly as the experiment highlights the maintenance of a potential difference across the current-carrying conductor.

7.13.2 A particle model of electrical conduction

After this introduction, attention can be given to the provision of a more detailed answer to the question asked at the beginning of Section 7.13.1. The situation illustrated by the above experiment can be idealized for purposes of discussion in the following way:-

/Imagine.....

Imagine that we have two spherical conductors, X and Y, each mounted on insulating stands; X is positively and Y negatively charged. Suppose that a piece of wire is connected between X and Y; there would be a momentary current in the wire as charge flowed along it to neutralize the charge on X and Y. If the free charge carriers in the wire were negatively charged, they would be repelled from Y and attracted towards X; and if X has a deficiency of negative charge carriers and Y an excess of them, then the excess charge carriers on Y would flow into the wire, while there would be a corresponding flow of negative charge from the wire on to X. If the flow of charge along the wire were to be kept up, it would be necessary to maintain positive charge on X and negative charge on Y. The continued presence of charges of opposite sign on X and Y would mean that an electric field was maintained in the wire and that there was a potential difference between its ends. We can regard the field as supplying the force that drives the free charge carriers down the wire, and can therefore state that as an essential condition for the maintenance of electric current in a conductor an electric field must be maintained in the conductor. (This can be compared with the case of an isolated charged conductor within which there is no electric field.)

For the preservation of opposite charges on X and Y charge will have to be supplied to them to replace what is lost (e.g. negative charge supplied to Y and removed from X). Pushing negative charge on to Y which is already negatively charged requires the performance of work, as does the removal of negative charge from or the supply of positive charge to the positively charged X. Electrical potential energy is thus being transferred to the free charge carriers by these operations, which immediately leads one to ask two questions:-

/ "What.....

"What is the source of this energy?" and "What becomes of it?" Since energy cannot be created from nothing, the electrical potential energy given to the electrons must be made available from some other form of energy. If a battery is the agent used to maintain the current, the energy is converted from chemical form. If X and Y are the main and auxiliary domes of a Van de Graaff generator, the electrical energy is converted from the mechanical energy driving the belt. Pupils could be invited to suggest and discuss other sources of electrical energy that are in common use. An instructive experiment in this connection is one in which a hand-operated dynamo is shown to need more energy to run it when it is delivering a current than when it is not (cf Nuffield Guide I).

7.13.3 The particle model and electrical heating

What becomes of the energy? If the charge carriers moved in a vacuum, the answer is simple: it becomes kinetic energy of the charge carriers. Now we know that current-carrying wires heat up. The accounts given in elementary textbooks of the origin of this "Joule heating" tend to be perfunctory. If such textbooks attempt an "explanation" of this phenomenon at all, they naturally do so in terms of the simple particle model which we have described above. They point out that the kinetic energy acquired by free electrons between successive collisions with ions of the lattice is given up to these ions in collisions, so that the energy acquired by the free electrons from the electric field becomes "thermal energy", i.e. energy associated with the lattice vibrations.

Frequently, this is all that is said about the process of energy transfer. Cf P.S.S.C. (1960,1965): "...as the charge makes collisions with the atoms of the wire, the

/energy.....

energy just goes into thermal energy heating that section of the wire"; Kip (1962): "The process is one in which electrons are accelerated by the applied field but continually lose their excess energy by collisions with the lattice of ions. This heats up the lattice by increasing the energy of thermal vibrations". The intelligent pupil, who is supposed by now to have some knowledge of the mechanics of collisions, may be curious to know more than what is contained in these quotations about the nature of the electron-ion collisions. Are they elastic or inelastic, for example? Sears and Zemansky (1964) state bluntly that the collisions are <u>inelastic</u>: "After momentary acceleration, the particle makes an inelastic collision with one of the fixed particles in the conductor, loses whatever velocity it has acquired in the direction of the driving force, and makes a fresh start". This makes it very easy to explain the transfer of energy; in view of the great difference in mass between the electron and a lattice ion, virtually all the kinetic energy of the electron becomes "heat", and that is that. But in an inelastic collision, the "heat" that is generated is <u>internal energy</u> of the system formed by the colliding particles, whereas the particle model of Joule heating requires that the energy received by the ions must become kinetic and potential energy associated with their <u>vibratory motion</u> and not any "internal motions" they might have. On this model, then, inelastic collisions between electrons and the lattice ions do not make sense. The collisions can only be elastic, which appears to raise the difficulty that we would normally expect very little energy to be transferred from an electron to an ion in a perfectly elastic collision because the mass of the ion is so much greater than that of the electron. This, however, applies to collisions between moving electrons and ions at rest. In the words of Purcell (1965), "...The average transfer of kinetic energy is small when a billiard

/ball.....

ball collides with a bowling ball. Therefore the ion (billiard ball) will continue to accumulate extra energy until its average kinetic energy is so high that its average loss of energy in a collision equals the amount gained between collisions." This explanation, which we have seen only in Purcell's book, is the one that should be given to the inquiring pupil who wants to know more of the process of Joule heating than is customarily given in simple accounts. The "true" story, concerned with electron-phonon interactions, is much more complicated; it is not realistic to try to discuss it in a school course.

One further point is perhaps worthy of comment. Some textbooks say a little of what happens to the ions when they acquire their extra vibrational energy. Halliday and Resnick (1966), for example, say that the collisions "increase the amplitude of the thermal vibrations of the lattice", while Atkins (1965) says that "the copper atoms vibrate more rapidly." Of these descriptions, the first is certainly to be preferred. The second presumably means that the mean kinetic energy of the vibratory motion is increased, but could be misinterpreted as meaning that the frequency of vibration goes up. This would not be justified in terms of this model of electrical conduction.

7.13.4 A useful analogy for electrical conduction

Halliday and Resnick (1966) and McLachlan and McNeill (1964a) both give an analogy for the maintenance of current in a circuit which is far more helpful than the water analogy. We suggest that teachers who wish to reinforce their teaching of electric currents at this level with the use of an apt analogy would do well to consider this one, which we shall describe in some detail.

/A.....

A man lifts balls from the floor to a shelf. The balls roll off the shelf into a vertical cylinder filled with viscous oil; they quickly attain their terminal velocity in the oil, sink to the bottom and are removed from the cylinder by a trapdoor mechanism. They then roll along the floor to the man's feet and are lifted up again. Here the oil-filled cylinder corresponds to the piece of wire in our idealized experiment above and the balls to the free charge carriers. In lifting the balls the man gives them gravitational potential energy, the energy coming from his own internal chemical energy; this corresponds to the battery removing charge carriers from one end of the wire and supplying them to the other, providing the necessary electrical potential energy from its internal chemical energy. The balls gain negligible kinetic energy in sinking through the oil, so that the gravitational potential energy they have been given is converted to internal energy of the oil and its container; this process is analogous to the generation of heat by the loss of electrical potential energy by the free charge carriers drifting down the wire. The process in which the man lifts up the balls and sends them on their way is <u>circuital</u>, i.e. the balls are kept moving around a closed circuit. Similarly a wire and a battery to which its ends are connected form a closed electric circuit in which the battery, in effect, takes free charge carriers from one end of the wire and transfers them to the other, from which point they drift down the wire again.

It will be noted that while we have made free use of the concepts of energy, charge and the electric field in the foregoing account, we have made no appeal to formulae or to mathematics of any kind. Nor have we had to make any use of the water analogy and the misleading terminology associated with

/it; .....

it; indeed it would do the pupil no harm if he forgot all about the water analogy from this point onward. The analogy we have described is seen to correspond closely with our particle model of electrical conduction; the one respect in which the correspondence is not so close is that in the gravitational analogy, the balls move in the gravitational field set up by the earth, whereas in the electrical circuit, the electric field in the wire is set up by concentrations of opposite charge at its ends. But this does not constitute an effective objection to the use of the analogy, which illustrates nearly all the important points that have to be made in a readily graspable manner. If the pupil can understand the above semi-quantitative account of the maintenance of a current in a simple electrical circuit, he will be well prepared for the quantitative treatment that will be given later in the course.

Jardine (1967) describes a simple working model which is closely related to the analogy just discussed. A vacuum cleaner is used to blow ping-pong balls up a vertical plastic tube to a plastic box with a hole cut in its base large enough to allow one ping-pong ball to fall through it at a time. The falling ping-pong balls knock a paddle-wheel on the way down, so that the kinetic energy they acquire is partly transferred to rotational kinetic energy of the paddle wheel, which is thus enabled to do various kinds of job. The balls finally fall into an inclined chute which leeks them to the base of the vertical tube in which they are once again blown up. This model has the advantage of providing a visible analogy for the transfer of energy from the electrons to useful work, and not just into "heat" as in the analogy described in the previous paragraphs.

Understanding of the role of the electric field in a conductor in maintaining a current in it might be aided if

/this.....

this electric field could be demonstrated in some way. Jefimenko (1962, 1966) has described a simple technique for showing such fields. Transparent two-dimensional printed circuit models of various conducting systems are made on glass plates using a transparent conducting ink. A power-supply giving a voltage of about 10^4 volts is applied to the model circuit and fine grass seeds are strewn over it. If the teacher does not think it worth the trouble to prepare the demonstrations, he can use Jefimenko's photographs of his own patterns, which are to be found in the references cited. They show up quite clearly the field pattern both inside and outside current-carrying conductors; in particular, the field inside is seen to be parallel to the sides of the conductors.

7.14 The nature of the electric current in metallic conductors, electrolytes, ionized gases and electron and positive ion beams

7.14.1 Introduction

It would aid the pupil in picturing the conduction process and add to his interest in the whole matter if more precise information could be given about the nature of the particles that carry the charges in typical instances of electrical conduction. As the P.S.S.C. showed in its new physics course (1960, 1965), much can be done in this direction without any serious departure from a logical presentation of the subject. Certainly there is a departure from a historical order of presentation, involving also the transplantation of certain topics conventionally dealt with in other sections of a physics syllabus, but as we have argued earlier, this fact does not in itself invalidate this approach.

/7.....

7.14.2 Metallic conductors

Experiment shows that whatever they are, the "electric particles" in a metal normally remain within the metal since, as the P.S.S.C. textbook points out, "a charged piece of metal in a vacuum holds its charge indefinitely". The possibility of persuading these particles to escape from the metallic surface by heating can be investigated in the experiment conventionally used to show the action of a thermionic diode. The EB91 diode made by Philips is very suitable for this experiment, since it will give plate currents of a few milliamps for a very modest (i.e. less than 10 volts) plate voltage. From this kind of experiment the following facts would emerge:-

(i) when the filament is "cold", there is no current in the plate circuit of the diode;

(ii) when the filament is heated by passing a current through it, there will be a current in the plate circuit, provided that the positive terminal of the battery in the plate circuit is connected to the plate and the negative terminal to the filament (or cathode).

From these facts one can infer that heating the filament boils "electrical particles" out of it, and that these particles must carry a negative charge. By themselves, these facts do not establish beyond doubt that all the free charge carriers in a metal must be negatively charged, or even that electric charge is transported by particles and not by some kind of fluid. But we can say that if there are "electrical particles" in metals, the ones that are boiled out carry negative charges. To simplify matters, we make the plausible guess at this stage that the charge carriers are particulate in nature and that those in the metal are all like the ones boiled out when the metal is heated; and we call the hypothetical particles electrons. We can then see how far this guess will take us. From other experimental facts we can

/immediately.....

immediately draw a further conclusion about our hypothetical "electrons". No chemical change is observed in either the filament or the plate of a diode as a result of the flow of electrons from filament to plate, even when these electrodes are made of different metals; none of the material of the filament is found to be deposited on the plate; when charge is allowed to flow from one wire to another of different metal, no change is observed in either wire as a result of the passage of a current, no matter how long the process is allowed to continue. All this suggests that electrons, whatever they are, are the same in any metal.

Since the results of the Millikan experiment are already to hand, we are in a position to argue that since electric charge is particulate in nature, it is plausible that the hypothetical "electrons" really are discrete entities, each carrying an amount of charge equal to the natural unit of charge. This idea could be checked if we could arrange to measure total charge on a known number of electrons. If one could produce a concentrated beam of electrons, it would be possible in principle to carry out this experiment. Therefore the production of electron beams can usefully be studied now. (N.B. We would not insist that electron beams should be introduced in this particular way; they could equally well be studied for their own sake as a form of charge flow.)

7.14.3 Electron beams

Since electrons are known to pass from filament to plate in an evacuated glass envelope when the filament is heated and the plate is made positive with respect to the filament it would appear that a narrow electron beam might be obtained if a small hole were drilled in the plate; presumably some of the electrons attracted towards the plate must go through

/the.....

the hole instead of striking the plate. This is indeed the case, and electron beams are produced in exactly this way in the conventional cathode-ray tube, of which the wall on which the electron beam finally impinges is covered with fluorescent material. Therefore, though the beam itself cannot be seen, the point at which it strikes the screen is rendered visible as a bright spot. Nowadays special "electron deflection tubes" are available in which an electron beam is produced in the manner described above, but instead of it impinging on a fluorescent screen it is allowed to "graze" past a suitably-mounted fluorescent screen, so that most of the path of the beam is rendered visible. In our opinion, pupils should preferably see both forms of electron-beam demonstration, the cathode-ray tube because it is the central constituent of the cathode-ray oscilloscope, and the "deflection tube" because it shows the path of the beam. Pupils should see the beam deflected both by an electric field applied between the deflecting plates and by the charge on an insulating rod held nearby. If a "fine-beam" tube is available, it could also be used at this point. This is a speical type of hot-filament discharge tube in which the entire path of the cathode-ray beam is rendered visible by the light from excited hydrogen atoms in the path of the beam. The apparatus is expensive, but the demonstration is so beautiful and enhances the teaching of the properties of electrons and of the effects of electric and magnetic fields so tremendously that the acquisition of a fine-beam tube would be a most worthwhile investment. It is significant that the Nuffield Physics Course makes a great deal of use of this piece of apparatus.

The discussion of the charge carried by electrons can be rounded off with a brief account of the principles of the experiment by which the average charge carried by an

/electron.....

electron can be measured (cf P.S.S.C. 1960, 1965). Such an experiment requires the use of a beam of electrons sufficiently weak for the arrival of individual electrons from it to be detected. The scintillation produced by the impact of an electron on a fluorescent screen would be too small to be detected by the naked eye, but it can be detected by a suitable photoelectric device or by an "electron multiplier." By the use of such devices, the number of electrons arriving per second from the beam can be counted. (Note that this type of experiment provides confirmation of the discreteness of electrons.) Then the same beam can be aimed into a metal bottle which acts as a collector of electrons. The excess negative charge acquired by the bottle per second can be measured; when this result is divided by the number of electrons arriving per second, it gives the <u>average</u> charge carried by each electron. This average charge is found to be identical within the limits of experimental error with the natural unit of charge; since we apparently cannot have a smaller amount of charge than the "natural unit" on anything, we conclude that the <u>actual</u> charge on the electron is the natural unit of charge.

7.14.4 Current in electrolytes

One can now pass on to consider other modes of charge transport, since it has been shown with reasonable certainty that the conduction of electricity in metals proceeds by motion of free electrons. This is perhaps not the best place to deal with charge transport in semiconductors; instead, the donduction of electricity in electrolytes can be dealt with, as the arguments that have to be used here can be based on easily demonstrable phenomena. In an experiment to demonstrate the conduction of electricity by a liquid, it will be found that if conduction does proceed to any marked extent,

/there.....

there are observable reactions at the electrodes dipping into the liquid. It is not clear how to account for these phenomena if charge transport in conducting liquids (electrolytes) is supposed to take place by movement of free electrons. (It has already been pointed out that chemical effects are not observed when a current is maintained in a circuit made up of wires of dissimilar metals.) On the other hand, if the liquid contained some number of positively and negatively charged ions, the observed behaviour at the electrodes is readily explicable. By ions are meant "molecules" (using this term in a wide sense) that have excess positive or negative charge; if we suppose that these ions can attain neutrality by gaining or losing electrons, we can give the following explanation of the phenomena associated with electrolytic conduction:-

In the external wires, electrons are being driven from the negative terminal of the battery to the cathode and pulled from the anode to the positive terminal. Within the electrolyte, the electric field set up between the electrodes by connection to the battery causes motion of positive ions towards the cathode and of negative ions towards the anode. The details of what happens when the ions reach the neighbourhood of the electrode are complicated and need not form part of a school physics course. What can be said is that whatever occurs at the electrodes must have the result that electrons are given up at the anode and removed from the cathode; each second, sufficient negative ions must lose their excess electrons at the anode and sufficient positive ions must gain the electrons they lack at the cathode, so that a steady potential difference may be maintained between the electrodes. In this way, the net rate of transfer of charge inside the electrolyte would equal the rate of charge

/transfer.....

transfer in the external portion of the circuit. The chemical reactions of the electrodes are the means whereby the liberation or removal of electrons there is accomplished. (cf. Gledhill (1961): "Electrons do not dissolve in water but ions do".)

If the pupils have sufficient chemical background, the above discussion of electrolytic conduction can be extended quantitatively so as to provide definite evidence in favour of the above picture of the conduction process. The necessary features of this background are the existence and significance of the "atomic weight" (relative atomic mass, and "equivalent" of an element. Provided these are understood, the conclusions to be drawn from the following "thought experiment" (cf P.S.S.C. 1960, 1965) should not present much difficulty:-

Connect electrolytic cells having platinum electrodes and containing hydrochloric acid, copper sulphate solution and silver nitrate solution respectively in series with a battery and allow a current to be maintained in the circuit for some arbitrary length of time. If the masses of copper and silver deposited and hydrogen liberated are determined, it will be found that these masses are in the ratio of the equivalents of copper, silver and hydrogen. When this fact is taken in conjunction with the significance of the equivalent and the view that charge is transferred by ions in liquids, one arrives at the possibility that the charges carried by the copper, silver and hydrogen ions are small integral multiples of a fundamental unit of charge. By itself, this experiment simply suggests that the <u>average</u> charge transported by these different types of ion is an integral multiple of a natural unit (cf Lewis 1964), but when we already know from Millikan's experiment that there is a natural unit of charge we can entertain a strong suspicion that the

/ions.....

ions really do carry charges in integral multiples of a natural unit. And this fits in with the idea that positive or negative ions in an electrolyte have become ions through the gain or loss of the right number of electrons.

7.14.5 Current in gases

In discussing the conduction of electricity in gases, one can start with the observation that under normal circumstances, the tendency of gases to conduct electricity is very slight indeed; this is shown by the fact that on a dry day a charged electroscope will retain its charge for a considerable time. Evidently the molecules of the air must be electrically neutral, and it can be shown in the same way that this must be true of the molecules of other gases as well. The presence of ions and of unattached electrons is the essential requirement for the establishment of a current in a gas. The ions may be formed with the aid of some outside agency (X-rays, radioactive source) or by the action of a strong electric field set up between the electrodes in a gas discharge tube. Once ions and unattached electrons are present, the transport of charge between the electrodes can proceed in the manner described above in connection with the electric spark (cf Section 3.14.3, pt. (vi)).

An important point to make in connection with the conduction of electricity in gases and in liquids is that in both cases the flow of charge takes place in two directions simultaneously; positively-charged particles move towards the negative electrode while negatively-charged particles move towards the positive electrode. Furthermore, the flow of negative charge in one direction is equivalent to the flow of positive charge in the opposite direction, so that the contributions to the current from the positive and the negative

/ion.....

ion flow must add. Gaseous conduction differs from electrolytic conduction in that electrons as well as ions participate in charge transport within a gas and no chemical reactions accompany the removal or liberation of electrons at the electrodes.

7.14.6 Positive ion beams

The knowledge of the processes of gaseous conduction can be used to explain the production of positive-ion beams. Suppose that we have a gas discharge tube so arranged that virtually all the enclosed gas is to be found between a pair of parallel plates which form the electrodes, and that there are small holes in each of the plates. When a strong electric field is applied to the electrodes, we can expect some negatively-charged ions to pass through the hole in the anode, and some positively-charged ions should pass through the hole in the cathode. The latter ions form a positive-ion beam. In the first experimental studies of positive-ion beams or "positive rays", as they were then called, these beams were produced in the manner just described. Nowadays it is more usual to obtain positive-ion beams by firing a beam of electrons at the neutral gas molecules and allowing those which then become positively charged to be attracted towards a negatively-charged plate with a hole in it (cf Rogers 1960.).

7.15 The magnetic field, treated quantitatively

7.15.1 Reasons for a quantitative treatment of the magnetic field in a school syllabus

There are a number of reasons why the magnetic field should be treated quantitatively, even in an elementary physics syllabus. They are:

(i) The magnetic field is an important part of fundamental physics; as such it deserves at least an indication of how it can be treated quantitatively.

/(ii).....

(ii) If the magnetic field vector is not introduced, one cannot begin to give a proper quantitative account of electromagnetic induction - which would be a pity, since that phenomenon produces such readily measurable effects (and is so important in our everyday lives).

(iii) Without the $\vec{B}$ vector, one cannot properly explain the principles on which an important method of measuring the mass of elementary (charged) particles and ions is based, so that a large area of modern physics will remain closed to the pupil.

7.15.2 Should the approach to "B" be "amperian" or "point-pole"?

There are probably few today who would dispute that it is important to include some quantitative treatment of the magnetic field in an elementary syllabus, for the reasons cited. But there is perhaps room for disagreement concerning the way in which this treatment should be approached. It is generally admitted that the magnetic field is more difficult to understand than the electric field. Consider, for example, the establishment of its fundamental field vector $\vec{B}$. If one does this by consideration of the "magnetic" force on a current-carrying conductor, one has to carry out two steps; first, one has to point out that there is a particular direction in a given region of the field along which we can align our conductor so that it will experience no force - this we call the direction of the field vector (having adopted some agreed convention about the way in which the "arrow" of the vector should point); second, the force on the conductor is found when it is at right angles to this direction (i.e. when it has its maximum value) and the magnitude of the field vector is specified in terms of the magnitude of this force, the current in the conductor and the

/length.....

length of the conductor. Compare this with the simplicity of the establishment of the electric field vector, where we place a small positive test charge at a point in the field, measure the magnitude and note the direction of the electric force on the charge and "define" the electric field vector as the force per unit charge on the test charge. The "definition" of the magnetic field vector is almost bound to strike the beginner as clumsy and awkward; as Chraplyvy (1952) remarks, it "is probably the weakest point of the so-called Amperian scheme of teaching electricity and magnetism."

But this is only the beginning of the difficulties that beset this approach. Suppose one wants to go on to derive the standard formula for the magnitudes of $\vec{B}$ set up at specified points in the fields of certain types of current-carrying conductors. To do this, one makes use of the Biot-Savart formula for the contribution to $\vec{B}$ from a "current-element"; but, although this will give the right answer, it does not alter the fact that a "current element" is a mathematical fiction whose use in the present context we ought at least to try to justify. Justification can be achieved by reasoning from the results of Ampere's experiments on the forces between current-carrying conductors, but the chain of reasoning is long (cf Tricker 1955, 1965; Shire 1960, Feather 1968) and not particularly easy to follow.

In view of these difficulties, advocates of the "point-pole" approach to magnetism might seem to have something of a case. On this approach, one develops "magnetostatics" to begin with in an exactly analogous fashion to electrostatics: a "Coulomb's Law" is established for the force between a pair of "point poles" (mathematical fictions, these, but let that pass for the moment), and "magnetic field strength" is defined in terms of the force per unit pole strength on an isolated point north-pole. If one wants to go as far

/as.....

as the Biot-Savart formula, one can derive it by making use of the experimentally established similarity between current-carrying circuits and magnetic shells and following a long but reasonably straightforward line of reasoning.

To discuss which of these approaches is best in a thorough-going treatment of electricity and magnetism would go beyond our terms of reference. The problem is fully and fairly discussed by the Coulomb's Law Committee of the American Association of Physics Teachers (1950), who suggest that there are advantages and disadvantages to either approach and conclude that "either approach is 'right' and the student who goes on in physics must eventually learn both languages". What does concern us here is which approach is "right" in an elementary presentation; and here it seems to us possible to come down firmly on one side of the fence. In elementary physics, there is no pressing need for the pupil to know how to find the force or couple exerted by one magnet upon another; on the other hand he does need to know the force exerted by a magnetic field <u>on</u> a moving charge, for without it he cannot understand the orbital motion of a charged particle in a magnetic field. The route to the expression for the "magnetic" force on a charged particle is long and tortuous via the point-pole approach; apart from anything else, it necessitates the introduction of a new "fundamental" concept, that of "pole strength", with the attendant problems of the units in which it should be expressed, and, if one follows the traditional mode of approach, a transition has to be made from the vector $\vec{H}$ to the vector $\vec{B}$. If one introduces $\vec{B}$ via the "Amperian" or "force-on-a-conductor" approach, the route to the force on a moving charge is short and direct. (Indeed it is possible to reach the desired expression in one step if one chooses to "<u>define</u>" the magnitude of $\vec{B}$ in terms of the

/force.....

force on a moving charge, as is done in some treatments. But this approach will not satisfy us if we want to introduce $\vec{B}$ with the aid of simple laboratory measurements.) Clearly we can avoid the difficulties associated with the introduction of the Biot-Savart formula in this approach if we so design our syllabus that there is no need to use this formula. In an elementary syllabus it is of primary importance to know how to measure $\vec{B}$ in a given field once it has been set up, and to be able to predict how a charged particle will behave if projected into the field; it is of less importance to know the quantitative details of how a given field may be set up. It is convenient for the pupil to know that the strength of the magnetic field around a long current-carrying wire falls off inversely as the distance from it, and that the field inside a long current-carrying solenoid is uniform; but these results can be established experimentally.

The conscientious planner of an elementary syllabus must be very much concerned with the kind of foundation which it will provide for future studies in the subject, if any. From this point of view, how does the use of the "Amperian" approach to the magnetic field appear? Such an approach places emphasis on $\vec{B}$ as the fundamental field vector, rather than on $\vec{H}$, where the latter is defined as the force-per-unit-pole-strength (where pole strength is measured in "Kennelly" units - cf. Coulomb's Law Committee 1950, Halligey 1966), and is rooted in the view that all known magnetic effects have their origin in electric currents of one kind or another. Since Ampere's hypothesis "that magnetic phenomena are due solely to electricity" (quoted in Tricker 1965) can, in principle, be checked by experiment, it is to experiment that we must turn for a decision on whether or not the Amperian approach to magnetism is really the more fundamental.

/First,....

First, we note that experiments to detect magnetic monopoles have so far proved unsuccessful (cf Ford 1963, Kolm 1968): if such entities do exist, they are very elusive. There are sound theoretical reasons for believing that magnetic monopoles could exist. As they stand at the moment, Maxwell's equations show an asymmetry between electricity and magnetism: the divergence of the electric field vector is equal to the charge density times a constant, while the divergence of the magnetic field vector is always zero; in the expression for the "curl" of $\vec{B}$, there is a term involving the electric current density, whereas in the corresponding expression for the curl of $\vec{E}$, there is no term involving a magnetic current density. Symmetry between electricity and magnetism could be restored by the plugging of these holes in the equations, and this would require the existence of isolated magnetic charges analogous to isolated electric ones. Dirac (1931) advanced arguments from quantum theory for the existence of such magnetic monopoles, and predicted a value for the amount of magnetic charge carried by them. In recent years determined efforts have been made to find magnetic monopoles. Various teams of research workers have looked for them in elementary-particle reactions produced by the most energetic accelerators available, in incoming cosmic rays, in magnetite deposits and in sediment from the ocean bed. All these efforts have so far proved unsuccessful, but the search continues. As Ford (op.cit.) remarks, "It can safely be predicted that physicists will go on searching until they either find magnetic monopoles or discover a good reason for their non-existence." In the meantime we should point out that even if the search for magnetic monopoles is ultimately successful, this need have no direct effect on the teaching of elementary electromagnetism. Magnetic monopoles are not needed to explain any of the presently known phenomena in electromagnetism;

/in.....

in particular, there appears to be little reason to suppose that isolated magnetic poles or charges may be found among ordinary matter, where by "ordinary matter" we mean matter ordinarily accessible to the experimenter.

Some twenty-five years ago, this last statement would have been challenged in at least one quarter. Ehrenhaft (1942, 1944 etc - there is no point in citing references to all the many papers he published on the subject) claimed to have observed "isolated magnetic ions" in the course of investigation of the phenomenon of "magneto-photophoresis" - (cf also Ehrenhaft and Banet 1942). Very small particles of elements such as Fe, Ni and Sb were suspended in a homogeneous magnetic field and irradiated with light; some moved towards the north "magnetrode" and others towards the south. Ehrenhaft also claimed that some particles would move in the field even when they were not irradiated by light; according to him, the effect of light was to liberate magnetic charge of one or other sign from the suspended particles, leaving them with a net magnetic charge; those particles that behaved like magnetic ions in the absence of light presumably were already the possessors of an uncancelled magnetic charge. Although Ehrenhaft published many papers in which he reiterated his claim to have discovered isolated magnetic poles in ordinary matter, he did not win much support. Instead, his claims were contested by other workers, who were unable to detect "magnetic ions" in circumstances in which, according to Ehrenhaft, they ought to have been detectable, or who pointed out that the effects observed by Ehrenhaft could be explained without recourse to the hypothesis of "magnetic ions". Hopper (1944) reported observing thousands of nickel particles suspended in a homogeneous magnetic field and finding that none of them moved in a way consistent with the "magnetic ion" hypothesis. G.Kane (1947) pointed out that some

/suspended.....

suspended particles could well be set in motion as a result of the current induced in them when the magnetic field was turned on, and Balkenkohl (1954) attributed the "Ehrenhaft effect" - motion of the particles in the field when irradiated with light - to a radiometer effect, the particles experiencing radiometer forces directing them along the field lines. In short, Ehrenhaft's experiments cannot be regarded as providing convincing support for the view that magnetic charges can be found in ordinary matter, and few people besides Ehrenhaft appear to have taken them seriously in this connection. The whole episode seems now to have been completely forgotten.

The second point to be considered in connection with the status of the Amperian approach concerns the problems raised by the origin of ferromagnetism and by the fact that the neutron, which is electrically neutral, has a magnetic moment. It cannot yet be claimed that ferromagnetism is fully understood, but enough is known about it for one to assert that the phenomenon has its origin in the "spins" of certain of the electrons in the material or, more specifically, in the intrinsic magnetic moments of these electrons. Only experiment can tell us whether the magnetic behaviour of microscopic entities like neutrons or "spinning" electrons can be more fittingly represented by picturing these entities as tiny permanent magnets (e.g. as uniformly magnetized spheres) or as tiny "current whirls" (e.g. as uniformly charged rotating spherical shells). If the latter picture is appropriate, then it can be predicted that a very fast-moving charged particle passing through a ferromagnetic specimen will experience a deflecting force proportional to $\vec{B}$ rather than $\vec{H}$, and that the behaviour of neutrons reflected and diffracted by ferromagnetic specimens will likewise be determined by $\vec{B}$ rather than by $\vec{H}$. (We are referring here to the "macroscopic" $\vec{B}$ and $\vec{H}$; cf Section

/3.....

3.16.2). Experiment has confirmed these predictions; in connection with the reflection and diffraction of neutrons, it has shown that both the neutrons and the ferromagnetic domains of the specimen behave as if they were infinitesimal "current whirls" and not permanent magnets. (These ideas are discussed in detail by Shire (1960); cf also Panofsky and Phillips (1962) and Phillips (1966).)

Experiment, then, favours the view that the Amperian approach is fundamentally correct and, in line with this, that $\vec{B}$ rather than $\vec{H}$ is the fundamental vector for the magnetic field. Therefore we need have no misgivings about starting a beginning pupil off on the Amperian road. As matters now stand in the teaching of intermediate and advanced electricity and magnetism, it is more than likely that a pupil who goes on to study physics at university will find magnetism taught there along Amperian lines; certainly this is the case in South Africa. As the Coulomb's Law Committee (1950a) puts it, "For the Amperian approach the broadest argument is that it proceeds as directly as is practicable for a freshman toward the physical principles adequate for a satisfactory comprehension of the whole range of magnetic phenomena including electromotive forces." The force of this argument is being more and more appreciated at university level. With a tendency now becoming evident of relating electric and magnetic fields with the aid of special relativity even in intermediate treatments (cf Sears and Zemansky 1964, Feynman et al 1964 and Purcell 1965), the emphasis on the Amperian viewpoint in university presentations of the subject is likely to become even more marked.

In view of the considerations advanced above, it is not surprising that recently-developed syllabuses in elementary physics (e.g. P.S.S.C., Association for Science Education and Nuffield) should have adopted the Amperian approach to magnetism. (The Nuffield Physics Course includes only the

/barest.....

barest minimum of quantitative magnetism, but what there is concerns the measurement of B by "force-on-a-conductor".) Following this lead, the new South African syllabuses for physics and physical science for the Joint Matriculation Board excluded "point-pole" magnetostatics altogether, concentrating instead on those topics in (electro)magnetism which are essential for the understanding of charged-particle motions in a magnetic field. The new Common Basic Syllabuses have virtually taken over these syllabus outlines as they stand, so that the Amperian approach is now firmly entrenched in South African physics teaching.

7.15.3 Preliminary ideas to the introduction of "$\vec{B}$"

As has been mentioned, the major difficulty in teaching magnetism along Amperian lines concerns the establishment of the $\vec{B}$ vector. In the present course, however, the pupil comes to this quantitative study of magnetism with a background of acquaintance with magnetic field patterns and with the knowledge that current-carrying conductors do experience forces when they are not aligned with the direction of the field. The Coulomb's Law Committee has recommended that "Enough attention must be given to the [field] lines, as lines, to make the student feel at home among them before any problem is raised about magnitudes." In effect, this recommendation is fulfilled for the South African pupil who has studied physical science in standards Six, Seven and Eight. With his qualitative familiarity with magnetic fields and their effect on current-carrying conductors, he should be ready to absorb the equation relating the strength of the magnetic field to the force on a current-carrying conductor held at right angles to it. It is advisable that before formal teaching of magnetism begins in the secondary school years, the pupil should see the relevant demonstrations once again.

/No.....

No doubt there is no uniquely satisfactory way of presenting the elements of (Amperian) magnetism at this level; various differing approaches are to be found in the literature. In the P.S.S.C. course the first steps toward quantitative magnetism are taken by showing that magnetic fields add vectorially. At this stage no rule is laid down for the measurement of the strength of the magnetic field; it is assumed that a current-carrying loop produces a field at its centre equal in strength to that of the earth if it exactly cancels the latter when directed opposite to it, and that two equal loops produce twice the field of a single loop, i.e. that the field strength produced at a given point in the field of a current-carrying conductor is proportional to the current. A compass needle is used to indicate the direction of the resultant field. The advantage of beginning in this way is that the pupil is not yet brought face to face with the relationship between the direction of the field and the direction of the force it exerts on a current-carrying conductor, and yet he does begin to develop a feeling for the quantitative aspects of the magnetic field. Yet the idea of field strength introduced like this may seem a little too abstract for comfort. In our view, there is no need to proceed in this way for pupils having the background of a South African senior secondary school pupil (which P.S.S.C. pupils have not, since they are assumed not to have met physics before.).

An alternative way of introducing quantitative magnetism is to show from considerations of the force between current-carrying conductors how a unit of current can be defined in terms of this interaction. Once one has placed the measurement of current on a proper footing, the concept of current can be used freely in the subsequent development of the subject,

/particularly.....

particularly in the actual definition of the magnitude of $\vec{B}$. This is the order of procedure suggested by the Coulomb's Law Committee (1950b) and by implication recommended by those responsible for drafting the new Joint Matriculation Board and Common Basic Syllabuses in South Africa. Yet another method is to show <u>first</u> how the direction and magnitude of $\vec{B}$ may be defined, without using an official definition of current, but assuming that n conductors parallel and next to one another and all carrying the same current will produce n times the magnetic effect of one of the conductors. The definition of current can then be dealt with in due course. Since a choice has to be made, we shall describe the first of the approaches suggested in this paragraph; but we appreciate that in this instance, the case for proceeding in this fashion rather than according to the other approach is not overwhelmingly strong.

7.15.4 <u>Force between parallel current-carrying wires: definition of the ampere</u>

Although this may already have been done in the junior years, it is advisable here to carry out demonstrations of the forces between current-carrying circuits or circuit elements, and to "explain" these forces by means of the argument that current-carrying circuit A must set up a magnetic field which will exert a force on any part of current-carrying circuit B that is placed in it. It is especially important to demonstrate the forces of attraction or repulsion between long, parallel current-carrying wires. The best demonstration apparatus for this purpose would be a two-wire current balance, e.g. that manufactured by Welch (cf Williamson 1958), with which the effects of the attractive or repulsive forces can be made very obvious by amplification with an optical lever. A cheaper version of a two-wire current balance is described in the 1966 Report of the Association for Science Education on the Teaching of Electricity (Association for Science

/Eudcation.....

Education 1966). Ampere's own apparatus for demonstrating the forces between parallel currents is also illustrated in this publication; it should be suitable for qualitative demonstration of the phenomenon. Alternatively one could show the force of attraction or repulsion between a pair of adjacent long, narrow, vertical strips of aluminium foil through which a current of several amperes is passed.

One's first important quantitative task in this section is to show how the force between a pair of parallel, current-carrying conductors depends on the current in them, the length of the conductors and the distance between them. Ideally, this should be treated experimentally but, even if suitable apparatus were available, the task would be far from easy, even for the experienced teacher. With currents of the order of 10 amperes, the forces between a pair of parallel conductors that are some 25 cm long are only of the order of 10^{-3} N when the centres of the conductors are separated by 2 or 3 mm. Thus it would take extremely careful experimental work to show convincingly that the force between the conductors was inversely proportional to the distance between them. Nevertheless, it should be possible to show that the "law of force" for the two conductors is more closely approximated by an inverse dependence on the first power of the distance between them than on the second; and this is, after all, the most important aspect of this force law. (Later, the dependence of the field strength of a current-carrying wire on the distance from the wire can be shown by means of a simple experiment to be of the form $B \propto \frac{1}{r}$.) The dependence of the force on the current in the conductors and on their lengths can be established by means of plausible arguments as follows:-

(i) Dependence on current Suppose that one of the wires under (hypothetical) test forms part of a very long wire in which a current is maintained. Then to obtain what amounts

/to.....

to twice the current in the wire under test one could take the long wire and bend it into the shape of a long flat coil having two turns, one part of the coil being straight and having the appropriate length. This procedure can easily be extended to give an effective current of any reasonable multiple of the original current. (The validity of this assumption is justified by the fact that current is a rate of flow of charge; two wires carrying the same current and placed next to each other will give twice the rate of flow of charge past a cross-section through both wires as is obtained for a single wire.) A length L of the straight portion of any of the turns composing the large flat coil will exert the same force on a given length of the <u>other</u> wire as would a length L of any of the other turns. <u>If</u> the Principle of Superposition applies to forces of this type, it follows at once that the force exerted on a given length of the other wire is proportional to the <u>effective</u> current in the flat coil; and from this we can conclude that the force on any length of either of a pair of parallel current-carrying conductors must be proportional to the product of the currents in the conductors.

(ii) <u>Dependence on length</u> If we assume, as seems reasonable, that the magnetic field around either "long wire" is cylindrically symmetrical, then we can conclude that the force exerted <u>on</u> any one portion of one of the wires is the same as that on any other portion of the same length; from which we conclude that the force <u>on</u> a given portion of one of the wires is proportional to its length.

(Rogers (1960) gives a more sophisticated argument than the above in order to establish the general form of the force law between two parallel current-carrying elements, which ought to be considered by those who are interested in pursuing

/the.....

the Amperian approach further with advanced pupils. The dependence of the force between the elements on the distance between them is "guessed" to be an inverse-square one.)

All these conclusions may be summed up in the statement that the force per unit length on either of two long, parallel, current-carrying straight wires is proportional to the product of the currents in the wires and inversely proportional to the perpendicular distance between them. For the special case of the same current in each wire, one obtains the following formula for the force F on length L of either of two long parallel, straight wires separated by a distance d:

$$F = k_m \frac{i^2}{d} L$$

(i is the current in each wire). If the unit of current is not yet defined, the value of the constant of proportionality k_m is a matter of arbitrary choice. By international agreement, the above formula yields the current in amperes if d and L are measured in metres, F in newtons and k_m is set equal to exactly 2×10^{-7} newton per (ampere).^{2} The formal definition of the ampere runs: "The ampere is that constant current which, if maintained in two, parallel rectilinear conductors of infinite length, of negligible cross-section, and placed at a distance of 1 metre apart in vacuo, would produce between those two conductors a force equal to 2×10^{-7} newton per metre length." (cf South African Bureau of Standards 1968).

Several comments about this definition are in order:-

(i) Like the definitions of the other fundamental SI units, e.g. the metre, the kilogram and the second, the definition of the ampere states a relation between a concept (the ampere) and a "thing" or set of experimental circumstances. According to Bunge (1967), such statements are properly called "referitions", which sets them apart from true definitions, which state concept-concept correspondences. Pupils need not be troubled with this semantic subtlety (cf Section 3.2.6).

/(ii).....

(ii) The reference to parallel conductors of <u>infinite</u> length may at first seem to remove the definition from the realm of reality. This is not really a serious difficulty, however. From fundamental electromagnetic theory it is possible in principle to derive a formula for the force of interaction between any pair of current-carrying circuits or conductors of any shape or relative orientation. The resulting formula will take the form of the product of three terms, viz. the product of the currents in the conductors, the arbitrary constant k_m and a factor depending on the size, shape and relative orientation of the conductors. When the conductors are supposed to be straight, parallel and infinitely long, the formula takes on the very simple form quoted above. For a realistic arrangement of conductors between which the force could be accurately measured, the formula, if exact will look much more formidable, but it will still take on the general form just described. If the same current were present in both conductors, this formula could be used to deduce the value of the current from the measured value of the force provided that a value has been chosen for k_m. Thus the real point of the official definition of the ampere is that <u>it fixes the value of k_m</u>; it does so with reference to the idealized situation of two parallel and infinitely long conductors, because only in this case is it possible to identify the proportionality constant that appears in the formula for the force with the arbitrary constant k_m that appears in all such formulae for the force between two current-carrying conductors. (Strictly speaking, the "arbitrary constant that appears in all such formulae" is conventionally put equal to half our k_m - it is usually written as " $\frac{\mu_c}{4\pi}$ " - but the pupil's attention need not be drawn to this.) It will be necessary to clarify this situation for the pupil, who might otherwise be disturbed by the thought of "infinitely

/long.....

long" conductors.

(iii) The peculiar-seeming value of 2×10^{-7} newton per $(\text{ampere})^2$ was chosen to make the new, internationally adopted ampere agree with the old ampere, which was defined as one-tenth of the absolute electromagnetic unit of current. Its presence in the definition of the ampere means that this definition conflicts with Lipson's idea of what such a definition should be (see earlier, section 7.4.4). But experience suggests that pupils, students and instructors can live with this definition and suffer remarkably little as a result. South African teachers may well feel that they can live much more easily with "2×10^{-7}" than "0.001118", which was the number that appeared in the only definition of the ampere that could be found in South African textbooks before 1965.

The coulomb can now be <u>officially</u> defined in terms of the ampere and the second as the quantity of charge that in one second crosses a cross-section of a circuit in which there is a constant current of 1 ampere. In the junior secondary school, the coulomb was also defined in this way, but there the definition of the ampere was temporary. The pupil may now wonder how he could carry out measurements to show that the officially defined coulomb has the same size as that temporarily defined with the aid of Coulomb's Law, or, what amounts to the same thing, how the value of the constant in Coulomb's Law could be <u>measured</u>. This question should certainly be answered for him: it is best discussed in full after the theory of electrical circuits has been discussed (cf Section 7.18). Another point on which the pupil may require assurance is how one can be certain that current measured in terms of the interaction between current-carrying conductors really does represent a true measure of "rate of flow of charge". As was noted above, we made use of the fact that current is rate of flow of charge in "deriving" the

/fundamental.....

fundamental formula of this section. A direct check that a "current balance" current really is rate of flow of charge is provided by Faraday's First Law of Electrolysis, which can be discussed and interpreted later.

Pupils should gain some experience in measuring a current or calibrating an ammeter with a parallel-conductor type of current balance. It should be made clear to them that the simple formula quoted above cannot apply accurately to these current balances: e.g. the forces between straight wires of finite length include "edge effects" arising from the ends of the wires. (The error introduced by this amounts to a few per cent only). The results of a calibration experiment would not be very valuable as calibration, but they would help to strengthen the confidence of the pupil in the validity of the "current balance" formula and in his own ability to take measurements somewhat more difficult than the average. (If desired, the two-conductor balance to be used in these experiments can be of the parallel-coil type: Seamer (1958) has given a description of such a balance that can be constructed from two single-turn coils of wire, one of which is connected to the baseboard of the balance and the other suspended above it from the balance beam.)

7.15.5 Introduction of the $\vec{B}$ vector

Since the pupil has had to study the forces between current-carrying conductors in some depth, it is logical for the introduction of the magnetic field vector to make use of this newly-acquired background. Proceeding along similar lines to those followed in the introduction of the electric field, the teacher can suggest that if current-carrying conductor X exerts a magnetic force on current-carrying conductor Y, this interaction can be pictured as follows:

(i) Since X has current in it, a magnetic field is set up in

/the.....

the surrounding space;

(ii) Y experiences a magnetic force because it is in the magnetic field of X. (Point (i) could be emphasized by once again demonstrating the magnetic field of a current-carrying conductor with the aid of iron filings.)

Whatever one decides to use as a measure of the strength of the magnetic field in the vicinity of Y, it must be related to the force on Y. To find out about this force and so provide a proper basis for the pupil's understanding of the $\vec{B}$ vector, some time should now be spent in experimental investigation of the force experienced by a current-carrying "element" in a magnetic field. A convenient and practical current element is a small strip of aluminium foil suspended fairly loosely between the prongs of a wooden yoke. The magnetic field to be explored by this "element" can be set up by a current-carrying coil or by a strong permanent magnet or electromagnet. This demonstration experiment should establish that when a current is maintained in the element,

(a) the element obviously experiences a force of non-mechanical and non-electrostatic origin when it is held in some part of the magnetic field;

(b) in any given region of the field, the element can be held in such a way that it experiences no magnetic force;

(c) the direction along which the element lies when condition (b) is realized is the same as that in which a compass needle would point when placed in the region, i.e. when the element is held in the direction of the magnetic field, it experiences no magnetic force;

(d) the magnetic force on the element seems to be biggest when the element is at right angles to the field direction.

One conclusion that can reasonably be drawn from these observations is that the maximum possible magnetic force experienced by the current element in a given region of the

/magnetic.....

magnetic field depends on the current in the element and the field itself; therefore measurement of this force would provide useful information about the magnetic field. At this point pupils could carry out the experiment of measuring the magnetic force on a current-carrying conductor. The latter could be in the shape of a U, so mounted that it forms one arm of a balance, enabling the magnetic force on the cross-piece of the U to be measured. Alternatively the U-shaped conductor could be suspended from one arm of a balance. If the balance is of the form in which the length of the cross-piece of the U can be varied (e.g. by using a set of "U's" of different widths), so much the better. (Ideally, one would also like to be able to vary the angle between the cross-piece of the U and the direction of the field, but this may not be possible to arrange). A permanent magnet or a solenoid can be used to supply the field, which should be as nearly as possible uniform over the width of the U.

This experiment is of fundamental importance. If done with sufficient care, it should show that the force on the element
current/is proportional to the length of the element and to the current in it. (There is no harm in using an ammeter to measure the current, as the ammeter may be assumed to have been calibrated by means of a current balance of the type mentioned in the previous section.) If it is possible to investigate the dependence of the force on the angle between the field and the conductor, it would be found that the relation between the force F and the angle θ was given by

$$F \propto \sin\theta$$

If the current and the length of the cross-piece of the U are denoted by i and L respectively, the results of these experiments show that

$$F \propto iL \sin\theta$$

or $$F = BiL \sin\theta$$

/Since.....

Since F, i, and L and θ can all be measured, the value of the proportionality constant B can be found; it is obviously characteristic of the magnetic field in the region in which F is measured, and can, indeed, be thought of as representing the "strength" of the magnetic field in this region. Now the magnetic field vector $\vec{B}$ may be "defined" as a vector whose magnitude is the value of the constant B in the above equation and whose direction is what has already been called "the direction of the magnetic field."

In the preceding sentence, we have enclosed the word "defined" in quotation marks because similar problems to those considered in Section 7.5.3 arise in connection with the definition of $\vec{B}$. As in our discussion of the definition of $\vec{E}$, we must conclude that $\vec{B}$ is in the last resort only "definable" by means of the axioms it obeys in electromagnetic theory. Although we can lay down a procedure of measurement for the magnitude of $\vec{B}$ and a means of specifying its direction, these do not add up to a definition of $\vec{B}$ in the strict sense of the term though they do, of course, constitute an "operational definition" of $\vec{B}$. As was the case with $\vec{E}$, the establishment of this "operational definition" is a crucial step in the teaching sequence, but the result does not have to be called a definition.

In its simplest form, the defining equation for the magnitude of the magnetic field vector is

$$F_{\text{maximum}} = iBL,$$

where F_{maximum} is the force on a current element of length L when a current i is maintained in it and it is held at right angles to a uniform magnetic field of strength $\vec{B}$. It would be as well to point out to pupils such similarities as exist between the operations that have to be carried out to find the magnitudes of the electric and magnetic field vectors. To find the magnitude of $\vec{B}$, one measures the force

/per.....

per unit length per unit current on a small current-carrying conductor placed at right angles to the field. It is reasonable that "force per unit length" should be the corresponding quantity for the magnetic field to "force" for the electric field, since the quantity in question must clearly be independent of the length of the conductor.

Up to this point we have been content to refer to $\vec{B}$ as "the magnetic field vector" or "the strength of the magnetic field". As is well known, the name "magnetic field strength" (or "magnetic intensity") has traditionally come to be associated with the vector $\vec{H}$, while $\vec{B}$ is most often known by the name given to it by Maxwell, viz. "magnetic induction", and sometimes by the term "magnetic flux density". "Magnetic induction" is an unfortunate name to use in an elementary course, in which it seems altogether pointless and, moreover, can be confused with the phenomenon of magnetic induction. "Magnetic flux density" would not convey any useful meaning at this stage, since magnetic flux has not yet been defined. We feel it best simply to call $\vec{B}$ the "magnetic field vector" or the "magnetic field strength" and forsake one's allegiance to history. The teacher who is worried about the effect this might have on pupils who will be continuing their studies at university, can always cover his tracks by mentioning the traditional names and pointing out that some still use them. In any case, there is a tendency for writers of textbooks nowadays to disregard the historical nomenclature for the magnetic field; the influential textbooks of Feynman et al (1964) and Purcell (1965) both call $\vec{B}$ simply "the magnetic field."

From the equation $F_{max} = iBL$, it is evident that the units of B are newtons per ampere-metre. The name "tesla" has been officially adopted for this unit and pupils should certainly be familiar with it. However, it is important for

/them.....

them to know the dimensions of B, and occasional use of the name "newton per ampere-metre" for the unit can serve as a reminder of these. Pupils should also hear of the gauss, defined as 10^{-4} of a tesla and a convenient unit for many practical purposes.

In summing up these facts concerning the magnetic field teachers should stress that it is possible (in theory though not always in practice) to dispense altogether with compass needles and iron filings as exploring tools in magnetic fields and that, in describing what one means by the magnetic field, one has no need to refer to such objects at all.

To be able to use a current-carrying element as a means of finding the direction of $\vec{B}$, one cannot avoid making use of the convention relating this direction to those of the current in the conductor and the force experienced by it. If, therefore, the teacher wishes to present a treatment of the magnetic field which makes some attempt at being conceptually independent of the behaviour of the compass needle - and we think this is desirable - then he will have to introduce one or other of the "hand rules" mentioned in our earlier discussion (see Section 3.16.3). This does not mean that he need depart from the principle enunciated earlier of not over-emphasizing the importance of these rules. The point here is that if we use a current-carrying conductor instead of a compass needle to tell us the direction of $\vec{B}$, we need a rule; and if we need a rule, we ought to know what it is. One should not make too much of the rule; e.g. one should not give it a prominent role in problems that might be set. One form of hand rule which, in the author's experience, goes down well with a class is Iona's FBI (or FBi!) gun rule (Iona 1962). Iona refers to "a generation of American students and teachers to whom the ordered sequence of letters FBI is quite familiar". American influence is sufficiently

/prevalent.....

prevalent in South Africa, it would seem, for this sequence to be familiar to South African pupils and students also. The letters F, B and I are respectively associated with the thumb, forefinger and second finger of the left hand if one wants to know the direction of the force on a current-carrying conductor in a field. (They can be associated with these fingers of the right hand if one wants to know the direction of the current induced in a conductor that is moving in a magnetic field.) (Note the occurrence of the letters "f" in "left" and "i" in "right".)

$\vec{B}$ is, of course, an axial vector or pseudovector, as opposed to a "true" vector or polar vector. That is, reflection of the coordinate axes through the origin changes the sign of $\vec{B}$, which it would not do to a true vector such as the electric field vector $\vec{E}$. The "axial" nature of $\vec{B}$ is readily deducible from the specified relation between the vectors $\vec{F}$, $\vec{B}$ and $\vec{L}$. ($\vec{L}$ is a vector drawn in the direction of the "current" (i.e. "current density") along the conductor on which the force acts.) The direction of $\vec{F}$ is related to those of $\vec{B}$ and $\vec{L}$ by a vector product rule ($\vec{F} = i\vec{L} \times \vec{B}$); the vector product of two "true" vectors must be an axial vector, because it is defined by means of a (right)-"handed" rule; but since $\vec{F}$ and $\vec{L}$ are both true vectors, and $\vec{F}$ is expressible in terms of the vector product of $\vec{L}$ and $\vec{B}$, it follows that $\vec{B}$ must be an axial vector.

The fact that $\vec{B}$ has this special character need not affect the teaching of the subject directly, in the sense that there is no need at this level to mention that $\vec{B}$ is a different kind of vector from the ordinary. Conventional elementary and intermediate treatments of electromagnetism manage very well without mentioning that $\vec{B}$ is an axial vector, even though they might have introduced $\vec{B}$ by means of a vector

/produet.....

product rule. Inevitably, however, the axial nature of $\vec{B}$ - its association with a "handed" rule - does have a bearing on what we may teach about electromagnetic interactions. The main point here - and this is something that could and should be said in class - is that the convention by which the direction of $\vec{B}$ is made to agree with a left-hand rule or a right-hand screw rule is only a convention. This particular choice of rule for specifying the direction of $\vec{B}$ accords with the direction in which the north-seeking pole of a compass needle points, but we could just as well have agreed on an opposite convention whereby we use a right-hand "N-S" rule or a left-hand screw rule. This would have turned all our $\vec{B}$'s round, but made no difference to the forces or accelerations we would measure (cf Feynman et al 1964).

Another way of looking at the situation is the following:- Suppose we imagine a "mirror-image" version of a particular electromagnetic experiment, such as the deflection of a charged particle in a magnetic field. Among the effects of the reflection of the experiment in the mirror is that right-hand rules for axial vectors must be replaced by left-hand ones, and vice versa, so that in the mirror-image space, $\vec{B}$ would be specified by a left-hand rule if it had been specified by a right-hand rule in real space. It turns out that Nature is so constructed that the mirror reflection does not change any observed electromagnetic phenomena. Therefore if a charged particle travelling in a given direction in ordinary space is deflected to "the left", say, by the magnetic field, a similar charged particle travelling in the same direction (i.e. parallel to the corresponding axis) in mirror-image space will also be deflected to the "left". Thus the change from right-hand to left-hand rules for $\vec{B}$ makes no difference to what happens.

/Martin.....

Martin Gardner (1967) has given an entertaining and instructive popular account of the symmetry of electromagnetic phenomena which makes no direct reference to polar or axial vectors and which would be worth the trouble of the interested teacher to consult. The point is far from trivial, as it is now known that not all phenomena in nature exhibit this kind of symmetry, and some pupils at least might be interested enough to pursue the matter further.

7.15.6 Field of a current-carrying straight wire, flat coil and solenoid

From the formula for the force between two long, parallel, straight current-carrying wires the expression for the magnetic field strength at a given distance from a single straight wire can easily be derived. A suitable form of the derivation is the following:

To find the magnetic field strength at a given point, we must in principle find the magnetic force per unit length per unit current on a current element placed at that point and held at right angles to the field direction there. From symmetry considerations it is clear that the magnitude of $\vec{B}$ will be the same at all points at a given perpendicular distance from a long, straight, current-carrying wire. To find the magnitude of $\vec{B}$ in this case one may therefore simply find the force per unit length per unit current on a second long straight wire held parallel to the first at the correct distance from it, since the second wire will be perpendicular to the magnetic field of the first and $\vec{B}$ will be the same at all points on it. From the formula

$$\text{Force per unit length} = k_m \frac{i_1 i_2}{r}$$

we then have directly that the magnitude of $\vec{B}$ at a distance r from the given wire which carries a current i_1 is given by

$$B = k_m \frac{i_1}{r}$$

/According.....

According to tne formula just derived, the magnetic field strengtn due to a long straight wire varies inversely as the distance from tne wire. This is a result well worth showing experimentally (it constitutes a partial verification of tne formula for the force between a pair of current-carrying conductors) and snould be attempted by pupils if time permits. If a compass needle is placed in a vertical plane containing the wire and tne direction of tne horizontal component of the earth's magnetic field, the tangent of its angle of deflection will be proportional to tne magnitude of $\vec{B}$ due to the wire. Therefore tne experiment consists in the measurement of tne angle of deflection of the compass needle when the latter is placed at different points on a line passing through the wire and drawn in the direction of the horizontal component of tne earth's magnetic field. The tangent of the angle of deflection can be plotted against tne reciprocal of tne distance from the wire and the resulting graph should be a straight line. (A current of 5 amps is suitable.) Note tnat in analyzing tne results in this way, use is made of two facts: (a) tne direction of tne resultant magnetic field is given by tne direction of tne compass needle (tnis is an experimental fact); (b) $\vec{B}$ is indeed a vector. No direct proof has actually been advanced in favour of the latter point; $\vec{B}$ has been "defined" as a quantity having both magnitude and direction, but no independent demonstration was given that $\vec{B}$ conforms to the laws of vector addition..
We do not suggest tnat time need be spent on such a proof, but tne point should be brought to tne attention of pupils, who can be invited to submit suggestions for a method of experimental verification.

The magnetic field of a current-carrying flat coil is of some interest in itself and of some relevance to what follows. Witn tne "fine-beam" tube mounted between Helm-

/holtz.....

noltz coils now becoming standard equipment, it is important for the pupil to have some idea at least of the "appearance" of the field associated with the flat coil. The pupil should see an "iron-filings" demonstration of this; even if he has seen this demonstration done in the junior secondary school, it would be advantageous to show it to him at this point, and to draw attention to its most important features, viz. that the field on the axis is parallel to the axis, that the field very near the wire resembles that of a long wire, and that the overall field pattern is similar to that which might be obtained with a disk-shaped magnet having a north pole on one face of the disk and a south pole on the other.

The effect of superposing the fields of two or more such coils could be discussed and demonstrated; in this way pupils can be given some preparation for the use of Helmholtz coils, and also for the form of the field inside a current-carrying solenoid. There would be no particular use in the course under consideration for the formula for the field at the centre of the flat coil. This is just as well, as the formula could not conveniently be derived without the aid of the Biot-Savart Law. Its establishment by experiment would require the use of several <u>different</u> coils of different radii; this could be set as a project for interested pupils to carry out, but in a course of the scope prescribed by the present South African syllabuses, the result of the experiment would have the status of an isolated fact rather than an important connecting link.

The appearance of the field of a current-carrying flat coil can also be used to show that a pair of parallel coils will attract each other if the current circulates in the same sense in each, and repel each other if the senses of circulation are opposite. Similar predictions can be made for the

/forces.....

forces between current-carrying solenoids. The tendency of current-carrying flat coils and solenoids to align themselves with their axes parallel to an external magnetic field in which they are placed can also be discussed and explained.

(The more advanced pupil might benefit from being given a private introduction to the Biot-Savart Law and its use. The current-carrying flat coil would be a useful circuit for such an introduction. The pupil can be asked to consider the idea that the resultant magnetic field at any point, and at the centre of the coil in particular, is made up of contributions from each "small piece" or "element" of the coil. It can be put to him that all these contributions to the field at the centre will have the same direction, viz. along the axis, that the magnitude of each contribution will be proportional to the current and to the length of the "element", and inversely proportional to some power of the distance between the "element" and the centre - the pupil himself might be prepared at this stage to suggest that the dependence on the distance is an inverse square one. It is easy to deduce from these assumptions that the strength of the field at the centre is inversely proportional to the radius of the coil and directly proportional to the current, and this is a result which could be checked experimentally by the same technique as that used for the field of a long straight wire.)

The formula for B inside a long current-carrying solenoid cannot be derived in this course, and no useful purpose would be served by quotation of this formula. Instead it can be argued from comparison of the appearance of the field pattern inside a solenoid with that of a uniform <u>electric</u> field that the field inside a long solenoid and away from its ends might be uniform also; and pupils can investigate the magnitude of the field strength inside the solenoid experimentally by measuring the force on a current-carrying conductor placed inside it.

/In.....

In the foregoing discussion, we have considered magnetic fields of a particularly simple kind. In representing them by means of field lines, it is the invariable practice to consider only the idealized versions of these fields, e.g. the field of an infinitely long straight wire and of an infinitely long solenoid. No problems arise in these idealized cases; the field lines can be drawn as closed loops. Evidently we have here examples of what is often asserted to be a general rule concerning magnetic field lines: "Lines of magnetic induction [i.e. of $\vec{B}$]... form loops without any beginning or end" (Kip 1962). In view of the cautionary remarks we have made earlier in these pages about the uncritical use of field lines, we ought to consider whether the quoted rule is of completely general application. To guide us, we need to consider the physics lying behind the field-line representation. The physical fact which the quoted rule is intended to represent, is that the "divergence" of the $\vec{B}$-vector is zero everywhere (everywhere, that is, except in the regions of space which may include the as yet undiscovered magnetic monopoles!). Put in other terms, this means that the magnetic flux through any closed surface in the magnetic field is zero. The convention usually adopted for the drawing of magnetic field lines requires that the number of lines per unit area in a given region of the field is proportional to the magnitude of $\vec{B}$ in that region. This makes it possible for magnetic flux to be pictured as the total number of lines passing through a given area, and for the vanishing of the divergence of $\vec{B}$ to be thought of as meaning that the number of field lines entering a closed surface in the field is the same as the number leaving it. This is a basic condition that has to be satisfied by a field line representation of the field; another is, of course, that the

/tangent.....

tangent to the field line at any point gives the direction of $\vec{B}$ at that point.

It is generally supposed that because magnetic field lines do not start or end on magnetic charges, and because they have to satisfy the "zero-divergende" condition, they must of necessity consist of closed loops. At the back of this supposition lies the belief that because the $\vec{B}$-field is continuous, the lines used to represent it must be everywhere continuous. As Slepian (1951) has pointed out, there are no grounds for this belief, and simultaneous adherence to it and to the principles of field-line representation outlined above will, in general, lead to contradictory results. Slepian uses the example of a steady magnetic field produced by a current-carrying loop not completely in one plane to illustrate this point. Lines starting from a given limited region would circle the wire and return to the region from which they began but, because the loop is not plane, could not return to their original starting points. Nor could they now be continued into the initial region, as this would alter the density (i.e. number of lines per unit area) of the lines there. They have thus to be stopped, so that the field-line representation of this field must consist of a collection of broken loops, drawn so as to meet the condition of continuity of line density. As Iona (1963b) remarks, "The representation of a source-free field by lines ending and beginning with equal density on opposite sides of an arbitrarily selected surface is just as good an indication of $\vec{\nabla}.\vec{B} = 0$ as closed lines would be."

The loss in easy picturability occasioned by the acceptance of this view of field-line representation is regrettable but inevitable, if the demands of rigour are to be satisfied. Fortunately this need not complicate the lives of the teacher and his pupils. Iron-filing patterns of magnetic field lines

/of.....

of current-carrying circuits look like closed loops, and are in any case formed in two dimensions and not three. Thus they do not arouse any suspicion of the troubles to which attention has been drawn above. They are clearly different in character from electric field patterns, and a crude but serviceable way of expressing this important difference is to say that whereas electric field lines begin and end on charges, magnetic field lines appear to have neither beginning nor end. We see no reason why the matter should not be discussed in these simple terms at this level. The pupil should be made aware of the difference in character between electric and magnetic fields, and the statement just given, while not rigorously true, does help the pupil to visualize this difference. It will not hamper his understanding of electromagnetism at this stage, unless it leads him to attach too much reality to the field lines. The teacher can guard against this by issuing suitable warnings from time to time.

7.15.7 Force on a moving charge in a magnetic field

To introduce the derivation of the expression for the force on a charged particle moving in a magnetic field, one might argue as follows:

Moving charges set up magnetic fields (cf Rowland's experiment), and conductors in magnetic fields only experience "magnetic forces" when the free charge carriers inside them are set moving. Perhaps it is because each moving charge in a conductor experiences a magnetic force that the conductor as a whole has a magnetic force on it when it is in a magnetic field and carries a current. If this argument is correct, we must expect that a beam of charged particles would be deflected by a magnetic field, and the left-hand rule should enable us to predict which way the beam will be deflected.

/We.....

We think it advisable that these predictions should be checked before the required formula is derived. This will help to create in pupils the desire to know more about the behaviour of charged particle streams in magnetic fields, so that they can see the advantage of having a formula for the force on a moving charge. (Cf Nuffield Guide V: "We show pupils that we and they need the result badly.") The required demonstration can be done with a cathode-ray tube, "deflection tube" or "fine-beam tube". For a demonstration with a cathode-ray tube, one could use the oscilloscope in the school laboratory, if the cathode-ray tube is not surrounded by a metal shield; a bar magnet held near the electron gun of the tube will very readily produce a deflection of the spot on the screen. Gledhill (1969) has warned that if the cathode-ray tube has much nickel in its electrode structure, it can be ruined by attempts to carry out this demonstration. In cases of doubt, the manufacturers should be consulted, but it may be mentioned that "Heathkit" oscilloscopes of the O-12U type have been used for this demonstration without apparently suffering any ill effects. If the laboratory oscilloscope cannot be conveniently adapted for use in this way, an independent cathode-ray tube can of course be used; Iona (1961) has suggested that such a tube should be connected by a long multiconnector cable to its power supply so that the tube can be mounted at any given angle and at some distance from disturbing influences such as the transformer in the power supply. The "deflection tube" and the "fine-beam-tube" are both mounted between coils which, when they carry a current, produce a substantially uniform magnetic field in the region traversed by the electron beam. The "fine-beam" tube has the particular advantage for this type of demonstration that the path of the electron beam is shown in three dimensions, so that the interesting effects obtained when the beam is projected at angles other than 90° to the

/field.....

field direction can readily be shown.

The derivation of $F_{max} = Bqv$ is a standard piece of theory; satisfactory simple versions of it are given by Rogers (1960) and the Nuffield Guide (V). It appears that this result was first derived (in a more fundamental manner than in the references cited) by Heaviside in 1889 (cf Shire 1960, Feather 1968).

There is one aspect of this type of derivation which might worry the teacher: how is the magnetic force on the free electrons in a wire transferred to the wire itself? This question has been answered in detail by English (1967), who points out that the mechanism of transfer operates through the Hall effect. The magnetic force on the free electrons produces a small separation of charge across the wire, bringing into being an electric field (the Hall field) which exerts a force on the free electrons equal and opposite to the magnetic force. This same field must exert a force on the positive ions of the metallic lattice which is equal and parallel to the magnetic force on the free electrons; hence it is the Hall field which ensures that the force felt by the <u>wire</u> is the same as the magnetic force felt by the free electrons. English also discusses the case of an intrinsic semiconductor in a magnetic field, which is a good deal more complicated. In this instance the net force on the conducting spedimen arises from variation of "hole" and electron concentrations across the specimen, giving rise to a pressure difference between the appropriate faces of the specimen, as well as the Hall effect.

In any presentation of the derivation of the expression for the force on a moving charge, it would become clear that one is saying, in effect, that whenever a charged particle is moving in the field of another moving charged particle, each charge must experience a "magnetic" force in addition

/to.....

to the electrostatic force between them. Various interesting consequences of this idea might be used to exercise the minds of brighter pupils.

(i) Consider the force between a pair of electrons at relative rest but moving with constant velocity relative to the observer in a direction normal to the line joining them. According to this observer, the force between the electrons will be less than the electrostatic repulsion between them as given by Coulomb's Law; naively, one can attribute this to the fact that there is an attractive magnetic force between these two "parallel currents". To an observer at rest relative to the electrons, the force between them is the normal repulsive force given by Coulomb's Law. There is nothing contradictory about this; in fact, application of special relativity to the problem shows that the force as measured in the one case must be a factor of $\sqrt{1-\frac{v^2}{c^2}}$ less than the force measured in the other (v is the speed of the electrons relative to the "stationary observer"; c is the speed of light.).

(ii) Consider two similar parallel wires, each carrying the same current. To an observer at rest in the laboratory, the wires are seen to be attracted towards each other. But suppose that the observer moves relative to the laboratory with the same velocity as the average drift velocity of the free electrons in the wires (he has to move very slowly to do this: the average drift velocity will typically be only a small fraction of a centimetre per second.). What will he now observe? According to him there is now no magnetic force between the electrons; but the positive ions of the lattices of the wire are now moving relative to him with the same magnitude of velocity as the electron drift velocity, and they must presumably now experience the same attractive force per unit length of wire as the electrons did in the first instance. Therefore the observer measures the same force as before.

/In.....

(In this case, application of special relativity to the problem is less straightforward than before, because there will be disagreement between the two reference systems as to what constitutes "unit length". If by "unit length" is meant "unit length in the reference system employed," there will in fact be agreement between the forces per unit length measured.)

The above two cases are discussed "naively" by Kolb (1962) and more rigorously by Ganley (1963). For textbook treatments of these and related matters, see Rosser (1964), Feynman et al. (1964), Sears and Zemansky (1964) and Purcell (1965).

7.15.8 Orbits of charged particles in magnetic fields: the charge-to-mass ratio of the electron

After the expression for the force on a moving charge in a magnetic field has been derived, it should be used to derive the equation describing the orbit of a charged particle projected at right angles to a uniform magnetic field. The derivation takes one line, so there is time for giving it an adequate background, which could take the form of a demonstration with a fine-beam tube that the orbit is a circle and a preliminary discussion of why the orbit must have this shape. In a discussion of this point, one must start with the fundamental fact that the magnetic force on a charged particle is always at right angles to both the velocity of the particle and the field. This force can have no component in the instantaneous direction of travel of the particle; therefore it cannot change the speed of the particle but only its direction of motion. This means that the magnitude of the force is constant, since the field is uniform and the magnitude of the velocity is constant. Therefore the particle must execute an orbit in a plane perpendicular to the field (it was originally projected at right angles to it), in such a way that its

/speed.....

speed is constant and it undergoes a constant acceleration at right angles to its direction of travel. The only orbit that fulfils these conditions is a circle.

To find the equation of the orbit, one simply puts the force on the particle (given by Bqv) equal to its mass times its acceleration ($= \frac{mv^2}{R}$).

$$\text{i.e.} \quad Bqv = \frac{mv^2}{R}$$

$$\text{or} \quad R = \frac{mv}{Bq}$$

Pupils now have all the necessary background for a determination of the ratio of charge to mass for the electron. In the Nuffield Guide (V), it is urged that teachers should "advertise the importance" of $\frac{q}{m}$ beforehand since "it may seem to pupils a dull thing for us to work out". The Nuffield Guide suggests "advertising the importance" of this quantity by pointing out that it is "a piece of information about extremely small things" and that $\frac{q}{m}$ for electrons can be compared with $\frac{q}{m}$ measurements for other particles. We would add to the first of these points the suggestion that if q is known for the electron (which it is, from Millikan's experiment), we can use $\frac{q}{m}$ to find the mass of the electron.

By far the most elegant and convincing method of determining $\frac{q}{m}$ for the electron in a school laboratory is that which makes use of the "fine-beam" tube. A suitable accelerating potential difference is applied to the electron gun of the tube, a suitable current is switched on in the magnetic field coils, and a circular path is seen to be traced out by the electron beam. The measurements that are taken are of the strength of the magnetic field (using a force-on-a-conductor balance placed in the field, with the tube removed), the potential difference applied to the electron gun (using a voltmeter) and the diameter of the orbit (see below). Because one cannot place a ruler inside the tube, measurement

/of....

of the diameter of the orbit raises a problem (cf Nuffield Guide V for a full discussion). One way of solving it is to estimate the diameter of the orbit, which would give a result known to within 20%. The Nuffield Guide author does not think this is worth improving on except with a "very fast" group of pupils: for such pupils he suggests forming in the plane of the orbit a virtual image of an illuminated scale, by placing a vertical sheet of plate glass between the scale and the tube so that the plate is exactly midway between the scale and the orbital plane. The reason for the reluctance of the Nuffield author to insist on the attainment of greater accuracy is that this might detract from what he conceives to be the main purpose of the experiment, viz. "to bring pupils into contact with a real measurement of electrons". Perhaps a suitable strategy for the teacher to adopt would be to let the measurement of the orbital diameter be rough in the first place, and then leave it up to the pupils to ask for improved accuracy if they want it. Besides the orbital equation derived above, the "theory" of the experiment needs one other equation, viz. that giving the kinetic energy of the electrons in terms of the accelerating potential difference:

$$qV = \frac{1}{2} mv^2$$

Again, the Nuffield Guide (V) suggests combining this equation with the other, first to yield a value for the electron speed and then to find $\frac{q}{m}$, on the grounds that v is a simple property, whereas $\frac{q}{m}$ is not, v is very easily obtained from the two equations, and the value obtained for v will be astonishing to the uninitiated. Most teachers will probably be glad to accept this recommendation on the grounds of the second reason alone, since this does seem to be the easiest way to do the algebra. We have

$$qV = \frac{1}{2} mv^2$$

$$\text{and} \quad Bq\,v = \frac{mv^2}{R},$$

/so.....

so that $\frac{Bv}{V} = \frac{2}{R}$, by simple division, and

$$v = \frac{2V}{BR} .$$

This value of v can then be used to find $\frac{q}{m}$ from the relation

$$\frac{q}{m} = \frac{2v}{BR}$$

Since the fine-beam tube is expensive, and only one is likely to be available at any one school, this experiment will have to be done as a communal experiment, either by the class as a whole or by groups. Teachers wishing to "individualize" the experiment will have to fall back on the P.S.S.C. version of it, which uses a 6AF6-GT or EM-34 "magic-eye" tube (cf P.S.S.C. Laboratory Guide). (The credit for suggesting that such tubes could be used in this way must evidently go to Mackay (1949).) The magic-eye tube is mounted inside and parallel to the axis of a suitable-sized solenoid. The electrons that are accelerated by the electric field in the space between the centre cap and the cathode travel with more or less constant speed after emerging from the cap over the remainder of their path to the anode. Since these electrons move at right angles to a magnetic field that is nearly uniform, they describe paths that are nearly arcs of circles. The edges of the shadow cast by the electron beam on the fluorescent material of the anode are thus circular. The radius of this shadow is "measured" by adjusting the current in the solenoid until the radius is judged to be the same as that of some circular object of known radius (e.g. pencil, coin or dowel rod). The strength of the magnetic field is measured by a force-on-a-conductor balance placed in the solenoid after the magic-eye tube has been removed, and the accelerating potential difference for the electrons is measured by means of a voltmeter. This experiment does give a cheap and simple method of finding the order of magnitude of $\frac{q}{m}$ for the electron, and could be used in the absence of the fine-

/beam.....

beam tube. If the latter is available, however, we doubt the value of letting pupils carry out the "magic-eye" experiment for themselves, except perhaps as a reminder that the things that come out of the cathodes of radio valves are of the same nature as those that emerge from the electron gun of the cathode-ray tube.

7.16 An historical approach to the electron

It will have been noticed that in the treatment of the electron outlined in this account, we first "hypothesized" it as the entity that might be responsible for carrying the charge in electrical conduction in metals; we then showed that the latter charge carriers conveyed negative charge, later identified in amount with the natural unit of charge revealed in Millikan's experiment; and finally we measured the ratio of the charge on the electron to its mass, enabling the latter to be calculated. All this has paid scant regard to the historical development of the subject, which began with the investigation of the "cathode rays" noticed in experiments on the conduction of electricity in gases at low pressures and culminated in J.J. Thomson's measurements of $\frac{q}{m}$ for the cathode-rays. We submit that if it is accepted that an elementary syllabus in physics should include some account of the electron, the above outline represents a treatment which is as logical and complete as can be managed within the limited space available in such a syllabus. To present an account of the electron along strictly historical lines would have required one not merely to trace in detail the work of Plücker, Johnstone-Stoney, Crookes, Perrin, Thomson and others on the "particle" nature of the cathode rays, but also that of Hertz, Lenard and Goldstein designed to support the view that "cathode rays" were electromagnetic waves, and we doubt whether anyone would now wish to defend such an

/approach.

approach. An excellent account of the history of the discovery of the electron is given by Anderson (1964) and good summaries are to be found in Arons (1965) and Feather (1968).

The new Common Basic Syllabuses do contain a somewhat ambiguous reference to "evidence [for the particle nature of the electron] from the passage of electricity through gases at low pressure." It is presumed that this is intended to refer to some of the historical work on cathode-rays and not, say, to the fine-beam tube. We suggest that in the age of the cathode-ray tube and the fine-beam tube, these syllabuses ought to be modified appropriately. We give a summary below of the relevant features of the phenomena observed in low-pressure discharge tubes; this outline might prove useful if the electron is to be studied along partly historical lines:-

(1) The traditional apparatus for the demonstration of the conduction of electricity through gases consists of a long closed glass tube with electrodes mounted inside it, one near each end. The tube is attached to a pumping system capable of reducing the pressure of the gas inside the tube to any desired value. When a sufficiently high potential difference is applied to the electrodes and the pressure of the gas is made sufficiently low, charge is transported between the electrodes, as indicated by a milliammeter connected in series with the tube. To begin with, the pressure of the enclosed gas may be a few mm of mercury; as the pressure is lowered, the appearance of the gaseous discharge changes strikingly (but the details of these changes are not relevant to the syllabus). When the pressure is less than 0.01 mm of mercury, the tube is almost completely dark but for a bright fluorescence on the wall near the anode. This fluorescence can be moved about by holding a magnet near the cathode. (This was first observed by Plücker in 1858). This suggests

/a.....

a direct association between the fluorescence and the charge coming from the cathode.

(ii) A screen placed in the path of the "rays" emanating from the cathode casts a sharp shadow in the fluorescent region, showing that the rays travel in straight lines.

(iii) If a modified form of discharge tube is used, the effect of magnetic and electric fields on the cathode rays can be clearly shown. The essential features of such a modified tube are (a) a means of making the beam of the cathode rays narrow and well-defined, and (b) the incorporation of a fluorescent screen (e.g. a coating of zinc sulphide) on which the beam can impinge. The first feature can be arranged with the aid of collimating slits. As was convincingly shown by J.J. Thomson, the application of magnetic and electric fields in the vicinity of the beam resulted in the deflection of the bright spot that marked the point of impact of the beam on the screen. The direction of the deflection of the spot established that the charge carried by the cathode rays was negative.

(iv) Points (ii) and (iii) strongly suggest that the cathode rays are in fact small negatively-charged particles. Careful measurements by Thomson of the deflection of the spot produced by electric and magnetic fields of known strength enabled him to make the first satisfactory measurements of the ratio of charge to mass of these particles which have come to be known as electrons.

(v) In one of Crookes's experiments on the particle nature of cathode rays, he showed that if the rays were allowed to strike a paddle wheel mounted on rails inside the discharge tube, the paddle wheel rotated. Today one may still see discharge tubes with paddle wheels mounted inside them, and some of these discharge tubes may be found in school laboratories. Crooks thougnt that the paddle wheel rotated because

/of.....

of the momentum transferred to it, but Thomson pointed out in 1903 that the amount of momentum transferred to the wheel by the electrons was much too small to rotate it at the observed rate, and that it was much more likely that the wheel rotated because its vanes were heated up on one side by the incident electrons (cf Anderson 1964). The rotating paddle wheel is thus a pretty but misleading demonstration.

Some schools may possess the cathode-ray demonstration apparatus described in (i) above. Because of the high voltages that must be employed to operate these "cold-cathode" discharge tubes, the tubes can act as X-ray tubes and thereby expose demonstrator and pupils to dangerous radiation. If the effects described above are to be demonstrated at all, it is both much safer and a great deal more convenient to use hot-filament discharge tubes which require far lower operating voltages. (Hot-filament tubes with internally mounted Maltese crosses are now commercially available: they are suitable for showing that cathode rays travel in straight lines.)

7.17 Electrical circuits - quantitative treatment

7.17.1 Objectives

In the junior secondary school, pupils were introduced to the rudiments of "circuit theory", including the concepts of current, potential difference and resistance. They should now have a deeper insight into the concept of current, and they have studied potential difference in the context of electric fields. They have also been given a general picture of circuit behaviour in which use is made of the concepts of the electric field and of "flow of charge". It remains now to "put the numbers into" this more fundamental description of the electric current: in particular to show how "potential difference" as defined for the electric field links up with the "potential difference" as defined in the junior years,

/and.....

and to introduce the vitally important concept of EMF. These must be regarded as the main objectives of this section; all else is secondary.

7.17.2 Further aspects of the particle model of electrical conduction

First, it would be as well to sketch in some more details of the model we are using to describe electrical conduction. We have seen that according to this model, the electric field applied along a wire causes the free electrons in it to experience a series of alternate accelerations and collisions. The net result is that the electrons drift further and further down the wire as time goes on. We may therefore define an average drift velocity for the free electrons as their average displacement down the wire per unit time. In terms of the kinetic theory of matter, we must imagine the electrons to be executing random motion all the time, whether an electric field is applied or not. The drift motion we have just described, which is carried out under the influence of an applied electric field, is superimposed on this random motion.

The magnitude of this average drift velocity is, generally speaking, extremely small. Suppose we have a circuit consisting of a battery, a switch, a torch bulb and some connecting wire. It has been calculated that the average drift velocity of the free electrons in a circuit like this would only be a tiny fraction of a centimetre per second. Yet when we close the circuit, the bulb lights up almost at once; the "message" that the current is on seems to have reached the bulb almost instantaneously. Some explanation of this ought to be given to pupils, some of whom are bound to be curious about it. The following elementary account ought to serve the purpose adequately:

/The.....

The electric field which sets the free electrons in the wires drifting reaches its steady value all round the circuit almost at once. What happens is this: closing the circuit brings about a change in the electric field pattern in and around the elements composing the circuit, and this change - from whatever it was before to the new steady condition - is propagated around the circuit as a kind of wave, concerning which more will be said later.

More details of the "change" in the electric field may be of interest to teachers. Prior to closing the circuit, there is no electric field inside its conductors, but there is an electric field in the surrounding space set up by the distributions of static charge maintained by the battery on the conductors. When the circuit is closed, free electrons flow through the connection made by the switch, initially under the influence of the electric field that arises from the potential difference between the connecting wires now joined by the switch. This upsets the previously established distributions of charge over the conductors of the circuit; charge starts to flow elsewhere in the circuit as a result, rapidly building up a new distribution of charge on the surface of the conductors in such a way that the electric field within the conductors is parallel to their surfaces. (Rosser (1963a) discusses this situation in some detail; incidentally, he complains that the topic is "never discussed adequately in elementary textbooks.")

Teachers and pupils should both be aware that despite the use in this model of electrical conduction of the concepts of electrons, charge, and potential difference, and the success with which we are able to use it to explain many aspects of electrical conduction, it is still only a model, which may yet prove deficient in certain respects. In fact,

/this.....

this is very much the case; there are subtleties concerning the nature of the electron and the free electron "gas" in a metallic conductor of which our model takes no account, and because of this, some of its predictions (e.g. those concerning the variation of resistance with temperature) are not confirmed by experiment. Nevertheless, despite its inadequacies this model has the very real merit of focussing attention on fundamental electrical concepts when the electrical circuit is studied.

7.17.3 Potential difference in electrical circuits

In the junior years, the potential difference between two points in a circuit was defined as energy transferred per coulomb of charge "passing through" this portion of the circuit. In returning to potential difference for the senior school syllabus, we have defined it for the electric field as the work done per coulomb against the field in taking a small positive test charge from the point of lower to the point of higher potential. The task now is to show that as far as the electrical circuit is concerned, these two definitions are equivalent. (This point is usually dealt with rather cursorily by textbooks and as it could conceivably give some difficulty to the beginner, we treat it in some detail here.) In effect, we must show that when an amount of charge q "passes through" a conductor across which a potential difference V is maintained, the amount of energy given up by the flowing charge within the conductor must be qV.

We must first clarify what is meant by a charge q "passing through" the conductor. What we mean is that in the time considered, an amount of charge q enters one end of the conductor and a similar amount leaves it at the other end. The pupil has now to be persuaded that as far as the energy given up by the charge is concerned, the effect is the same

/as.....

as though the amount of charge q was all concentrated on one particle which "falls through" a potential difference V in the electric field (which we suppose not to be disturbed by the presence of charge q on the particle). To some this may be obvious. For those for whom it is not, we suggest that the matter could be explained by considering in detail the process of energy loss by the charge carriers. Suppose that a potential difference V is maintained between the ends of a uniform metallic conductor in which there are N free electrons, each having charge e. In the time interval t that all of these electrons would take to flow out of the conductor, how much electrical energy is given up within the conductor? Only the free electrons at the high-potential end of the conductor will have travelled the full length of the conductor in this time; the others will "escape" from the conductor before the end of the time interval, and during the remainder of it will give up electrical energy elsewhere in the circuit. But as these electrons move out of the conductor, others move in to take their place and give up electrical energy within the conductor, so that the overall effect as far as the transfer of energy is concerned is as though all N free electrons had simultaneously started at the high-potential end of the conductor. Therefore the total amount of electrical potential energy lost by the flowing charge in the conductor during the time interval t is given by

$$W = N \times (eV) = q\,V,$$

where $q = Ne$ is the total charge that flows into one end of the conductor and out of the other end in time t. It follows directly that the <u>rate</u> at which energy is given up within the conductor is given by

$$P = \frac{q\,V}{t} = \left(\frac{q}{t}\right) V$$

$$= iV$$

/This.....

This result is extremely important - of much more fundamental importance than Ohm's Law, for example - and should be thoroughly understood. Pupils should be led to realize its complete generality - that it applies to the potential difference between any two points in any circuit. If, for example, the conductor across which one wants to know the potential difference is the winding on the armature on a D.C. electric motor, it is still true that the rate of electrical potential energy loss is given by the current times the potential difference. In this case the "lost" energy becomes mechanical energy of the motor as well as internal energy of the conductor itself, but this does not affect the fundamental relationship between rate of electrical energy loss, current and potential difference.

In discussing the significance of the term "potential difference" Warren (1965) finds it necessary to introduce a concept which he calls "current potential difference". He says: "The energy used in moving a charge in a vacuum is in principle recoverable unless it hits a body, but the energy used in moving a charge along a wire is dissipated as heat. It is therefore desirable to define P.D. in current electricity in terms of the dissipation of energy, that is, the P.D. across a conductor carrying a steady current is the rate of heat production divided by the current." With his first sentence we can agree, subject to the limitation that the "wire" spoken of should not form part of, say, a motor or a generator. But we fail to see the force of the "therefore" in his second sentence. In principle, there is no objection to the introduction of a new concept, defined in terms of the irreversible energy loss per unit charge undergone in a conductor; what we have to decide is whether such a concept is really needed. As we have pointed out, what Warren calls "electrostatic" P.D. covers all cases of energy transfer in circuits, including irreversible ones. As Warren himself

/points.....

points out, it is the "electrostatic" P.D. that one measures with a voltmeter. It is difficult to see that many advantages may accrue from defining a new type of "P.D." to deal with irreversible energy transfers only. The only possible one that we can think of is that we are thereby enabled to say that the "electrostatic" terminal potential difference of a cell (say) is equal to its EMF minus (or plus, depending on the direction of the current) its current potential difference. But against this we have to set the possibility that the use of <u>two</u> kinds of potential difference, one measured with a voltmeter and the other with a calorimeter and a stopwatch, may increase rather than lessen the beginner's confusion. To sum up: the concept of "current P.D." is not "wrong", but its use may create more problems than it solves (cf Helm 1966).

Ericson (1968a,b) goes even further than Warren in introducing different kinds of potential difference. He gives the name "thermal p.d." to what Warren has called "current P.D."; he also introduces a quantity which he calls simply "p.d." (to distinguish it from electrostatic potential difference), which he defines as $\int_A^B \vec{P} . \overrightarrow{dl}$, where $\vec{P}$ = "total force per unit charge on the conventional positive charge in the circuit excluding those dissipative forces which produce Joulean heating." (In simpler terms, Ericson's "p.d." is the work done per unit charge by this "total force" when a positive test charge is taken from A to B). He uses the term "voltage" to refer to the reading on a voltmeter, which is not necessarily equal to one or the other of his three types of "p.d." in particular circumstances. His argument for introducing three types of p.d. is one of expediency. "In circuit analysis," he says, "it is inconvenient to consider that there is a potential difference across a metal-metal junction or p-n junction when no current flows." (There will,

/of.....

of course, be an electrostatic potential difference across such junctions.) Therefore he introduces his "p.d.", which is equal to zero across a metal-metal or p-n junction when there is no current. Also he claims that the "thermal p.d." is useful, although he does not say how. However, he does conclude that "up to O.N.C. and A level physics one can satisfactorily use...the concept of electrostatic p.d." and that his other categories of p.d. are best left to "more specialized courses". As is apparent from our discussion above, we are in agreement with Ericson's recommendation regarding p.d. in elementary courses. Situations in which there is a need to deal quantitatively with metal-metal or p-n junctions do not arise in the syllabus under discussion and it is clearly advantageous to use only one concept of p.d. if this is possible. Whether or not it would be advantageous to use other concepts of p.d. in more specialized courses is a question which falls outside our terms of reference.

The equation P = iV provides the basis for the fundamental calibration of a voltmeter in terms of the joule and the coulomb, whereby one measures the energy dissipated as internal energy in a resistor when a known number of coulombs "passes through" it. The number of coulombs is measured with an ammeter (calibrated with a current balance) and a stopwatch. In effect, this procedure will have been dealt with in the junior years, but pupils could profitably be reminded of it here.

7.17.4 EMF: The nature and definition of the concept

There is nothing awesomely difficult about the concept of "EMF". Yet, as has been pointed out at various times (cf Alexander 1939, Kenworthy 1941, Eldridge 1948 and Ericson 1968a,b), there has been (and still is) much confusion about the meaning of the concept and how to define it. Then, too,

/amongst.....

amongst those who plainly understand the concept well enough, there is disagreement about the best way of formulating certain of the underlying ideas in elementary teaching (cf Pugh 1961, Phillips 1962 and P.P. Kane 1962), while the name itself has been a thorn in the flesh of instructors for many years. It will be best to consider these issues before dealing with the teaching of the concept. We confine ourselves initially to circuits containing only single sources of EMF.

Qualitatively, EMF (or "electromotive force" as it was originally called) is associated with "what makes the current go round" in a circuit. Charge cannot be kept circulating round a circuit under the action of macroscopic electrostatic fields alone. Referring to the example discussed in Section 7.13.2, we can say that charge can only continue to flow from conductor X to conductor Y under the action of the electric field if at the same time some agency can keep pushing charge round from Y to X <u>against</u> the electric field. The agency which achieves this and thereby maintains a current in the circuit is called the seat or source of EMF. In the analogy described in Section 7.13.4, the man lifting the balls from the floor to the shelf corresponds to the seat of EMF in an electrical circuit.

It is clear that any agency which pushes charges <u>against</u> an electrostatic field must transfer electrical potential energy to the charges, and from the principle of conservation of energy it follows that this energy must be converted from some form other than the potential energy of a macroscopic electrostatic field. This fact leads to the association of EMF as a <u>quantitative</u> concept with electrical potential energy made available to the circulating charges from some form other than "electrical", and all definitions of the EMF of a localized source which can make some claim to acceptability can be related to the amount of energy converted per unit charge to

/electrical.....

electric form.

It has been pointed out above (Section 7.6.2) that potential difference can be defined constitutively as the line integral of the electric field vector along any path linking the specified pair of points, and that this definition can readily be linked to potential energy transferred per unit charge. One might hope to formulate a corresponding definition of EMF as a line integral of some appropriate "field vector." This "field vector" should evidently be associated with the force which pushes the charges <u>against</u> the electrostatic field within the source of EMF and which cannot therefore be electrostatic. Examples of such nonelectrostatic forces in sources of EMF are the so-called "chemical forces" that push positive charge carriers from the negative to the positive poles of chemical cells, the mechanical forces that transport charges up to the main dome of a Van de Graaff generator, and the nonelectrostatic <u>electrical</u> forces that arise from the induced electric fields set up by changing magnetic fields. By analogy with the procedure adopted for electrostatic fields, one can take as one's measure of the field vector of a given nonelectrostatic force field the force per unit charge exerted by this field on a small positive test charge. (This presupposes that the magnitude of the nonelectrostatic force exerted on a charge is likely to be proportional to the magnitude of the charge; the measured performance of sources of EMF bears this out.) The "line integral" definition of the EMF of a source of EMF can then be formulated as follows: the EMF of the source is equal to the line integral of the nonelectrostatic field vector from the negatively-charged to the positively-charged terminal of the source. (Normally the actual path of integration does not matter.)

/Let.....

Let us consider the meaning of this line integral, to see whether it is consistent with the idea of "energy converted per unit charge" that we are trying to convey. Translated into operational terms, this line integral gives us the work done per unit charge by the nonelectrostatic field on a positively-charged "test" particle as it is pushed from the negative to the positive terminal of the source. If the only force opposing the motion of the particle were the electrostatic force arising from the electrostatic field between the terminals, and if the particle after being started off moved with constant velocity (i.e. the field vectors of the nonelectrostatic and electrostatic fields were equal in magnitude and opposite in direction), then the work done on the particle by the nonelectrostatic field would transfer only electrical potential energy to it. Therefore in these circumstances the line integral of the nonelectrostatic field vector from one terminal to the other would be equal to the electrical potential energy per unit charge made available to a positively charged particle that is driven through the sourde of EMF. Suppose the situation is altered so that the particle is driven in the opposite direction through the source of EMF, but still at constant velocity. (To bring about such a change, the strength of the electrostatic field must momentarily be made infinitesimally greater than the strength of the nonelectrostatic field). Now the transfer of energy takes place in the reverse direction: energy is converted from electrical potential form to whatever non-electrical form is appropriate to the particular source of EMF. However, the magnitude of the energy converted per unit charge would be the same as before, provided that when the source was operated in the original situation, nothing happened that was likely to change its nonelectrostatic field.

/We.....

We see that if there are negligible dissipative forces in a source of EMF, and if the charge may be driven in either direction through the source without sensible effect on its nonelectrostatic field, then the magnitude of the line integral of the vector of this field from one terminal of the source to the other is equal to the net energy converted per unit charge from nonelectrostatic to electrostatic form or vice versa. It is plain that in the idealized conditions prescribed above these energy conversions take place reversibly in the strict thermodynamic sense of the term; that is, they occur in processes in which the system passes through equilibrium states and the direction of which can be altered by making an infinitesimal change in the environment of the system (cf Zemansky 1957). Therefore the line integral of the nonelectrostatic field may be set equal to "the reversible work per unit charge" or "the amount of energy transferred reversibly per unit charge", and these phrases are often used as alternative, operational definitions of EMF. We have in fact shown that the constitutive, line-integral definition of EMF takes on meaning as "energy converted per unit charge" provided that the conversion takes place reversibly.

If there are appreciable dissipative forces operative within the source (i.e. if the source has appreciable "internal resistance"), then the energy conversions are no longer completely reversible; whichever direction charge passes through the source, some energy will be dissipated as "heat". In other words, "Joule heating" - the common form of heat generation within a conductor - takes place within the source, and the line integral of the nonelectrostatic field vector through the source cannot now be equal to the net amount of energy converted per unit charge from nonelectrostatic to

/electrostatic.....

electrostatic form (or vice versa). (It is worth noting here that reversible conversion of "heat" energy to electrical energy is possible in the so-called "thermoelectric devices" which have nowadays assumed considerable importance.).

Our discussion so far has emphasized the role of a source of EMF as a means of making electrical potential energy available in a circuit, and has shown that a constitutive definition of the concept of EMF as the line integral of the nonelectrostatic field vector of the source from one terminal to the other is directly related to the "energy-converting ability" of the source. The definition immediately makes clear that there is a conceptual difference between EMF and potential difference; the latter is defined as the line integral of the electrostatic field vector, whereas the former is the line integral of the characteristic vector of a nonelectrostatic field.

Our discussion has confined itself so far to situations in which the source of EMF could be a particular circuit element, such as a cell, battery or generator. In certain cases, it may turn out that the entire circuit is effectively the source of EMF; this would be so, for example, if the circuit was a closed loop of wire held stationary in a magnetic field that was changing with time. Here the "line-integral" definition of EMF must refer to the entire circuit; the EMF is the line integral of the nonelectrostatic field vector taken around the entire circuit. It is not meaningful to talk about "potential difference" in reference to such a circuit; a voltmeter connected across a portion of it would give a reading which would depend on the placing of the voltmeter and its leads relative to the circuit and not just on the positions of the points of connection (cf Phillips 1963 for a full discussion.). Strictly speaking, there is no electrostatic field in this circuit, and we can no longer talk

/about.....

about energy converted in it from nonelectrical to electrical potential form. In the circuit under consideration, the energy derived from the changing magnetic field is transferred directly to kinetic energy of the charge carriers of the wire, whence it is converted to internal energy of the wire. (If only a portion of the circuit lay in a region of changing magnetic flux, we would have a genuine electrostatic field established in the circuit and our original descriptions of energy conversion would apply.)

Some authorities take the view that even in cases where the sources of EMF are localized, EMF is properly definable for an entire circuit or closed loop only (cf Panofsky and Phillips 1962, Phillips 1962, Warren 1965). Phillips (1962) goes so far as to say that "Only the loop definition of EMF distinguishes it from an electrostatic difference of potential, and thus makes it a useful additional concept." She does not give a detailed defence of this statement, but the underlying argument is presumably something like the following:-

The fundamental point of difference between EMF and potential difference lies in the difference between a non-electrostatic and an electrostatic field. The line integral of the electrostatic field vector around any closed path must be zero; that of the appropriate nonelectrostatic field vector round a closed path passing once through a source of EMF cannot be zero. In fact, the latter line integral will be equal to the EMF as previously defined if the closed path passes through no other sources of EMF. It follows from these facts that the line integral of the net force per unit charge (excluding dissipative forces from consideration) around the above type of closed path must also be equal to the EMF. The latter statement could indeed be treated as a perfectly satisfactory definition of EMF.

/It.....

It seems to us that, unexceptionable though the above reasoning is, it does not compel one to adopt closed-loop line-integral definitions (or their "work-per-unit-charge" translations) for EMF in place of line-integral-through-source ones when one is dealing with localized sources of EMF. In the last resort, the approach one prefers to adopt here is a matter of taste. It is, of course, essential to make the point in teaching EMF that the net reversible work done on a charge transported right round a circuit is not zero, while the work done on the charge by the electrostatic field alone must be zero.

One should stress that in the "loop" definition, the portion of the "closed path" lying outside the source of EMF does not have to coincide with any actual conducting path; any imaginary path will do for the purpose of formulating the definition of the EMF of the given source. Thus if one prefers the "loop" definition of EMF, this does not mean that one regards the EMF of a cell or a generator as a property of the circuit of which the source is part, rather than of the source itself. Only if the circuit as a whole is the source of EMF, which is sometimes the case, is one justified in regarding EMF in this way.

Warren (1965) has produced a misleading version of the "loop" definition of EMF, viz. "the line integral of the electric field around the loop" (or the work done per unit charge by the electric field on a charge traversing the loop). It all depends on what one means by "electric field". If by this is meant the macroscopic electric field (i.e. that which appears in Maxwell's equations), then the definition is only valid for nonelectrostatic electric fields, such as the electric fields induced by changing magnetic flux. If the definition is intended to include microscopic electric fields,

/then.....

then as Phillips (1962) has pointed out, it can be claimed to be generally valid, because microscopic electric fields are in fact responsible for the other "nonelectrostatic forces" discussed above. It is clear that Warren's definition, even in the down-to-earth version of it given in parenthesis, is not suitable for an introductory treatment of EMF.

The application of the various definitions of EMF given above to such devices as cells and generators is simple and clear. Can it also be applied to what Shire (1960) calls "passive devices", viz. inductors, capacitors and resistors, in which energy conversions are known to take place? If a current increases in an inductance, some of the electrical potential energy associated with the circulating charge is converted to increased energy of the magnetic field associated with the inductance, and if the current decreases, the conversion takes place in the reverse direction. In this case we can legitimately extend the concept of EMF to a "passive" circuit element. If we have charge flowing out of a capacitor, electrical potential energy associated with the electric field between the plates of the capacitor becomes electrical potential energy associated with the circulating charge; if charge is flowing into the capacitor, the reverse happens. Here we certainly have reversible energy transfer, but no nonelectrostatic fields are involved. Certain authors have been happy to treat capacitors as sources of EMF (e.g. Warren 1965), but this practice is in definite conflict with the definitions of EMF adopted above. There is no nonelectrostatic field between the plates of a capacitor, so that there can be no line integral associated with this field. If we were to find the reversible work per unit charge on a charge transported right round a closed loop passing through a capacitor, the net contribution from the electrostatic field

/associated.....

associated with the charges on the capacitor is indubitably zero. It is because there is no literal flow of charge between the plates of the capacitor that the device can give the appearance of functioning like a source of EMF.

If one drops the qualification that the energy transformation characterizing a source of EMF must be reversible, then it is possible to define an EMF for a pure resistance; Shire (1960) has done just this. (In this case the "non-electrostatic field" is the field that one would have to associate with the "frictional forces" that act on the charge carriers in the conductor). But this way confusion lies, at least in elementary teaching. If the concept of EMF is to be extended to passive circuit elements, then it seems to us to be completely safe to do this only for inductances, since only in them do we have genuinely nonelectrostatic fields (induced electric fields, in fact) set up.

The EMF of a source is occasionally defined thus: "The potential difference between the terminals of a source which is not supplying current is called the electromotive force e.m.f. [sic] of the source" (Prinsloo et al 1968). This definition is open to objection on at least three counts. First, it blurs the distinction between the concepts of potential difference and EMF. Second (related to the first), it confuses a numerical equality with a logical identity (cf Section 4.7.4). Third, the numerical equality is not valid in all circumstances. Shire (1960) has pointed out that it does not hold in general for the EMF of a passive device, and may break down altogether for active devices. The EMF of a battery may depend on the current in it on account of polarization effects, and a dynamo whose armature and field coils are connected in series will have zero EMF when the current in it is zero. In view of these objections, we do not think that a good case can be made out for defining the

/EMF.....

EMF of a source in terms of the potential difference between its terminals on open circuit.

Pugh (1961) has proposed that the EMF of a device should be defined as the work done per unit charge by the electric field for a charge transported through the device from the negative to the positive terminal, "provided that there is no current through the device converting energy to heat". This will be recognized as a thinly-disguised form of the "potential difference on open circuit" form of definition, and as such it is unacceptable.

"Electromotive force", the traditional name of the concept under discussion, is a potential source of confusion to both pupils and teachers. The author once heard an experienced, well-qualified physics teacher trying to formulate a definition of EMF beginning "The electromotive force is that force which..."; the teacher was persuaded to proceed differently, but he might unwittingly have confused his pupils about EMF prior to this incident. It is not surprising that other names for the concept are suggested from time to time. Duff (1938) has proposed the name "electromotance", which has been used by some authors of textbooks (e.g. by Corson and Lorrain 1962); Major (1960) has suggested "electrokinetic potential" and P.P. Kane (1962) "electromotive potential". However the SUN commission is content to list "electromotive force" as the official name of the concept. A compromise favoured by many authors and adopted in this thesis is the use of the abbreviation "EMF" or "emf" as the name of the concept. Obviously this is not entirely satisfactory, but until there is general agreement concerning what should replace the traditional name, it provides what is probably the best stopgap solution.

/7.17.5.....

7.17.5 EMF in the elementary syllabus

From the foregoing discussion, it has emerged that EMF can be defined as reversible work per unit charge round a closed loop including the source, or as the energy converted reversibly per unit charge within the source (subject to the limitation that these are in fact operational definitions, introduced in order to make the underlying "line integral" definitions intelligible in an elementary course.) Potential difference is fundamentally defined in terms of the electric field, but in electrical circuits we can relate it to energy conversion. However in this case we are not confined to reversible energy conversion; the potential difference across a pure resistance carrying a steady current, for example, is equal to the rate at which energy is converted from electrical form to "internal energy" per unit current in the resistor, and the energy conversion referred to in this description is entirely irreversible.

In teaching EMF, the basis of its correct definition must be made clear, and the concept must be clearly distinguished from potential difference. The essential ideas can be brought out by consideration of energy relations in a circuit consisting of a battery and a pure resistor. Applying the Principle of Conservation of Energy to this circuit, we have that the electrical energy made available by the battery from chemical form each second must be equal to the energy dissipated each second as internal energy in the resistor and possibly in the battery itself. The latter point raises the possibility that the battery has internal resistance; for the time being we treat this as a prediction whose consequences are to be studied. We need to write down an equation expressing the fact that energy is conserved in this circuit. For this equation, it will be convenient to express the rate of evolution of internal energy in resistor and battery in

/terms.....

terms of the current in them. That in the resistor is given by iV, where i is the current and V the potential difference across the resistor. From the definition of resistance $V = Ri$, where R is the resistance of the resistor, hence the expression for the power dissipation in the resistor is Ri^2. (Note that the validity of this expression is independent of whether or not the resistor is "ohmic"). If we call the internal resistance of the battery 'r", then the power dissipation in it can presumably be set equal to ri^2, so that the total rate of dissipation of electrical energy in the circuit is given by

$$P = ri^2 + Ri^2.$$

Now this must be equal to the rate at which the cell converts energy from chemical to electrical. We must expect that this rate of energy conversion will depend in some way on the current in the circuit. The simplest assumption we can make about this is that it is proportional to the current. We proceed to make this assumption and work out its consequences. So we now say

$$\text{Rate at which energy is converted in battery} \propto i$$

$$\text{or Rate of energy conversion} = \mathcal{E}i,$$

when the proportionality constant $\mathcal{E}$ is characteristic of the battery. (In principle, one could check this assumption directly, for example by carrying out an experiment in which several different conductors are connected in turn in series with the same battery and ammeter for the same length of time, and the amounts of internal energy developed in each conductor are compared. Provided that the resistances of the conductors were large enough, these amounts of internal energy would be found to be proportional to the currents. See prediction (iii) below.)

/The.....

The proposed "energy equation" for the circuit now reads

$$\mathcal{E}i = ri^2 + Ri^2$$
$$= (r + R)i^2$$
$$\text{or} \qquad \mathcal{E} = (r + R)i$$

This equation, then, is derived from the assumptions that the battery has internal resistance and that it supplies electrical energy at a rate proportional to the current in the circuit. From this equation certain predictions can be made.

(i) If the internal resistance r of the battery is very much smaller than the resistance R in the external circuit, the potential difference across R will be very nearly equal to the quantity $\mathcal{E}$. If r is comparable in magnitude with R, then the potential difference across R will be appreciably less than $\mathcal{E}$. Therefore the potential difference across R will be bigger the bigger we make R, with a maximum value (equal to $\mathcal{E}$) when the battery is on "open circuit", i.e. when the battery delivers no current.

This prediction is readily verified with a voltmeter.

(ii) If one were to connect a number of high resistances of different values across the battery terminals, the potential difference across the battery would be much the same in each case and very nearly equal to $\mathcal{E}$.

This prediction could be verified if a suitable voltmeter were used to measure the potential difference.

(iii) If the rate at which internal energy were evolved in the resistances mentioned in (ii) were to be measured, it should be found that the rate of evolution of internal energy was proportional to the current.

It would be awkward in practice to carry out the experimental confirmation of this prediction, but pupils can be assured that careful performance of the experiment would yield

/this.....

this confirmation. Those sufficiently interested could be encouraged to attempt the experiment. It is not usually prescribed in elementary courses, nor does there seem to be any "official" account of such an experiment; but it would be well worth doing.

It can now be concluded that a battery does have internal resistance, and that it does convert chemical energy to electrical energy at a rate proportional to the current it delivers. The quantity $\mathcal{E}$ can now be called the EMF of the battery, defined as the energy converted from chemical to electrical form per unit charge passing through the battery or as the rate at which this conversion takes place per unit current. The concept of EMF can now be generalized to include other devices and complete circuits, and the reversibility of energy conversion in true sources of EMF can be stressed. The difference between the concepts of potential difference and EMF can then be brought out. To a first approximation one can state the difference as follows: the potential difference between two points in a circuit has to do with the net energy converted from electrical to some other form between the points, whereas EMF is a quantity associated with <u>sources</u> of electrical energy in circuits. By way of amplification of the remark about potential difference, one could say that it may turn out that between a given pair of points in the circuit there is a net <u>gain</u> of electrical potential energy, in which case one is concerned with a potential rise and not a potential drop between those points. Be it energy gain or loss, however, it is always computed by taking into account <u>all</u> the conversion processes that take place, including the irreversible ones. Also, for a complete circuit there is no net gain or loss of electrical potential energy - when we go right round the circuit we come back to

/our.....

our starting point where the potential is the same as we started out with. Regarding EMF, we can point out that the "sources" of electrical energy might be so connected as to function as "sinks" instead, in which electrical energy is being reversibly converted to some other form. For the entire circuit, we should in general expect to find that there is a nonzero net amount of electrical energy converted reversibly per unit charge; this is equal to the net EMF for the entire circuit.

It would be instructive to say something to pupils about the conditions prevailing in a source of EMF on "open circuit". One could point out that in a cell, for example, that is not connected to any circuit elements, the chemical action of the cell would ensure that positive charge would be transferred from the negative to the positive terminal inside the cell until the potential difference between the terminals numerically equalled the EMF of the cell. The intensity of the electrostatic field set up by the charges on the terminals of the cell would then be equal and opposite to the intensity of the effective nonelectrostatic field set up by the "chemical forces" within the cell.

7.17.6 Ohm's Law and resistance

The topic of "resistance" has already been treated in the junior years (Sections 3.15.6 and 3.15.13). A few additional remarks on the logic of tests of Ohm's Law are fittingly included here. The simplest way of testing a given specimen for ohmic behaviour is to use an ammeter to measure the current in it and a voltmeter to measure the potential difference across it, and to take corresponding readings over as wide a range of currents as possible. The principle of operation of the ammeter and voltmeter used in the test will be based in part on the validity of Ohm's

/Law.....

Law for the conducting materials used in the meters. If the specimen does not consist of one or other of these materials, then there can be no logical objection to the use of these meters to test whether its behaviour is ohmic or not. The result of the test can be formulated in the following terms: "Since the behaviour of the materials used in the meters is known to be ohmic, the constancy/non-constancy of the ratio of p.d. to current for the specimen under test shows that it must/cannot be ohmic." But suppose the meters have been calibrated by procedures which do not depend on Ohm's Law; e.g. that the ammeter has been calibrated with a current balance, and the voltmeter with a current balance and the heating effect of a current. (As Warren (1961, 1965) has pointed out, the calibration of a voltmeter with a potentiometer is also independent of Ohm's Law, as the current in the potentiometer remains constant throughout. On the other hand, calibration of an ammeter with a standard resistance and a potentiometer does depend on Ohm's Law: the standard resistance has to be ohmic and kept at constant temperature if the calibration is to be conveniently carried out.) Then the ohmic character of the materials in the meters does not have to be invoked in order to interpret the measurements carried out with them, and we may use an ammeter-voltmeter method to check the resistive behaviour of any material we please.

To observe the behaviour of a definitely nonohmic conductor, pupils should use the ammeter-voltmeter method to test a semiconductor diode for both possible directions of the current in the diode. Thermionic diodes could also be studied in this way if time permitted.

The well-known expressions for the effective resistance of a number of resistors connected in series and in parallel are so easily derived that pupils should see them derived

/or.....

(or derive them for themselves). These derivations are not among the major intellectual achievements of the course and should not be treated as such. Warren (1965) has pointed out that the validity of the formula for the effective resistance of a number of resistors connected in series does not depend on the assumption that the resistors must be ohmic; the formula will work just as well if the resistors are non-ohmic, although the values used for the resistances must be the correct ones for the particular current in them. Warren also says: "By contrast the reciprocal rule for resistances in parallel does depend upon Ohm's Law since the current generally divides unequally between the branches". But surely the truth of the matter is rather that although the "reciprocal rule" is still valid for nonohmic conductors (as an examination of the derivation will immediately show), no-one would dream of using it in such cases, since in order to know the resistances of the individual branches, one would have to know how the current divided between the branches - and to find this out the resistances would have to be known! In practice, one would calculate the effective resistance of a number of parallel non-ohmic resistors, with the aid of graphs showing the variation of current with potential difference for these conductors.

The Common Basic Syllabuses include mention of measurement of resistance by the Wheatstone Bridge. On this topic the Nuffield Guide (IV) bluntly says: "...modern measurements are made with multimeters rather than bridges. The Wheatstone bridge fell out of use generations ago, except in teaching laboratories." This is too sweeping. Impedance bridges, based on the principle of the Wheatstone bridge, are still very much in use in research and industrial laboratories, as are platinum resistance thermometers, the resistances of

/which.....

which are measured with Wheatstone bridges. To say the least, it would do no harm for pupils to meet the principle of the Wheatstone bridge in a school science course; it is an intrinsically instructive piece of circuit theory which pupils could study with profit, and is also valuable as an illustration of a null method of measurement. The same can be said of the principle of the potentiometer which undeniably is still widely applied. Pupils should be familiar with it, as is required by the Common Basic Syllabuses.

7.18 The proportionality constant in Coulomb's Law

The bridge between static and current electricity will not have been completely erected until pupils know how, in principle, to measure the charge on a particle in "official" coulombs and thence how to show that the proportionality constant in Coulomb's Law is indeed very nearly 9×10^9 N-m^2/C^2. Once the pupil understands that the "potential difference" measured by laboratory voltmeters is indeed the potential difference defined in terms of the electric field, the completion of the bridge presents no problems. To measure the charge on a particle one could proceed as follows (cf Rogers 1960, P.S.S.C. film "Coulomb's Law Constant"):-

Attach the charged "particle" (a charged conducting ball) to one end of a simple balance equipped with an insulating beam. Place the ball between two large parallel plates and apply a strong electric field to the plates with a high-voltage power supply. Use the balance to measure the force on the charged ball, and a voltmeter to measure the potential difference between the plates. Also measure the distance between the plates. If the ball is well inside the region of the parallel plates, the electric field in its vicinity will be uniform and have an intensity given to a good approximation by the potential difference between the plates divided by the distance

/between.....

between them. Therefore the force per unit charge on the ball is known from measurements with a voltmeter and a metre stick, and the actual electric force on the ball is measured directly with the balance. Combination of these measurements yields the actual charge on the ball. To find the constant in Coulomb's Law, the ball and balance can be removed from the parallel plates, the ball can be made to share its charge with a similar ball, and the balance used to determine the force between the balls when they are separated by a known distance. (Of course the two stages of this experiment can be performed in inverse order.)

It should be mentioned to pupils that the proportionality constant in Coulomb's Law is determined in practice by more convenient but also more sophisticated methods, understanding of the principles of which requires knowledge of electrical theory beyond that outlined in their course.

7.19 Definitions of current, resistance and resistivity

The definition of current as "rate of flow of charge past a given cross-section" is an unexceptionable constitutive definition of this concept in terms of charge and time. It is a logical identity: this and nothing else is what we mean by "current"; according to this definition, if there were no flow of charge, there would be no current, which is correct. (Compare the definition of (average) speed as $\frac{\text{distance travelled}}{\text{time taken}}$, which is a similarly acceptable constitutive definition of average speed.)

For the junior years, we have defined resistance as the quotient $\frac{\text{potential difference}}{\text{current}}$ (cf Section 3.15.13). It will be recognized that this definition is essentially operational in character. The existence of "resistance" does not, in general, depend on the presence of a potential difference across the conductor or a current in it; on the other hand, if the conductor is nonohmic the value of its resistance does

/depend.....

depend on the value of the current in it. This makes the analysis of the concept of resistance a more difficult task than might at first appear. For homogeneous ohmic conductors, we could argue that resistance was a truly intrinsic property, determined only by the material of the conductor and its temperature and by the dimensions of the conductor. Even this is an oversimplification, since it ignores the dependence of resistance on external physical conditions and on other factors, such as ~~the~~ previous history. At any rate, if we accept this view of the essentially intrinsic character of resistance for such conductors, we could argue that for them resistance must be given a constitutive definition in terms of the resistivity of the specimen and its dimensions. Indeed, this is the manner in which resistance is approached in some modern treatments (e.g. Sears and Zemansky 1964). But even here we have to be careful; the resistance of a conductor will depend on the flow lines of the charge through the conductor, which will depend on how we arrange the terminals of the conductor. (The resistance of a hollow cylindrical conductor, for example, will depend on whether the charge flows from one end to the other, or radially from the inner to the outer surface (or vice versa.).)

If we attempt to define resistance in terms of resistivity, we are naturally led to consider the definition of resistivity itself. Clearly we cannot now define resistivity in terms of resistance, so that the usual possibilities here, ranging from the ridiculous "resistance of unit cube" (cf Warren's (1965) exposure of the fatuity of this pseudo-definition) to "the proportionality constant in the equation $R = \frac{\rho L}{A}$", are ruled out. Sears and Zemansky (1964) give "electric intensity per unit current density". This does seem to have an operational character: it is akin to defining mass as force divided by acceleration, or distance as speed multiplied

/by.....

by time. As such, it is suspect, but we shall not be able to improve on it in simple terms. For bulk nonohmic materials such as thyrite, this definition of resistivity would be easier to justify, since the current density is in fact a function of the electric intensity.

To gases and nonlinear devices the concept of resistivity is not really applicable. In such cases the statement that "resistance is potential difference divided by current" can be regarded as a true constitutive definition of resistance, for we cannot treat "resistance" as an entirely intrinsic property for conductors of this description. In sum, this definition is the only one which covers all these cases and, while we might have doubts about its status as a definition in respect of ohmic conductors, for example, we can offer no convenient alternative for use in an elementary course.

7.20 Electromagnetic induction

7.20.1 Faraday's Law: "motional" and "variational" EMF's

"Faraday deduced experimentally the following rule:-

Whenever the total net magnetic flux through a closed circuit varies there is induced in the circuit a voltage [i.e. an EMF] whose magnitude is proportional to the rate of diminution of the flux through the circuit.

"This law is universally true, whether either or both the magnetic system and the circuit are moving, and even if the circuit changes its configuration as well as position - so long as the circuit remains closed and unbroken. The flux density itself at any point may change either because of a variation in the magnetomotive force which caused it, or a physical displacement of the magnetic system, or any other cause. The circuit need not be metallic or conducting. In fact it need be nothing more than an imaginary closed line in space, and the circuit may be as large or small as we

/please."

please." Thus does Bewley (1964) neatly sum up the content and domain of applicability of Faraday's Law of Electromagnetic Induction. Put like this, the facts of the matter seem plain enough. Yet even today, much of the teaching and textbook presentation of electromagnetic induction is characterized by serious confusion concerning these facts - confusion which, as we shall see, is traceable to Faraday's own writings on the subject. But while there was ample excuse for the discoverer of these phenomena to have ideas about their interpretation which are now considered invalid, no excuse can be advanced for those who, wilfully or otherwise, ignore the work that has been done by Maxwell and others in clarifying the subject. Let us consider the predictions of Maxwell's electromagnetic theory, which was developed in part from the experimental researches of Faraday, as far as they apply to electromagnetic induction, for they will enable us to isolate the major sources of confusion.

Consider a closed loop of flexible wire which is placed in a magnetic field. (For simplicity, we shall suppose the sources of the latter to be at rest relative to the laboratory.) Suppose that the strength of the magnetic field is varying with time, and that the loop is simultaneously being made to change its shape and size. According to the quotation above, Faraday's Law can be applied to this circuit. Straightforward but sophisticated mathematical analysis of the application of the Law in this case shows that the induced EMF in the circuit is made up of two contributions: a "variational" or "transformer" component and a "motional" or "flux-cutting" component (cf Bewley 1964, Birss 1961, Corson and Lorrain 1962). The "variational" component is calculated by treating the circuit as instantaneously stationary and finding the rate of change of magnetic flux through this stationary circuit. The "flux-cutting" component is calculated

/by.....

by treating the magnetic field as instantaneously constant and finding "the sum for all the elements of the circuit of the product (length of the moving element of circuit) x (component of velocity perpendicular to the length of the element) x (component of $\vec{B}$ normal to the plane of motion of the element)." (The wording of the statement inside quotation marks is essentially Bewley's.)

The variational component of the EMF arises from the induced electric field set up by the variation of the magnetic field; the motional component arises from the motion of the elements of the circuit in the magnetic field. The latter is particularly easy to understand: the conducting material composing the circuit contains free charges, and as the circuit elements move in the field, so do the free charge carriers, which must then experience "magnetic" forces.

In a circuit held stationary in a changing magnetic field, the only contribution to the EMF is the variational component. This is the situation in a transformer, whose secondary coil encircles a changing magnetic field set up by the changing current in the primary coil; hence the alternative name "transformer component". In pictorial language, we may say that the "transformer" EMF arises from the change in the number of field lines (or "flux lines") "linking" or "threading" the circuit; i.e. the EMF is brought about by change in flux linkage. In a circuit part or all of which is moving in a constant magnetic field, the only contribution to the EMF is the motional component. This is the situation in a dynamo. In pictorial language, we may say that the motional EMF arises from the elements of the circuit "cutting across" the field or flux lines; hence the expression "flux cutting."

We must emphasize that the two types of contribution to the induced EMF given by Faraday's Law appear to owe their origin to two distinct physical causes - the transformer EMF

/to.....

to an induced electric field and the motional EMF to the motion of charged particles in a magnetic field. Feynman et al (1964) have commented as follows on this state of affairs: "We know of no other place in physics where such a simple and accurate general principle requires for its real understanding an analysis in terms of two different phenomena. Usually such a beautiful generalization is found to stem from a single deep underlying principle. Nevertheless, in this case there does not appear to be any such profound implication. We have to understand the 'rule' as the combined effects of two quite separate phenomena."

It is not at all surprising that Faraday should have imagined that he had discovered such a "single deep underlying principle", viz. that all phenomena of electromagnetic induction arise from the relative motion between conductors and magnetic field lines and more especially from the cutting of lines by conductors. Consider the following quotations from Faraday's writings (reproduced in Tricker 1966): "...the inducing wire and that under induction were arranged at a fixed distance from each other, and then an electric current sent through the former. In such cases the magnetic curves themselves must be considered as moving (if I may use the expression) across the wire under induction, from the moment at which they begin to be developed until the magnetic force of the current is at its utmost; expanding as it were from the wire outwards, and consequently being in the same relation to the fixed wire under induction as if it had moved in the opposite direction across them, or towards the wire carrying the current...On breaking the battery contact the magnetic curves (which are mere expressions for arranged magnetic forces) may be conceived as contracted upon and returning towards the falling electrical current, and therefore move in the opposite direction across the wire, and cause an opposite induced current to the first."

/Following.....

Following Faraday, many teachers and writers of elementary textbooks have been content to explain all aspects of electromagnetic induction in terms of "flux-cutting"; we have already noted attempts to explain the principle of the transformer in this way (cf. Sections 2.2, 3.16.4). As late as 1965, we find an attempt being made in a first-year university textbook (Hazen and Pidd 1965) to account for the induction of EMF's by time-varying fields, in terms of a supposed "magnetic" force on electrons resulting from the relative motion between these electrons and magnetic field lines moving inwards or outwards from current-carrying conductors. For example, these authors explain the action of a betatron with the aid of a diagram showing magnetic field lines moving radially inwards with velocity V_B past an electron; they say that this is equivalent to the electron moving radially outward, and by this means explain the fact that the electron experiences a tangential force. (For the purpose of actually deriving the "betatron equation", they drop this picture and just use Faraday's Law.) Now according to what was said above, "transformer action" (the induction of EMF's by time-varying magnetic fields) is to be attributed to induced electric fields. Can it be that we have here two equivalent ways of looking at the phenomenon—that the one is as good as the other? It is easy to show that they are not. First of all, the idea that a charged particle must experience a force because the field lines of a time-varying magnetic field move across it, is based on the assumption that these lines can be thought of as moving. To Faraday this might have seemed natural enough - his writings show that he was a firm believer in the physical reality of his lines of force - but nowadays we should be much more hesitant in making such an assumption. What it amounts to

/is.....

is that lines of force are supposed to have individuality - that as new lines are created round a conductor in which the current increases, the "old" lines retain their identity and are pushed outwards. Now this is a physical picture which must surely be justified by experiment if we are to use it with confidence. Perhaps one can argue that the facts of electromagnetic induction do provide this justification. We shall leave aside the objection that in that case, one would wish to see a detailed quantitative theory which successfully predicts the value of EMF's induced by time-varying fields, in terms of the above physical picture. Instead, we shall raise a more fundamental objection. We have already pointed out that in a well-designed transformer, virtually all the magnetic flux is confined within the primary coil, so that the secondary coil is not even in a magnetic field. There can be no question here of outwards-moving or inwards-collapsing lines of force sweeping across charged particles and so giving rise to magnetic forces on them, so that in this instance the flux-cutting picture breaks down completely. Now this example was presumably not known to Faraday; but it has been known to theoretical physicists and electrical engineers for decades, and one can only suppose that those elementary teachers who did know of it and persisted in teaching the "flux-cutting" explanation of transformer action did so in the mistaken belief that they were thereby making life simpler for their pupils.

We conclude that the induction of EMF's by time-varying fields cannot be attributed to "flux-cutting", and it is therefore downright misleading to teach the subject as though the phenomenon can be so explained. We shall see below (Section 7.20.3) that in certain cases we can have induced electric fields set up when there is relative motion between the magnetic field sources and the chosen reference system, but that it would be incorrect in such cases to ascribe the physical

/course.....

cause of the induction to "flux-cutting". Page and Adams (1935) long ago drew attention to the danger of exclusive emphasis on the "flux-cutting" view of electromagnetic induction: "...the student learns to rely upon [this] point of view... and is greatly disconcerted when this leads to wrong predictions." Obviously if the subject is to be correctly taught, even at elementary level, both the "induced electric field" and "motion in a magnetic field" aspects of electromagnetic induction must be discussed and applied where appropriate. There are certain examples of electromagnetic induction where it is not immediately obvious what is happening (e.g. Faraday's disc generator, Hering's experiment and Blondel's experiment), but these paradoxes can always be resolved by careful consideration of the fundamental processes that might be applicable - "induction of an electric field" or "motion in a magnetic field". (A summary is given by Scroggie (1955, 1963); a number of paradoxes are analysed in detail by Bewley (1964). Certain aspects of these paradoxes are usefully discussed by Scanlon et al (1969) from a viewpoint which lays emphasis on the mathematical and physical basis of the general rules given by Bewley for the resolution of the paradoxes.)

Pugh (1964) has gone so far as to suggest that it would facilitate understanding of all aspects of electromagnetic induction if one treated as fundamental Maxwell's equation relating the intensity of the induced electric field to the rate of change of magnetic field strength, and the equation giving the resultant force on a charged particle moving in magnetic and electric fields. From these the "loop" form of Faraday's Law can be derived. There is much to be said for this point of view, for it does lay stress on the fundamental physics of electromagnetic induction; by considering whether there can be an induced electric field acting at a

/given.....

given point in a circuit, or a magnetic force on a moving charge, or both, and by making the simplest choice of reference system, an intelligent student can get to the bottom of many difficult cases of electromagnetic induction. It is unfortunately not practicable to apply this recommendation fully in school teaching; whereas Faraday's Law in its loop form can be presented in terms intelligible to a pupil, this cannot be done with the expression for the intensity of an induced electric field. A sensible compromise would be to introduce Faraday's Law in its familiar, "loop" form, to point out the role of the induced electric field and of the force on a moving charge in various selected cases of electromagnetic induction, to make fully clear when one or the other of these causes can be expected to operate, and to avoid consideration of paradoxes, except possibly for brighter students. The last suggestion is made not with any base motive of wanting to pursue a "treacherous ostrich policy" (cf Scroggie 1955, 1963), but simply because time will be short and the prime emphasis should be placed on those cases which are most readily understood.

7.20.2 Faraday's Law: derived or "experimental"?

Is Faraday's Law in its more general form (i.e. embracing induction of both "variational" and "motional" forms) a derived law or an experimental one? It is very easy to derive the law for simple examples of motional EMF, e.g. a straight conductor sliding along the long arms of a U-shaped conductor mounted perpendicular to a uniform magnetic field, and the extension of this type of derivation to the general case of different portions of a flexible circuit moving in a non-uniform field presents no difficulty in principle. Thus for motional EMF's it is possible to regard Faraday's Law as a

/derived.....

derived law, the basis of the derivation being the expression for the force on a charged particle moving in a magnetic field. But what of the variational EMF? One may find conflicting views on this in the literature. On the one hand, we have Feather (1968): "... we shall take it for granted that the result, in the general case, derives directly from 'Ampere's Law' for the 'ponderomotive' force on a current element in a magnetic field - and the law of conservation of energy"; and Macfadyen (1958): "... if $\vec{B}$ is defined in the traditional manner it is possible to predict the law of electromagnetic induction from Ampere's experiments combined with the principle of energy conservation." Now one might have reason to doubt the generality of the proofs derived in the way these authors suggest, in view of the established fact that EMF's can be induced in circuits no part of which lie in a magnetic field and upon which no ponderomotive forces can act. Panofsky and Phillips (1962) state: "Note that [Faraday's Law] is an independent experimental law and is in no way derivable from any of the relations that have been previously used. In particular, contrary to the statement that is sometimes made, Faraday's law of induction is not the consequence of the law of conservation of energy applied to the overall energy balance of currents in magnetic fields." Shire (1960) has examined some of the alleged derivations of Faraday's Law and concludes: "None of these derivations seem to me to be completely satisfactory. I think it best to regard them, together with Faraday's thorough, if qualitative, experiments, as pointers to the right direction and to treat the law as a basic assumption, justified by its success in its many applications." Tricker (1966) has studied the problem in some detail; his summing-up is: "The deduction [of Faraday's Law] can only be made on the basis of certain

/other.....

other assumptions also, and, with the disappearance of the possibility of a mechanical aether, these have lost their plausibility. Modern writers thus appear to prefer to treat the law of induction as established independently by experiment." We have quoted two modern writers who prefer to treat the Law in this way; we could have quoted many, because the tendency now, as Tricker suggests, is to accord Faraday's Law a similar status to that of Coulomb's and Ampere's Laws. In advanced treatises on electromagnetism it is becoming common practice to take Maxwell's equations (one of which gives an expression for the electric field intensity induced by a time-varying magnetic field) and the Lorentz force Law as the axiomatic basis of electromagnetic theory.

Tricker (1966) has also taken up the question of the accuracy with which Faraday's Law may be said to have been established experimentally. It would seem that in recent times, no experiments have been carried out with this specific end in view. In an experiment to test Faraday's Law, the method of measurement of EMF and of rate of change of flux would both have to be independent of the Law. Tricker has considered various measurements carried out to standardize the volt or the ohm from this point of view. Two of these seem particularly promising: (i) comparison between the "standard ohm" established with reference to a calculable capacitance and a "standard ohm" established from induction experiments; (ii) comparison between the "standard ohm" established from heating experiments and that established from induction experiments. He points out that for method (i), the value of the permittivity of free space used to calculate the capacitance must not depend on measurements of the speed of light, since use of the latter would amount to assumption of the validity of Maxwell's electromagnetic theory, which

/has.....

has Faraday's Law as one of its cornerstones. Tricker concludes that "There is now ample experimental evidence to support Faraday's Law of the induction of currents to within an accuracy of about one in ten thousand", and adds "The law is, however, applied in circumstances where its accuracy is implied to some one in a million." By implication, he regards this state of affairs as unsatisfactory; elsewhere (Tricker 1967) he has suggested that the assessment of certain aspects of this matter could be a suitable problem for a student "to exercise himself on."

In an elementary course, it will certainly not be possible for teacher or pupils to undertake even a moderately accurate experimental verification of Faraday's Law. It will be possible to derive the law for a simple case of motional EMF; for the rest, pupils will have to be told that experimentally, it is found that the law is true in the general case also. There is certainly something to be said for the development of a simple experiment which will give senior school pupils a direct appreciation of the general validity of the law. As far as South African schools are concerned, the theory of such an experiment cannot depend on employment of differential or integral calculus. This imposes severe limitations on the method that can be used. It rules out the possibility of using conventional alternating current in, say, a long solenoid to provide a rapidly changing flux, since the expression for the rate of change of flux depends on the rate of change of current and so requires differentiation of the sine function. It also rules out the possible use of a ballistic galvanometer to measure the charge displaced in a circuit enclosing a changing flux. Ideally, what is wanted is a current varying linearly but rapidly with time in a circuit element which can produce an easily calculable flux for a given current in it. Perhaps pupils (and those interested in the teaching of physics)

/should.....

should be invited to "exercise themselves on" this problem.

7.20.3 Motional EMF's and "relativity"

We deal here with certain problems concerning motional EMF's which, while they need not concern pupils directly, do have some importance for teachers in that they draw attention to the need for care in the use of the concepts required for specifying or calculating motional EMF's.

(i) A straight conductor moving at right angles to its length and to a uniform magnetic field

We view this problem first from a reference system at rest relative to the sources of the magnetic field. Suppose the velocity of the conductor (rod) has magnitude v and the magnitude of the field strength is B. Consider any of the free charge carriers in the rod. Let it have charge q, which for convenience we shall suppose to be positive. This particle is moving at right angles to the magnetic field with speed v; therefore it must experience a force pushing it along the rod. This is true for any of the free charge carriers in the rod, so that the net effect of the motion of the rod in the field is for positive charge to accumulate at one end of it, leaving a deficiency of positive charge at the other. The accumulation of opposite charges at the end of the rod will continue until the electrostatic field set up by these charges exactly cancels the effect of the "magnetic" force on the free charge carriers. When this happens, the intensity of the electrostatic field must be equal in magnitude and opposite in direction to the force per unit charge exerted on the charge carriers by the magnetic field. The latter is equal to $\frac{Bqv}{q} = Bv$, so that the intensity of the electrostatic field must have magnitude Bv.

The "equilibrium" condition of the rod, then, is

/characterized.....

characterized by the following:-

(a) There is an accumulation of positive and negative charges at opposite ends of the rod.

(b) Because of this, there is an electrostatic field within and outside the rod. Inside the rod, this field is uniform, has intensity Bv and is directed from the positively- towards the negatively- charged end of the rod.

(c) The existence of this field means that there is a potential difference between the ends of the rod. If the length of the rod is L, this p.d. is equal to BvL.

(d) The intensity of this field is equal and opposite to that of the effective nonelectrostatic field acting on the charges. Because of the presence of this latter field, the rod can be thought of as a seat of EMF on open circuit. The EMF of the moving rod must/be equal to the energy per unit charge that would be transferred to a charged particle by the nonelectrostatic field if the particle were moved from one end of the rod to the other. This is just equal to $\frac{BqvL}{q} = BvL$, i.e. it is numerically equal to the p.d. between the ends of the rod, as we would have expected. (Alternatively, since the nonelectrostatic and electrostatic fields are oppositely directed within the rod, the line integrals of the characteristic vectors of these fields will be equal in magnitude and opposite in sign if evaluated in the same direction.) Thus a motional EMF of amount BvL is induced in the moving rod; if we insisted, we could ascribe the cause of this motional EMF to "flux cutting" - the rod sweeping across the lines of force.

So far, all is straightforward and familiar. Many would recognize in this situation the problem that frequently turns up in the following guise: "An aircraft is flying horizontally at constant speed v in a region in which the vertical component of the earth's magnetic field strength is B; find

/the.....

the p.d. between the wing tips [or: find the EMF developed in the wings]. " But let us now consider the entire situation from the viewpoint of a reference system in which the rod is at rest (Phillips 1963; Feynman et al 1964; Purcell 1965). How would we now describe what is happening? Our instincts (i.e. our ingrained belief in "Galilean" relativity) tell us that the change of reference system from one inertial frame to another ought to make no difference to our account of the fundamental facts of the situation. This might seem to be an open invitation to apply the idea of "flux-cutting" again, this time by saying that "the field lines are now moving relative to the rod, so that they are still being cut by the rod, in effect, and so we shall still get a motional EMF of amount BvL induced in the rod." (Cf Houston (1939): "If the lines of induction are regarded as moving with the magnet, the [rod] is cutting them and experiencing an e.m.f.") But we should be cautious. First, because the rod is now stationary in our chosen reference system, we are compelled to say that the free charged particles in it can experience no magnetic force - the velocity that appears in the expression for the force on a moving charge in a magnetic field is the velocity relative to the chosen reference system, not the velocity relative to the sources of the magnetic field. But somehow we would like to account for the fact that charges of opposite sign accumulate on the ends of the rod, and for the fact that free charges within the rod experience no net force. This we can do by supposing that in the surrounding space there is now a uniform electric field whose intensity has magnitude Bv and whose direction is parallel to the rod and pointing towards its positively-charged end. The rod is an uncharged, isolated and stationary conductor in this field, which induces equal and opposite charges on the ends of the rod. This means that inside the rod there is no electric

/field.....

field, and that the electric field in the neighbourhood of the rod is not uniform, since the presence of the induced charges on the ends of the rod causes some field lines to terminate on one end of the rod and others to start from the other end, and other lines to be distorted from straightness and parallelism with one another. This resultant field is, of course, the superposition of two fields - a uniform field extending through all space (including the inside of the conductor) and the field set up by the charges induced on the surface of the conductor. Inside the conductor, these fields annul one another exactly. (These field patterns are illustrated in Purcell (1965).)

Most interesting of all in the present context is the fact that <u>all points of the conductor are at the same potential</u> - to an observer at rest with respect to the rod, there can be no potential difference between the ends of the rod. So "the p.d. between the wing-tips" calculated in the "standard problem" referred to above is essentially a <u>calculated</u> quantity - it would not be measurable in the aeroplane itself. As Feynman et al (1964) have pointed out, it is possible in principle to measure the potential difference between two points in the <u>surrounding</u> field, but this is not a practical proposition as the fluctuations in the atmospheric electric field are enormous in magnitude compared with the magnitude of the induced electric field.

What is the origin of the induced electric field? We have seen that our instinctive view that a change from one inertial system to another should not alter the basic physics has led us to postulate the existence of this field. In effect, we have reasoned thus: Viewed from the original reference system, a particle carrying charge q and having velocity v experiences a force Bqv at right angles to its motion and the direction of the magnetic field. Viewed from the new

/reference.....

reference system at rest relative to the particle, we should expect the particle to experience a force in the same direction and of the same magnitude (if its velocity is small compared with that of light.) This time, however, we cannot attribute this force to the magnetic field, since the particle is not moving in it. The only agency to which we can conceivably attribute this force is an "induced" electric field. Application of the special theory of relativity readily confirms the above account of the situation, subject to the restriction to small relative velocities (cf Feynman et al 1964, Purcell 1965 and many other sources).

In a sense we still have not answered our opening question fully. An electric field must presumably have sources of a kind. In the present instance, what are they? (The authorities cited above make no attempt to answer this question, at any rate not in this context). Again, special relativity can come to our aid, if we direct our attention to the sources of the magnetic field. Suppose that this field, which we have supposed to be uniform, is set up by an enormous current-carrying conducting sheet, appropriately orientated. Viewed from a reference system moving relative to the sheet, the latter will appear to have a net charge, on account of the different length contractions undergone by the positive and negative charge distributions in the sheet (cf Sears and Zemansky 1964, Feynman et al 1964 or Purcell 1965 for a discussion of the corresponding situation for a current-carrying wire). It is this net "polarization" of the sheet that sets up the observed electric field. Alternatively, if the magnetic field is produced by a huge permanent magnet, one can show that this device will similarly appear to be "polarized" when viewed from a reference system moving relative to it (cf Cullwick 1959, Panofsky and Phillips 1962), and so is

/capable.....

capable of producing an electric field. Thus the setting-up of the induced electric fields in situations where we have motion relative to the sources of a magnetic field, which would have been necessary but inexplicable in terms of "Galilean" relativity, is seen to be both necessary and explicable in terms of Einsteinian relativity.

From a reference system in which the rod is at rest, the description of the rod as a seat of EMF seems less certain. We can still regard the rod in this way if we consider the energy that would be transferred to a charged particle by the induced electric field as the particle goes from one end of the rod to the other. This will give us the same answer as before, but the procedure may seem somewhat artificial, since the induced electric field in this case is just as much electrostatic as is the field set up by the accumulation of opposite charges on the end of the rod. The position will become clear when we study below the case of the rod as part of a closed circuit.

We give below a summary of the conditions in and around the rod as seen from the reference system in which the rod is at rest:-

(a') There is an accumulation of positive and negative charges on opposite ends of the rod, as the rod is a stationary conductor in an electric field.

(b') There is no net electric field inside the rod; within the rod, the induced electric field and the field set up by the charges on the surface of the rod exactly cancel.

(c') The potential difference between any two points on the rod is zero.

(d') The intensity of the induced electric field has magnitude Bv. If we consider the effect of this field alone, we can calculate from it an "EMF" of BvL for the rod, representing the amount of energy that it would transfer per unit charge to a particle displaced through the rod. But this seems

/a.....

a dubious procedure if we insist on defining EMF in terms of the energy transferred by nonelectrostatic fields, since the induced electric field here is unquestionably electrostatic in character.

(ii) Conducting rod (XY) sliding along the long arms of a U-shaped conductor mounted with its plane at right angles to a uniform magnetic field

(Cf. Rosser 1968a and Scanlon et al 1969 for discussions which partly overlap with that given here.) We now allow the rod we have been discussing to form part of a closed circuit. Again we shall view the problem first from a reference system at rest relative to the field sources (and the U-shaped conductor) and then from a reference system in which the rod is at rest.

From the reference system at rest relative to the field sources, the situation is particularly easy to describe (see Fig. 7.20.3(a), page 790A). Of the free charge carriers in the circuit, initially only those in the rod will be moving in the magnetic field; and only those in the rod will experience forces pushing them along the rod. A charged particle taken round the circuit will only have energy transferred to it from a nonelectrostatic field while it is within the rod, and the amount of energy per unit charge it will receive there is BvL, which is thus the EMF induced in the circuit as a whole. In this case the seat of the EMF in the circuit is the moving rod.

From the reference system at rest relative to the rod, the situation is more intriguing. (see Fig. 7.20.3(b), page 790A). In the rod itself there is no net electric field or nonelectrostatic field, and within it no energy is transferred to the circulating charge from non-electrical sources. The same is true for the long arms of the U-shaped conductor,

/since.....

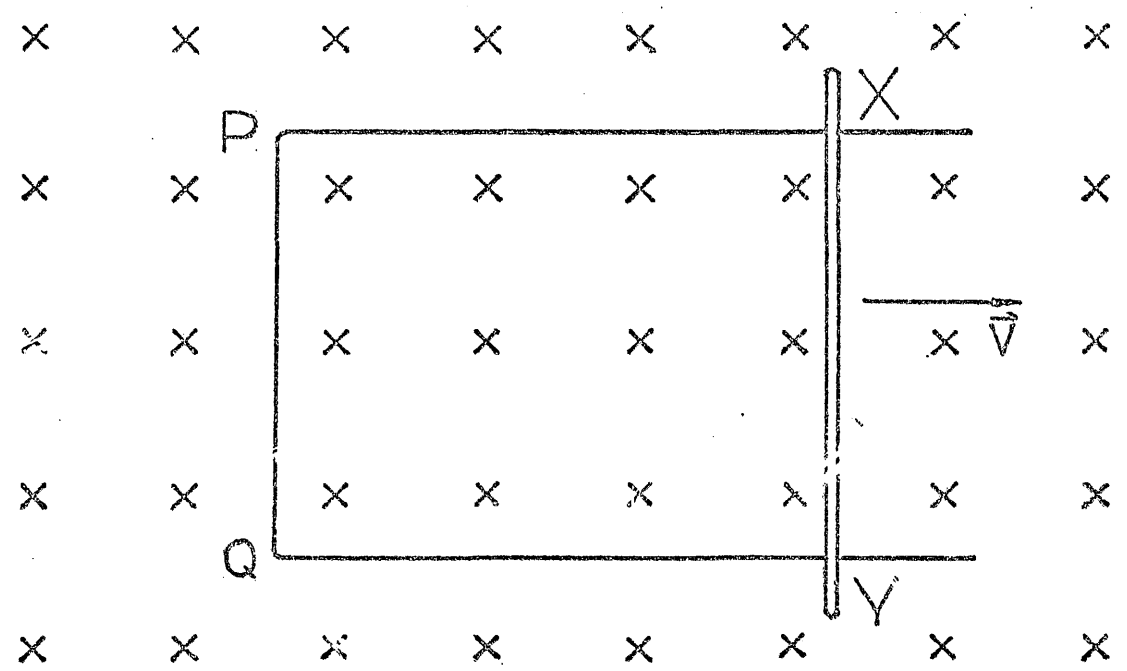

Fig. 7.20.3(a)

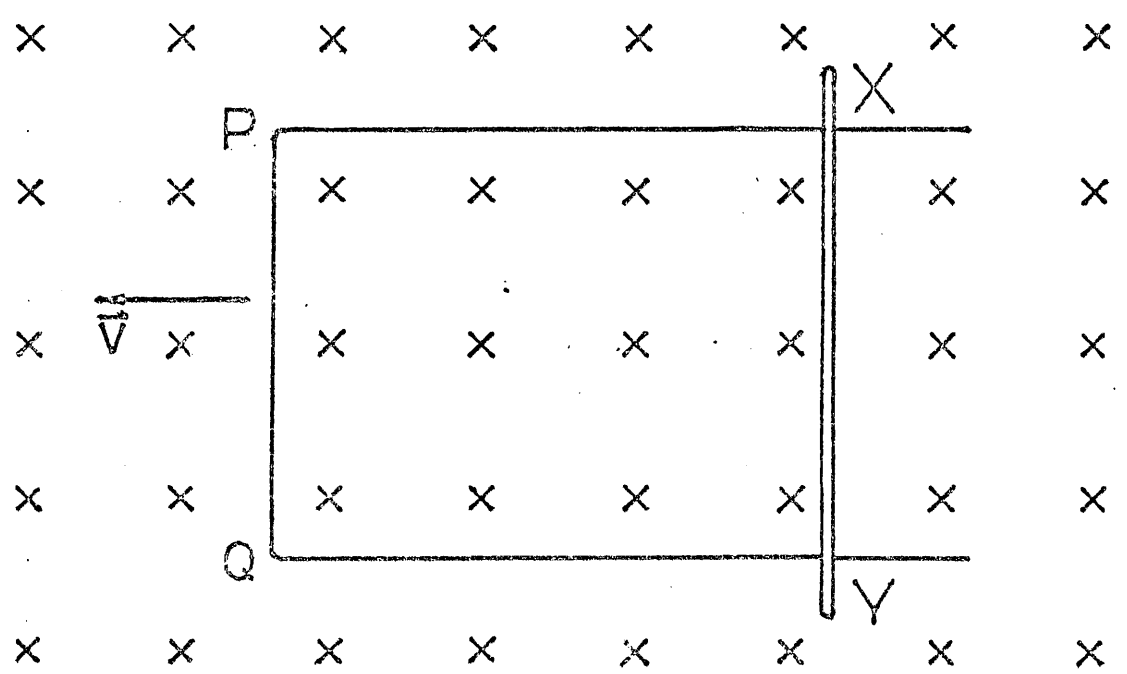

Fig. 7.20.3(b)

since although the free charges in them are certainly moving in the magnetic field, the "magnetic" forces they experience do not displace these charges. This leaves us to consider the conductor PQ forming the crosspiece of the U. This is also moving in the magnetic field, and the charge carriers in it will experience magnetic forces impelling them towards one end (Q) of it. However, these forces will be opposed by the effect of the induced electric field directed from Q to P, which will in fact exactly cancel them. Hence the charges in PQ experience no net force, and there is no accumulation of opposite charges at the ends of this conductor. This does not mean that no EMF is induced in the circuit; it would be alarming if it did! There is an accumulation of opposite charges at the ends of the rod, XY, as a result of the induced electric field, and this will be maintained as long as the relative motion between field sources and reference system is maintained. Thus charge will keep circulating around the circuit and will acquire energy reversibly from nonelectrical sources in PQ, which is moving in the magnetic field. The amount of the EMF is easily shown to be BvL, just as before; in this case it is found by calculating the energy transferred per unit charge by the "magnetic" force Bqv to charges displaced through PQ. Viewed from the reference system in which the rod is stationary, PQ and not the rod, is the seat of the EMF induced in the circuit. This illustrates Phillips's (1963) observation that "one can localize the $\vec{v} \times \vec{B}$ fields [i.e. the "magnetic force" fields] which contribute to [the EMF] only when one specifies the observer."

In the simple case which we have considered, in which one part of a circuit moves relative to the rest in a magnetic field, we have seen that the induced EMF depends only on

/the.....

the relative motion between the "moving" part of the circuit and the sources of the field. To this extent one might feel justified in talking about relative motion between a conductor and the lines of force, or "flux-cutting". However, we have seen that the seat of EMF was that portion of the circuit which was moving in the magnetic field relative to the reference system, so that we should not be too hasty in leaping to this conclusion. There are cases in which the "motional" EMF can be localized in the same part of the circuit, whichever inertial system the problem is viewed from; e.g. a fixed circuit dragged across a magnetic field of limited extent. Here the seat of the EMF can only be that part of the circuit that is in the magnetic field. From a reference system at rest relative to this portion of the circuit, one sees an induced electric field which extends only over that region of space that is filled by the magnetic field, and so acts only on part of the circuit, instead of on all as in the above example.

7.20.4 The teaching of electromagnetic induction

(i) Motional EMF's Pupils come to this material with some familiarity with the basic phenomena and with some understanding of how motional EMF's are caused. Since motional EMF's can be readily understood in terms of concepts learned earlier, it is sensible to begin with them. A convenient example to introduce the subject is that dealt with in the foregoing section, viz. a circuit consisting of a U-shaped conductor and a conducting rod sliding on it, the whole being mounted at right angles to a magnetic field. This should be treated from the point of view of a reference system at rest relative to the magnetic field sources. It is not essential to discuss the case of the isolated moving conductor

/first,.....

first, though if time permitted this could be done; in a course of this type the pupil's main concern will be with closed circuits, and no complications are introduced by the treatment of closed circuits right from the start. The discussion should be fully quantitative; that is, the expression BvL for the induced EMF should be derived, not so much for its intrinsic interest as for its identification later with the rate of change of flux through the circuit. The velocity v of the sliding conductor should be referred to as the velocity "relative to the laboratory" (in which the magnetic field sources are stationary), or occasionally as the velocity "relative to the field sources", but preferably not as the velocity "relative to the field" and never as the velocity "relative to the field lines" (cf Burling 1963). To speak of a velocity "relative to the field lines" begs the question of whether or not field lines can be said to have identity and "move with the magnet" (or other field source); to speak of a velocity "relative to the field" raises questions of what one means by the field and how one can be said to move relative to it. The careful teacher will surely want to avoid the use of unfortunate expressions like these, as well as the misleading expression "flux-cutting".

It is worthwhile analysing the above example of motional EMF further to show the operation of the Principle of Conservation of Energy. As should be realized by pupils, the direction of the induced current is such as to give rise to a magnetic force on the rod opposing its motion across the field. (This is a special case of Lenz's Law, a fact of which pupils can be reminded later on.) If the induced current is i, the magnitude of this force is iBL, where B and L have their usual meanings, and an equal and opposite force must be exerted on the rod to keep it moving with constant velocity. Since this velocity is v, the amount of work done per unit

/time.....

time in moving the rod is iBLv, i.e. energy is being transferred to the rod at the rate of iBLv joules per second. But, from the definition of EMF, an amount of energy BLv joules per second is being transferred per unit charge to the charge carriers in the rod. Then if an amount of charge q flows in a time t, the energy supplied per second is BLv x $\frac{q}{t}$, or BLvi. Therefore, as one might expect, energy is being supplied to the charges in the circuit at the same rate at which it is being transferred to the rod.

There is a further point worth mentioning here, since some pupils might raise it. (The first person to raise it in the author's experience was a professor of physical chemistry.) Since magnetic forces act at right angles to the direction of the displacement of a particle, they cannot presumably transfer energy to it. Yet in cases of motional EMF such as the above we <u>are</u> concerned with energy transfer from the magnetic field to the charged particles. Is there not a contradiction here? The contradiction is only apparent, however; it is removed when we rectify a slight looseness in wording in the above description, where we spoke of energy being transferred "to the rod" by the mechanical force pushing on the wire. Strictly speaking, the energy is not being transferred to the rod, which gains neither kinetic nor potential energy; it is being transferred to the magnetic field, which supplies the force opposing the mechanical force. The magnetic field, as we have seen, transfers energy at the same rate to the charge carriers in the wire. Therefore on balance the magnetic field neither gains nor loses energy; it simply acts as a channel through which energy is transferred from the external force to the charge carriers.

(ii) <u>Introduction of magnetic flux: Faraday's Law</u>

Next, the above example should be analysed in terms of the concept of magnetic flux. It should be apparent to

/pupils.....

pupils that the induction of the EMF in this circuit is associated with a change in the area of the circuit and with this an increase in the "amount of magnetic field" enclosed by the circuit. This idea must now be made more precise by the introduction of the concept of magnetic flux. This is easily defined for a uniform field: the magnetic flux through a given plane area in a uniform magnetic field is the product of the area and the component of the magnetic field strength normal to the area, i.e. the flux $\Phi = BA\cos\theta$, where B is the magnetic field strength, A the area and θ the angle between the normal to the area and the direction of the field. When the field is normal to the area, $\Phi = BA$, and B is seen to be equal to the magnetic flux per unit area, or flux density. In SI units, flux is measured in webers and flux density can obviously be expressed in weber/metre^2 (Wb/m^2), i.e. an alternative, equivalent unit to the tesla or newton/amp-metre.

To apply the concept of magnetic flux to the study of the above circuit, one can prodeed as follows:-

In a time Δt, the wire moves a distance $v\Delta t$, so that the area of the portion of the circuit within the magnetic field is altered by an amount $vL\Delta t$ and the flux through the circuit changes by $BvL\Delta t = \Delta\Phi$. The change of flux per unit time is $\frac{\Delta\Phi}{\Delta t} = BvL$, which is just the expression derived earlier for the induced EMF. (Introduction of the sign convention whereby we are able to write the induced EMF equal to minus the rate of increase of flux would seem to be a needless complication in an elementary course. The direction of the induced current is easily deduced for motional EMF's from one of the hand rules; for transformer EMF's it can be deduced from Lenz's Law, if required.) It can next be pointed out that Faraday's Law - the fact that the induced EMF is equal to the rate of change of magnetic flux through the circuit -

/appears.....

appears to apply to all cases of induced EMF, including those in which there is no relative motion between a circuit and the sources of a magnetic field. It should be stressed that in the general case, Faraday's Law seems to have the status of an experimental law. (cf Section 7.20.2.)

(iii) "Variational" or "Transformer" EMF's

In discussing "transformer" EMF's, the teacher could first deal with a simple example of a circular loop of wire placed with its plane normal to a changing, uniform magnetic field. There is no need to deal quantitatively with this case, but a few important points will have to be made:-

(a) The circuit need not be in the magnetic field for an EMF to be induced in it; all that is necessary is that it must surround a region of changing flux.

(b) The entire circuit must be considered the seat of EMF.

(c) Since the circuit is stationary (in the laboratory frame of reference) and also because of point (i), we cannot ascribe the induction of an EMF in the circuit to forces on charges moving in a magnetic field. The inescapable conclusion seems to be that a changing magnetic flux is responsible for inducing an electric field which, in turn, gives rise to the movement of charges which constitutes the induced current.

(d) This induced electric field is not like an electrostatic field: its lines must form closed loops (in idealized situations, at any rate - consider our earlier discussion in Section 7.15.6) if it is to be able to supply electrical energy to a charge taken right around a closed loop. This means that it is not possible to specify a potential difference between any two points of this circuit, since the work done in transporting a charge from one point to the other on the loop will depend on the path chosen. (Phillips (1963) has discussed this case in some detail).

/(e).....

(e) The field lines of the induced electric field need not coincide with conductors, just as the field lines of electrostatic or magnetic fields do not have to coincide with material objects.

7.20.5 Further discussion of "variational" or "transformer" EMF's

Certain implications of the induction of a transformer EMF in a region where there is no magnetic field are worth examining more closely. We consider an idealized version of this case, namely the EMF that would be induced by a changing current in a very long solenoid. Outside the solenoid, there is virtually no magnetic field but, according to Faraday's Law, there will be an induced EMF in a closed loop surrounding the solenoid, produced by the induced electric field associated with the changing flux. The symmetry of the situation suggests that the field lines of the induced electric field should be circular closed loops concentric with the axis of the solenoid. (In this idealized case we may treat the field lines as closed loops without running into any difficulty.) This induced field is clearly not conservative; that is, if we were to take a test charge around certain closed paths in it, there would be a net transfer of energy from the field to the particle (cf part (iii) (d) of the preceding section.). But it follows from Faraday's Law that one can find certain closed paths in the induced field around which we could take a test charge without it receiving energy from the field. Indeed, this would be true for any closed path which lies outside the solenoid and does not encircle it. In more sophisticated language, the curl of $\vec{E}$ vanishes everywhere outside the solenoid. Thus we have an example of an electric field that is definitely not conservative, but for which it is not true that its curl is everywhere non-zero; this serves as a reminder that for a field to be classed as

/conservative.....

conservative, its curl must vanish everywhere. (Inside the solenoid, the curl of $\vec{E}$ does not vanish.)

From Faraday's Law, it also follows that the magnitude of the electric field vector outside the solenoid is inversely proportional to the radial distance from the centre of the solenoid. By considering a closed path of suitable shape, or by direct differentiation, it is easy to show that this result is consistent with the curl of $\vec{E}$ being zero outside the solenoid. From more advanced electromagnetic theory, it is possible to show that the "vector potential" of the magnetic field of the solenoid does not vanish outside the solenoid (cf Feynman et al 1964, Scott 1959), although $\vec{B}$ does. In any induced electric field, the magnitude of $\vec{E}$ at a point is given by the partial time derivative of the vector potential at that point; applicability of this result to the present example confirms the result stated above, and provides an alternative way of conceiving of the production of the induced electric field, viz. the changing vector potential at a point "sets up" an electric field there.

7.20.6 The field concept and electromagnetic induction

Is the pupil in need of any further "explanation" of the phenomenon considered in the previous section? Sears (1963c) considers that the main puzzle for students is how the electrons in a wire loop surrounding a region of changing flux "get the message that something is going on" in the source of magnetic flux. In Sears's view, this puzzle is solved for the student once one has drawn attention to the existence of the induced electric field; it is the latter which accelerates the electrons and so gives them the message that something is going on elsewhere. On this view, there would be no need to take the discussion further than we have taken it above. But some may still be puzzled: how does a changing

/magnetic.....

magnetic field produce an electric field? Is the question correctly put, in any case? What lies behind it is the feeling that while electrostatic fields have an obvious "cause", namely the mere presence of excess charge on an object, and magnetic fields can clearly be associated with current-carrying circuits, the induced electric field seems to represent a further degree of abstraction - it arises when a magnetic field changes, and there seems to be no obvious "mechanism" linking the two. It is natural to try to view the problem in these terms, and those who wish to make the attempt should be asked to look at how the magnetic field is made to change. As Edwards (1965) puts it, "... the flux-linking rule gives no information about the physical mechanism of induction until we inquire how it is that the function Φ[magnetic flux] comes to be changing". The result of the inquiry can be incorporated into the following account of the situation:-

The "changing magnetic field" of the type under discussion is produced by a changing current. If a current is changing, the drift velocity of the free charges responsible for it must be changing; i.e. the averaged-out motion of the free charges must be <u>accelerated</u>, not uniform motion. It turns out that the force exerted by an <u>accelerated</u> charged particle on a stationary charged particle is not the same as that which would be predicted by Coulomb's Law. It is this extra contribution to the electric force arising from the acceleration of the charged particles which corresponds to the induced electric field <u>associated with</u> the changing magnetic field.

From this "action-at-a-distance" account, it might seem that it is not strictly correct to refer to the changing magnetic field "producing", "creating", "setting up" or "generating" the induced electric field. Instead, "The induced electric field and the varying magnetic field have a common cause in the past, namely, moving charges" (Rosser 1963b).

/Should.....

Should one, then, discourage the use of such expressions as "a changing magnetic field produces an electric field" or "a changing electric field produces a magnetic field"? Warburton (1954) has strongly urged this, and specifically suggested that the relation between the rate of change of electric flux and the strength of the associated magnetic field should be derived directly from the motions of the charges producing both fields. (His main reason for adopting this attitude appears to be his belief that Maxwell's "smoothed-over" field theory allegedly suppresses certain aspects of electromagnetic theory suggestive of the ballistic emission of radiation.) If it were practicable and helpful always to refer induced electric and magnetic fields back to the motion of the charges from which they ultimately derive, then there would be a case for adopting these suggestions. The practical situation is very different, however. The field equations (Maxwell's), among which are those whose physical content is often expressed by the phrases under discussion, can be used to derive formulae which relate the electric and magnetic fields at a point to the motions of charged particles. (It is this derivation which lies behind the "explanation" of the induced electric field given in the previous paragraph.) These formulae are undoubtedly useful - they are applied to the solution of certain problems concerning electromagnetic radiation - but it is fair to say that for many physicists, the habitual approach to problems in electromagnetism is by way of the field equations. It is natural that it should be so, for the equations are so simple.

But is not this approach "artificial", concealing the "reality" of the situation, which lies in the interactions between charged particles? One's answer to a question like this will depend on the view one has of the "reality" (or otherwise) of electric and magnetic fields. When one pauses

/to.....

to consider the status of other "abstract" concepts in physics, such as energy, one should certainly hesitate before attempting to deny "reality" to the electric and magnetic field concepts. These concepts occupy a central position in an established physical theory; for this reason alone one might feel justified in ascribing as much "reality" to them as to other physical concepts, including electric charge. Feynman et al (1963) have remarked that "..., the laws of force are so complicated that it turns out that the fields have a reality that is almost independent of the objects which create them...If the force upon some charge depends upon where another charge was yesterday, which it does, then we need machinery to keep track of what went on yesterday, and that is the character of the field. So when the forces get more complicated, the field becomes more and more real, and this technique [the separation of a particle - particle interaction into two parts - 'particle A creates field' and 'field acts on particle B'] becomes less and less of an artificial separation". We would prefer to say here that the reality of the fields becomes more and more apparent as the "forces get more complicated". However, the message of the quotation is clear: when we have to deal with complicated situations in electromagnetism, we still have a simple basis for the description of what happens, namely the behaviour of the fields, and in such situations we are particularly conscious of the "reality" of the fields.

Without becoming dogmatic on this question, we can at least state that there are no secure grounds for relegating the "field" description of particle-particle interactions to a lower status than that of an "action-at-a-distance" description directly in terms of forces. With this in mind, we can now make the following points concerning the elementary teaching of electromagnetic induction and matters related to it:

/(i).....

(i) It is usually much more convenient to describe the processes of electromagnetic induction in terms of magnetic and electric fields then of forces between charges. There is nothing wrong with describing or thinking of these processes in these terms, and the obvious and indeed the only practical way for a teacher to make quantitative sense of these processes for his pupils is to do so in terms of the fields and the physical laws which relate them.

(ii) When the pupil studies electromagnetic waves, he will have to understand them entirely in terms of the field concepts. For him to be ready for this experience, he must be familiar with the description of electromagnetic induction in terms of <u>fields</u>.

(iii) If we adopt Rosser's (1963b) view that "the induced electric field and the varying magnetic field have a common cause in the past, namely, moving charges", then we would have to insist that to speak of a changing magnetic field "setting up", "creating", "producing" or "generating" an electric field is not strictly correct, and that it would be better to say that the one field is <u>associated with</u> the other. But we have reasoned above that the "field" description of electromagnetic processes does not necessarily have a lower claim to validity than an "action-at-a-distance" description in terms of interacting charges; and in the present instance there seems to be nothing to prevent one taking the view that what happens in electromagnetic induction is that the changing magnetic fields associated with the accelerating charges <u>set up</u> the induced electric fields. There seems to be no way of establishing the superiority of one or other of these descriptions by experiment. (In this connection, see also Section 7.21.2.) The teacher who prefers to be cautious here could say that "it follows from Faraday's Law of Electromagnetic

/Induction.....

Induction that whenever we have a changing magnetic field in a certain region of space, there will be an induced electric field in that region." This is a factual statement which could be accommodated within either view of the phenomenon. (Cf. the "neutral" title of the present section.) On the other hand, the teacher who prefers to say that "the changing magnetic field produces the electric field" cannot be accused of seriously misleading his pupils; this positive statement might be a very useful prop for pupils to cling to at this stage, and at the very least it represents a defensible point of view.

(iv) It is natural enough that a pupil should be curious about the "physical mechanism" of electromagnetic induction, as we have suggested at the beginning of this section. To satisfy this curiosity, one could give the pupil the "action-at-a-distance explanation" cited above of the origin of induced electric fields, but at the same time one should make clear to him that the details of the force set up by a collection of moving charges on a stationary charge are obtained in practice from application of the field equations. Thus there is but little comfort to be gained from this "explanation"; it presents an alternative view of the phenomenon, but it would be too much to say that it is the "true" view.

7.20.7 Self-induction

This is an important aspect of electromagnetic induction which cannot be ignored, even in an elementary course, since it leads to the concept of energy stored in the magnetic field and is essential to the derivation of the equation relating the input and output voltages of a transformer to its turns ratio. Self-induction is best introduced experimentally, with observation of the slow rise of current switched on in, say, a large electromagnet, and of what happens when the circuit is broken. In analysis of what is observed, it can be

/explained.....

explained that an induced EMF, opposite in sense to the potential difference applied to the electromagnet, is present in the coils of the magnet while the current grows in view of the change in the flux linked by the coils, and the effect of this induced EMF is to ensure that while the current is increasing not all the energy supplied to the coils is dissipated as internal energy in them. The other energy is evidently expended in increasing the current, thereby building up the magnetic field of the coil. That this energy is stored and can be recovered is demonstrated when the circuit containing the coil is broken; the large EMF that results from this very sudden change of flux through the coils can cause an impressive spark across the contacts of the switch. If a neon indicator lamp is connected across the terminals of the magnet, it will light up briefly when the switch in the circuit is opened.

7.20.8 The dynamo and alternating current

These topics will have been studied qualitatively in the junior years. If he wished, a teacher could show a bright group of pupils that the EMF generated in a simple A.C. dynamo must vary sinusoidally with time, by pointing out that the component perpendicular to the magnetic field of the velocity of the appropriate parts of the rotor must vary with time in this way.

7.20.9 The transformer

Explanation of the principle of operation of the transformer should present no difficulty, provided that the distinction between motional and variational EMF's has been properly clarified earlier. In deriving the equation $\frac{V_2}{V_1} = \frac{N_2}{N_1}$ (the symbols have obvious meanings), the assumptions made should be clearly stated, viz. that the transformer is so

/designed.....

designed that the same magnetic flux links each turn of the primary and the secondary, and that the resistances of primary and secondary coils may be neglected. Let $\frac{\Delta\Phi}{\Delta t}$ be the rate of change of flux in each turn of the primary and secondary. Then, since the induced EMF's in the turns of each coil are in series and must add, it follows that the <u>self-induced</u> EMF in the primary coil is $N_1\frac{\Delta\Phi}{\Delta t}$ and the induced EMF in the secondary is $N_2\frac{\Delta\Phi}{\Delta t}$. If the instantaneous potential difference applied to the primary is V_1, then this must be numerically equal to $N_1\frac{\Delta\Phi}{\Delta t}$ if the resistance in the primary is negligible. If the secondary is open-circuited, the potential difference V_2 across it must be $N_2\frac{\Delta\Phi}{\Delta t}$. Hence $\frac{V_2}{V_1} = \frac{N_2}{N_1}$. It should be argued from the Principle of Conservation of Energy that the output power from the secondary must be less than or equal to the input power to the primary so that, at best $i_2 = \frac{V_1}{V_2} i_1$.

Applications of the transformer have for long been a standard feature of syllabuses and the standard discussions of them do not seem to contain any noteworthy errors. Accordingly, we shall not consider them here.

7.20.10 <u>Alternating current circuits</u>

In view of the ubiquity of A.C., it is highly desirable for pupils to receive some sort of introduction, chiefly experimental, to the rudiments of the theory of A.C. circuits. It would be easy for pupils to carry out an experiment to show the effect of the inclusion in an A.C. circuit of inductors and capacitors. For this purpose, a lamp could serve as indicator of the current; a more sophisticated indicator could be an oscilloscope connected so as to show the variation of potential difference across a resistor in the circuit. The contrast between the behaviour of these circuit elements

/in.....

in A.C. and D.C. circuits can be brought out; a capacitor does not "block" alternating current, whereas it does do this to direct current; an inductor has a much bigger effect on alternating current than it would have on direct current established in the same current. Also, it can be shown that the inclusion of a capacitor or an inductor in an A.C. circuit will reduce the current, just as the inclusion of a "pure" resistance would, so that capacitors and inductors can be said to have a property analogous to resistance. It is important to point out that the energy conversions that take place in "pure" inductors and capacitors when an alternating current is established in them are <u>reversible</u>, unlike the energy conversion in a resistor.

It can be briefly indicated that variations in current are not in step with variations in potential difference in inductors and capacitors, and this can be demonstrated with the use of a current of very low frequency (1 Hz) and centre-zero indicating instruments. The reasons for the distinctive behaviour of inductors and capacitors in A.C. circuits can be briefly explained; charge can flow into and out of a capacitor, and the potential difference across an inductor must depend on the rate of change of flux in it, which in turn will depend on the rate of change of current.

The use of a lamp as an indicator of current gives a clue to the way in which current and potential difference can be measured in A.C. circuits. It should be obvious to pupils that while an oscilloscope could be used to measure maximum values of these quantities, it would often be more convenient to use some kind of meter instead, and that the kind of meter used cannot be the conventional D.C. ammeter or voltmeter. Before deciding on the kind of meter that should be used it is necessary to decide what it is to measure; if it is to be

/some.....

some sort of "average value" of the current or p.d., how is this average to be defined? It obviously cannot be what is ordinarily meant by "average," since the average value (in the usual sense) of a sinusoidal alternating current or potential difference is zero when evaluated over an integral number of cycles. For many practical purposes it would clearly be sensible to choose as the "average" values of current or p.d. those D.C. values of these quantities which would produce the same power dissipation in purely resistive circuit elements as the actual currents or p.d.'s. In a resistor whose resistance did not depend on the current, the equivalent D.C. value of the alternating current in it would simply be the average value of the square of the current, taken over a whole number of cycles, or "root mean square" current. Without their having to go into the details of the relation between the R.M.S. and peak values of current and potential difference, the mathematics of which will be beyond them, pupils should know that A.C. ammeters and voltmeters are customarily calibrated to read R.M.S. values, and they should know something of the principle of operation of these instruments, e.g. "hot-wire", "moving-iron" and "thermocouple" meters. In this connection the use of ordinary moving-coil instruments in conjunction with rectifiers to measure alternating currents and voltages should also be mentioned, since these are nowadays in common use.

7.20·11 The induction coil

This topic - another carry-over from traditional syllabuses but perhaps worthy of inclusion because of its use in motor-vehicle engines - is still listed in the Common Basic Syllabuses; it is discussed here principally because of the part played by the capacitor connected in parallel across

/the.....

the "points" in the primary circuit. The standard textbook account of this aspect of the behaviour of the induction coil is along the following lines: "...for a short time after opening the contacts current continues to flow charging up the capacitor. This gives time for the contacts to move far enough apart for arcing not to occur" (Bennet 1968). This is certainly part of the story, and would be adequate as far as pupils were concerned, but teachers might wish to know a little more about it. The essential point is that during the "break", the primary circuit is essentially an "oscillatory" circuit consisting of an inductance and a capacitance in series and, provided that the effective resistance in this circuit is small enough, conditions exist in this circuit for damped oscillations of current (and p.d.) to be set up. Because these oscillations are extremely rapid, the induced EMF in the secondary can attain very large peak values. (Page and Adams (1931) have analysed this situation in some detail.)

7.20.12 Interchange of energy between electric and magnetic fields

Reference was made in the previous section to the behaviour of an "oscillatory" circuit containing an inductance and a capacitance in series. It is useful for the pupil to understand the action of this commonly-used circuit from the viewpoint of the interchange of energy. He already knows about energy stored in electric and magnetic fields and the phenomenon of self-induction, and this should form an adequate background. Suppose that a previously-charged capacitor is allowed to discharge through a coil (assumed to have negligible resistance). Because of the effect of self-induction, the current is initially zero; it then builds up gradually

/until.....

until the capacitor is completely discharged. At this stage all the energy originally stored in the electric field of the capacitor has been transferred to the magnetic field of the coil. Self-induction keeps the current going in the same direction, the current decreasing in magnitude until the capacitor is charged up to the initial extent but in the opposite sense; the energy has now been transferred back from the magnetic field to the electric field again. In this idealized circuit the interchange between electric and magnetic field energy would go on indefinitely; in practice the unavoidable presence of resistance in the circuit (if it is not superconducting!) means that energy is constantly being dissipated as internal energy in the circuit, so that if such oscillations are to be maintained, energy will have to be supplied to the circuit to compensate for the energy that is "lost" as internal energy. (Of course, if the frequency of oscillation is high enough, there will also be appreciable energy lost from this circuit by radiation.)

There are a number of mechanical analogies for the behaviour of this circuit; the oscillations of a mass suspended from a spring, a flywheel attached to a clock spring and oscillating on its axis, or even the oscillations of a simple pendulum. The damped oscillations of potential difference in a circuit like this can be displayed on the screen of an oscilloscope. If an audio-frequency oscillator with a "square-wave" as well as a "sine-wave" output is available, the square-wave output can be applied via a semi-conductor diode to a capacitor (capacitance 0.001 to 0.01 μF), across which an inductance (approximately 0.1 henry is suitable) and a resistance (preferably variable, between 100 and 1000 ohms) are connected in series. The p.d. across the resistor, inductor or capacitor can be applied to the Y-input terminals of the oscilloscope.

/7.21.....

7.21. The nature and propagation of electromagnetic waves

7.21.1 Electromagnetic waves in the school syllabus

Water waves and longitudinal and transverse waves on springs and stretched strings can be observed directly and are thus susceptible of direct study, whereas only the effects of sound and electromagnetic waves can be observed. In practice, the inability to see sound waves does not prove an insuperable obstacle to understanding their mode of propagation, as it is easy to visualize longitudinal waves being propagated in solids and liquids in much the same way as they are propagated along springs, and it is likewise easy to visualize them being propagated through a gas conforming to the model provided by the kinetic theory. No such simple physical picture is available to anyone wishing to explain the propagation of electromagnetic waves. In this case, the disturbance that is propagated is a change in the electric and magnetic fields at a point in space, and those who wish to explain what is happening must do so in terms of these relatively abstract concepts. It is scarcely surprising, therefore, that until quite recently this topic has either not been included in elementary physics syllabuses or, if it has, teachers have been expected to give it only a most perfunctory treatment. And yet it is natural for pupils to feel curious about a phenomenon of which they hear so much and which is of such general importance; and it is equally natural for the keen teacher to want to explain it to his pupils. But in the past, such a teacher will have been handicapped by the lack of an account of electromagnetic waves which was at once convincing, easily understood at an elementary level, and readily available to him. Nor did the attitude of the planners of syllabuses encourage the emergence of such an account. The Physical Science Study Committee are therefore to be congratulated on taking the bold step of including this topic in

/their.....

their course and on making a carefully considered attempt to produce in their textbook the kind of account of it that is required. As we shall see, this attempt was not without minor errors in exposition which, however, the P.S.S.C. were quick to rectify in the second edition of their textbook. (It is pleasant indeed to be able to record an instance of prompt action being taken to prevent the perpetuation of a faulty tradition in the teaching of a branch of physics).

What approach should be adopted in trying to explain the propagation of electromagnetic waves in an elementary physics course? Before it became an accepted practice to include this topic in such courses, a number of elementary textbooks and popular books sought to "explain" it by considering radiation from a dipole antenna to which a rapidly-alternating potential difference is applied. (One of the best of these accounts is to be found in Rogers (1960.).) They described the formation of closed-loop lines of force in the electric field of the antenna and represented these as being pushed out by new lines of force created at the antenna. Now "fanciful" accounts of this kind (the adjective is Rogers's) can help to make the idea of electromagnetic waves seem plausible, but they raise a very big question: how do we know that the lines of force really do "snap off" the antenna as closed loops? In other words, can electromagnetic disturbances really travel through space and if so, how do they start off and how do they keep going? These are the issues which are really fundamental and with which any treatment of the subject, even an elementary one, ought to concern itself.

7.21.2 The association of a magnetic field with a changing electric field

It is impossible to explain the propagation of electromagnetic waves without reference to Maxwell's "Theorem D", viz. that a magnetic field is associated with (or "produced

/by".....

by"[a])/changing electric field, or, as Maxwell himself put it, "When the electric displacement increases or diminishes, the effect is equivalent to that of an electric current in the positive or negative direction" (Maxwell 1868, quoted by Bork 1963. See also article 610 of Maxwell's "Treatise on Electricity and Magnetism", 1881.). Understandably enough, the P.S.S.C. authors tried in the first edition of their textbook to make this idea plausible in terms of[b] simple (hypothetical) experiment viz. investigation of the magnetic field near a charged capacitor which is slowly discharging on account of slight ionisation of the air between its plates. According to these authors, the magnetic field produced by the motion of the ions between the plates, which constitutes a current, would be exactly cancelled by an opposing magnetic field set up by the changing electric flux between the plates. In principle, this cancellation of the magnetic field of the flowing ions should be observable but, as the P.S.S.C. authors explain, the practical difficulties would be formidable. However, as French and Tessman (1963) have pointed out, the above argument is erroneous. Suppose one tries to calculate the value of $\vec{B}$ at a point in the field of a circuit containing the capacitor, when its charging or discharging current is slowly varying or "quasi-steady", as in the above example. On the face of it, one should do this by using the Biot-Savart Law, taking into account the contributions from the actual moving charges in the circuit and from the changing electric field of the capacitor. (Obviously an approach of this kind lay behind the reasoning of the P.S.S.C. authors.) As various authors have shown, the "curl" of the "displacement current density" is negligible in quasi-steady conditions, from which it follows that the net contribution to $\vec{B}$ from all parts of the changing electric field must be negligible, if we calculate this contribution with the aid of the Biot-

/Savart.....

Savart Law (French and Tessman 1963, Purcell 1965, Rosser 1968a). Therefore we can account for the magnetic field associated with this circuit by applying the Biot-Savart Law solely to the real currents in it, i.e. to the currents associated with moving charges. This means that there can be no "magnetic field of the displacement current" which can cancel the contribution to the field from the ions moving between the plates of the capacitor.

Clearly one cannot appeal to this kind of thought experiment as a justification of Maxwell's hypothesis. In the second edition of the P.S.S.C. textbook (1965), the faulty discussion of this phenomenon has been replaced by a treatment based on Ampere's Circuital Law, i.e. the standard treatment given in intermediate courses and in Maxwell's own approach to the problem. In both editions of the P.S.S.C. textbook, the final summing-up is that "Even today, the existence of electromagnetic waves, behaving just as Maxwell predicted, is the most convincing proof of the equivalence between current and the changing electric flux."

There is an aspect of the above discussion which needs to be considered further, lest the wrong impression be gained from it. We have referred to the proof that the magnetic field of a "quasi-steady" current can be accounted for entirely in terms of the moving charges in it. As Tricker (1969) has recently emphasized, this does not mean that the magnetic field has to be accounted for in this way, i.e. that the proof cited above establishes that the "displacement current" does not or cannot produce a magnetic field in the above circumstances. Tricker's argument may be rendered (and expanded) as follows:-

It can be deduced from the application of the principle of relativity to the laws of electromagnetism that the magnetic field of a charge q moving with a uniform velocity $\vec{u}$ having a magnitude very much less than the speed of light,

/is.....

is given by the formula

$$\vec{B} = \frac{q\vec{u} \times \vec{r}}{4\pi\varepsilon_0 c^2 r^3}$$

This is just the Biot-Savart Law, with the moving charge being treated as a "current element" $q\vec{u}$. We can have two views of the origin of this field. We can say either that the moving charge is the "sole originator" of the field, or that "the moving charge is accompanied by a distribution of displacement currents", in which case the above formula for $\vec{B}$ represents the field produced by the "whole system", moving charge plus distribution of displacement currents. As various authors, including Scott (1959) and Rosser (1968a), have shown, it is possible to deduce this formula by working directly from Maxwell's equations, in particular from the equation which includes the "displacement current" term. One might reasonably say that in this approach to the calculation of the field, one is treating the "changing electric field" resulting from the motion of the charge as the "originator" of the magnetic field. (Rosser himself firmly rejects this view: see below, this section.) We have no way of telling which of these views of the origin of the field is "correct"; either is consistent with both theory and experiment.

Now we consider Rosser's (1968a) proof that the magnetic field of a displacement current cannot contribute to the magnetic field of a quasi-steady current. He bases the proof on the electric and magnetic fields associated with a single moving point charge. Putting ourselves in Rosser's shoes, we would seem initially to have to reason thus: "It is possible that the magnetic field of a slowly-moving charge is made up of two contributions, one from the charge itself and the other from the distribution of 'displacement currents' associated with the varying electric field of the moving

/charge".

charge." We would then use the Biot-Savart formula to find the latter contribution, and show that it must sum to zero. This would leave us with only the contribution from the charge itself. The latter we would also obtain by using the Biot-Savart formula. But as this is known to give the correct result for the actual field, it should not surprise us that the "displacement current" contributions to the field as calculated from the Biot-Savart Law should sum to zero. In fact, it really has been assumed all along that the moving charge is the sole originator of the magnetic field; as Tricker remarks, "We have simply left nothing for [the displacement currents] to do in our initial assumptions."

What this "proof" really establishes is that it is "wrong in principle" (to use Rosser's own phrase) to substitute the "displacement current" term into the Biot-Savart formula in order to calculate a contribution to the magnetic field from the displacement current. The correct method of finding such a contribution is to work directly from Maxwell's equations; as already noted, this would lead us in the case of a single moving charge back to the above-quoted formula, only this time we could interpret the derivation as implying that the changing electric field of the moving charge is the "sole originator" of the observed magnetic field.

It may be argued that all this is not of much practical interest, since it is usually much simpler to obtain the magnetic fields of quasi-steady currents by correct use of the Biot-Savart Law than by working with displacement currents and Maxwell's equations, and the answer is more important than the way we interpret it. The real cases of interest in regard to "displacement currents" are those in which conditions are far from quasi-steady (e.g. radiation from oscillating charges), and here we may expect to see magnetic effects

/produced.....

produced by the rapidly changing electric fields. But this way of describing the latter situation would still be found objectionable by Rosser (1963b, 1968a). According to him, "In itself, [the displacement current] produces no magnetic field. It is associated with magnetic fields, only in as much as both the electric and magnetic fields in a system arise from a common cause, namely the moving charges in the system." (Cf his remark about induced electric fields quoted in Section 7.20.6.) In fact, "it is wrong in principle to say that the 'displacement current'...produces a magnetic field" (both quotations from Rosser 1968a). These assertions are open to the same objections as were raised in Section 7.20.6 against Rosser's remarks concerning the origin of electromagnetically induced electric fields. We do not and apparently cannot know whether the magnetic field associated with a changing electric field is "really" produced by the accelerating charge or by the changing electric field. Either view is permitted by the circumstances, but for the reasons advanced in Section 7.20.6 one might well find the latter view more attractive.

In our view, the introduction of Ampere's Circuital Law for steady currents and Maxwell's modification of it to make it apply to varying currents as well, are subtleties which go beyond the limits of the kind of elementary physics course under discussion in these pages. How do we avoid them if we are nevertheless to introduce Maxwell's hypothesis? We feel that it would be reasonable to take a short cut and start the discussion off in the following way: "We now know that an electric field is produced by a changing magnetic field. Can we interchange the two key adjectives in this sentence?" One could then go on to relate that in 1864 Maxwell had argued that the answer to this question must be in the affirmative; that he had been thinking about problems connected with the

/behaviour.....

behaviour of varying currents in circuits and had pointed out that unless one assumed that a magnetic field could be associated with a changing electric field, one encountered serious logical difficulties in dealing with these problems (cf. Section 3.3). (This piece of history is not essential to the presentation of our recommended "short cut", but it might be psychologically advantageous to introduce it at this point, even though it would not be possible to give any details of the "logical difficulties" referred to.) In pursuing the point, it would be wise for teachers to point out that it is to all intents and purposes impossible to check this assumption about electric fields directly and that the best justification for it is furnished by the predictions based upon it, concerning which details will shortly be given.

It is tempting to use as a justification of Maxwell's postulate the simple example that has been discussed above in considering whether or not a changing electric field can produce a magnetic field. Take a charged particle at rest in a given inertial system. In this system it is surrounded by a spherically symmetrical electric field, and there is no magnetic field associated with it. But suppose we now view it from a reference system moving with uniform velocity relative to the first. At a given observation point in this system, one will "see" an electric field that is changing in intensity, and according to Maxwell's prediction, this should be accompanied by a magnetic field. But we know already that a moving charge sets up a magnetic field - this idea comes to the fore in the course of our study of steady-state electromagnetism. So in this instance Maxwell's prediction leads to a conclusion which agrees with something we have already decided to accept. This example could be mentioned, in order to confront the pupil with an alternative way of looking at a

/familiar.....

familiar situation, at least in the case of a slowly-moving particle.

To render Maxwell's postulate in more precise terms, it is necessary to introduce the concept of electric flux. Since this is defined in analogous fashion to magnetic flux, it should present no difficulty. As with magnetic flux, the "electric flux" through a given loop can be defined for a uniform field: it will only be used in connection with uniform fields in what follows. Then for our purposes we can formulate Maxwell's postulate as follows: if the electric flux through a given area is changing, a magnetic field will be set up with field lines which are perpendicular to the electric field lines and which form closed loops.

The term "displacement current", which denotes the effective current corresponding to a rate of change of electric flux, should not, in our opinion, be introduced to pupils. Since they will not meet Ampere's Law or its modification, it is not essential for them to have the concept of "rate of change of electric flux" related to "current", though no harm would be done by pointing the parallel between the two. Maxwell's term "electric displacement", from which the term "displacement current" derives, is meaningless in terms of present-day interpretations of electromagnetic theory, and to use the word "displacement" in connection with the phenomenon discussed here could easily bewilder and confuse a pupil.

7.21.3 The propagation of electromagnetic radiation

To explain the mechanism of propagation of electromagnetic radiation, it is obviously wise to deal with the simplest form of electromagnetic wave. Consciously or unconsciously following Heaviside's discussion of the problem (cf Bork 1967), a number of writers have begun their treatment of it with a consideration of the advance through space of a boundary between a region of uniform electric and magnetic

/fields.....

fields and "no field" (cf Carr and Sells 1960, P.S.S.C. 1965, Feynman et al 1964, Arons 1965, Pierce 1965, Mara 1966). That this is indeed the simplest choice of wave to make is evident from the facts that the field configurations are uniquely easy to deal with and that the electric and magnetic fields are everywhere in phase (compare the radiation from an "oscillating" dipole in this respect). Purcell (1965) prefers to use "gable-"shaped pulses in his introductory treatment of electromagnetic waves; he points out that abrupt changes in $\vec{E}$ or $\vec{B}$ (such as are implied by the existence of a sharp boundary between "some field" and "no field") cannot occur in empty space. In our view, this is not a serious objection to what seems to have become the standard treatment; in order to make the latter as "respectable" as Purcell's, it is only necessary to specify that the change from "some field" to "no field" is not abrupt but takes place over a small but finite distance. One can always make clear that the "standard" treatment deals with an idealized situation which can only be approximated to in practice.

The authors cited differ slightly in their manner of setting the stage; Carr and Sells, the P.S.S.C. authors and Arons all start with the wave already on its way and make no reference to what could have been its source, whereas Feynman et al (and Orear (1961, 1967), who says he is using the Feynman treatment) do describe the source - a very large (strictly-speaking, infinitely large) plane conducting sheet in which a current is suddenly turned on. In our view more stage-setting than this is needed to help the average pupil to focus on the problem. We suggest that a suitable introduction could start from the phrase "radio waves" and the knowledge that "transmitters" and "receivers" exist. This terminology implies that a disturbance of some sort is propagated from transmitting antenna to receiving antenna. Pupils can also be told,

/if.....

if they do not already know, that when a transmitting station is broadcasting, a high-frequency alternating current is generated in the wires of which the antenna is constituted and that when the programme being broadcast is "received", a corresponding alternating current is produced in the antenna of the receiver. So whatever the disturbance is that is propagated, it must have its origin in the alternating current in the transmitting antenna and must somehow give rise to a corresponding alternating current in the receiving antenna.

As a first attempt at an explanation of this phenomena, one can try the theory that the transmitting and receiving antennas are like the primary and secondary coils of an enormous transformer. In the transmitting antenna, there is a changing current which must set up a changing magnetic field in the surrounding space; in the receiving antenna, there is a changing current which must presumably result from the induction of a changing electric field in the wire. Since a changing magnetic field gives rise to an electric field, it looks as though we might now have established the desired link between what happens in the transmitting and receiving antennas. But this is obviously only part of the story, since in the argument just presented it is not clear how there could be a time lag between a change in current in the transmitting antenna and a corresponding current in the receiving antenna. If one really is dealing with a kind of wave motion, there must be such a time lag.

It can now be pointed out that the other induction principle of which the pupil has by now heard, viz. that a changing electric field can "produce" a magnetic field, can also be expected to apply in a situation such as this. Now the pupil is ready to study wave propagation from an idealized "plane-current-sheet" transmitter of the type mentioned above. First, he has to realize that the magnetic field of such a

/current.....

current sheet must be uniform, parallel to the sheet and perpendicular to the direction of the current, and oppositely directed on the two sides of the sheet. The teacher can then argue as follows:

"When the current is switched on in the sheet, the effect will be that we have a change in the surrounding space from 'no magnetic field' to 'some magnetic field' (assuming that there was effectively no field present before the current was switched on). We shall now assume that this change does not occur simultaneously at all points in the surroundings; instead it travels out from the sheet as a kind of wave motion. If our assumption is correct, then shortly after the current has been turned on, the boundary between 'field' and 'no field' will be at a certain distance from the sheet; a little later it will be further away, and still later even further away. If we imagine a series of observers placed on a line perpendicular to the sheet, each armed with an instrument for detecting a magnetic field, then these observers will successively detect the magnetic field of the current in the sheet; the further the observer is from the sheet, the later he will detect the field."

It will be seen that the approach here is to assume that the "message" that the field has been turned on travels out through space at a finite speed, and to pursue the consequences of this assumption. All the authors cited above make this assumption; at this level, there seems to be no alternative. As Arons (1965) remarks, one does much the same sort of thing when one derives expressions for the speeds of travel of sound waves or waves on strings without deriving the "wave equation". There is, however, this difference: that when one sets out to derive these expressions in the mechanical cases mentioned, one does so with a good idea of how the disturbances in question are propagated, whereas in the present instance the mechanism of propagation is the very thing we are

/trying.....

trying to find out. By the same token, one has a good idea of how sound waves or waves on strings "take off" from their sources - one can see it happen in the case of the latter - whereas this knowledge is not available for the assumed electromagnetic waves. It is as well for this difficulty to be recognized and squarely faced from the outset, so that the teacher has no illusions about what he is trying to do.

In the account given by the "hypothetical teacher" above, reference was made to the advance of the boundary between "field" and "no field". This is the mode of expression adopted in the second edition of the P.S.S.C. textbook, in contrast to that employed in the first, where the authors talk about slabs of magnetic and electric fields moving through space. In the latter case one sees what is meant readily enough, but, as we have already noted, it is desirable to proceed with some care when talking about "moving" magnetic and electric fields. It is certainly preferable to refer to the advance of the boundary of the field than to say that the field is moving outwards. We are concerned here with the propagation of changes in fields and not of the fields themselves.

To return to our hypothetical teacher, who continues his summary of his approach:- "Let us pursue the consequences of our assumption a little further. Consider an imaginary rectangular loop drawn with its plane perpendicular to the direction of the magnetic field and to the current sheet, and placed so that one pair of sides is parallel to the sheet. As the wave front separating 'field' from 'no field' crosses the left-hand side of this loop, the magnetic flux through the loop increases, so an electric field must be induced. A detailed study of this point would show that the electric field is induced only along the side of the loop crossed at

/that.....

that instant by the 'wavefront'. When we apply this principle to any similar loop being crossed by the wavefront, we realize that there must be an induced electric field everywhere between the wavefront and the sheet; in other words if we have a magnetic field whose boundary advances in the way we have described, it must be accompanied by an induced electric field at right angles to it."

Again, we need to make certain comments before proceeding. The above account leaves a certain amount unsaid; in particular, as is indicated in it, it fails to provide a proper justification for the fact that the induced electric field is to be found only in the region occupied by the magnetic field. Rigorous justification can only come from Maxwell's equations themselves; these show that the induced electric field is as described, with its lines of force closing up at "infinity", just as those of the magnetic field do in the present instance. In the first edition of the P.S.S.C. textbook the induction of the electric field is attributed, not to the change of the flux through a loop, but directly to the passage of the magnetic field past a given line. This is perilously like "flux-cutting", though if ever there was a situation to which "flux-cutting" is not in the least applicable, it is the present one. In the second edition the treatment has been altered and it now agrees with that to be found in Carr and Sells (1960) and Arons (1965). The rate of change of flux through a loop is considered, and the far side of the loop is taken to be well in advance of the wavefront. Because of this, the contribution to the induced EMF (and consequently to the induced electric field) is taken to be zero along the far side of the loop, while the contributions to the induced EMF from the long sides are assumed to cancel each other because of symmetry. This leaves the contribution to the EMF from the near side of the loop as the only nonzero and

/uncancelled.....

uncancelled one, and shows that it is only in line with or behind the wavefront that we will find a component of the induced electric field perpendicular to the magnetic field and the direction of advance of the wavefront. Now these arguments can scarcely be called rigorous, and those who use them do not pretend that they are. Perhaps we can accept that the approach of our "hypothetical teacher", who omits these arguments altogether, will do almost as well in an elementary course: with brighter pupils the matter can be gone into in more detail.

The "hypothetical teacher" could now complete his summary as follows:-

"According to the principle of induction suggested by Maxwell, when the electric flux through a loop changes, we should expect to find a net component of magnetic field around the loop. We have shown that if we have a boundary between a magnetic field and no field advancing through space, this boundary also separates a region of induced electric field from no field. So let us now study what happens as this boundary advances across a rectangular loop whose plane is parallel to the magnetic field and perpendicular to the induced electric field. Using the same argument as before, we can see that everywhere between the wavefront and the sheet, the electric field must be accompanied by a magnetic field which it generates by virtue of its advance; and this magnetic field must be at right angles to the directions of both the electric field and the advance of the wavefront. In fact, if we are careful about how these directions go, we shall find that the direction of this magnetic field is exactly the same as the one we started with at the beginning of this discussion. Now we see where all this has led us: if the advance of the wavefront means that an induced electric

/field.....

field must accompany the magnetic field, and a magnetic field in the same direction as the original must accompany the electric field, then perhaps it really means that each field 'keeps the other going' - that such a wavefront advances by a mechanism of mutual regeneration of the two kinds of field at right angles to each other and to the direction of travel of the wavefront. For this to be true, each field must generate the other in exactly the right strength; and for this to happen we might expect that there is a particular critical speed at which such a wavefront must travel. Maxwell worked out this speed and found it was equal to 3×10^8 m/s, i.e. the speed with which light is known to travel. After Maxwell's death, Hertz showed that if he set up an electromagnetic disturbance in one part of the laboratory, it travelled with the right sort of speed and that these supposed 'electromagnetic waves' could be reflected and refracted."

Thus the story of the mechanism of propagation of electromagnetic waves is complete. Or rather, not quite complete, since no attempt has been made to carry out an actual calculation of the speed with which the electromagnetic waves have to travel. Since the relation between the induced magnetic field strength and the corresponding rate of change of electric flux is assumed not to be given in the form of an equation in the syllabus under discussion, it will not be possible to derive the expression for the speed of the waves. Although this will deprive pupils of the pleasure of seeing the magic number 3×10^8 emerge from calculations performed by themselves or their teacher, this loss is not too serious; it is more important for them at this stage to know that there is a characteristic speed of propagation for electromagnetic disturbances.

If he desires, the teacher could continue the argument to explain that any arbitrary "plane wave" electromagnetic

/disturbance.....

disturbance could be synthesized from an appropriate combination of "step functions" of the type whose propagation has just been discussed (cf P.S.S.C. 1965). We would not, however, suggest that it is essential to make this point. It is more important to stress the propagation of energy by electromagnetic waves. It is easy to do this with reference to the above example. Since pupils are already familiar with the idea of energy being stored in electric and magnetic fields, they should appreciate that the advance of the wave-front dividing "field" from "no field" means that energy is streaming out into space, and it is easy for the teacher to draw the analogy between this situation and the corresponding one in the propagation of mechanical waves. The mechanism by which energy is fed into the fields in this instance should not be left out of account. A check of the direction of the induced electric field will show that it is necessarily opposite to that of the field which is applied to the conducting sheet to maintain a current in it. So in driving the charge carriers through the sheet, the applied field does work against the induced electric field and so transfers energy to the combination of electric and magnetic fields. (Pierce (1965) is the only one of the modern authors cited above who makes this vital point.) This will be recognized as being identical in all essentials with what happens when a current is switched on in a coil - we are dealing with the same phenomenon.

7.21.4 The generation of electromagnetic radiation

There remains the matter of the initiation or generation of electromagnetic waves. Again one can make the essential point by referring to our simple example of electromagnetic wave propagation. The wave front separating "field" from "no field" started to travel out from the sheet after the current had been switched on. Originally the free electrons

/in.....

in the sheet had zero average drift velocity; when the current was switched on they were accelerated until they acquired some steady average drift velocity. Therefore to start an electromagnetic wave off from the sheet, it was necessary to accelerate charged particles in it. Since it can now be accepted by pupils that radio transmitters send out electromagnetic waves, one can refer to what happens in the antenna of a transmitter as additional evidence of the connection between the acceleration of charged particles and the initiation of electromagnetic waves. In the antenna of a transmitter, a high-frequency alternating current is maintained; the motion of the charge carriers in such a current is accelerated motion. Experimentally, then, there is a clear connection between the acceleration of charged particles and electromagnetic radiation; can we show that there must be such a connection? The P.S.S.C. authors attempt this by means of a "relativistic" argument (P.S.S.C. 1960, 1965). They point out that "a charge in constant-velocity motion is really no different from a charge at rest"; we would not expect radiation to be emitted from a charged particle if we simply move away from it at constant velocity, or, what amounts to the same thing, it moves away from us at constant velocity. We have referred above (Section 7.21.2) to the electric and magnetic fields associated with a uniformly-moving charge; as the P.S.S.C. authors put it (though they do not try to justify the statement directly), "...the electric and magnetic fields stay with a charge in constant-velocity motion". (A study of the Poynting vectors associated with these fields does confirm this statement, but this is hardly elementary physics.) So, it would seem that it is only when one accelerates a charge that one has the chance of persuading it to radiate. This still does not answer the question we put above; we have shown that accelerated

/charges.....

charges do radiate, and that it is only when a charge is accelerated that it can radiate, but we have not shown that it must radiate. To do that, we seem to require the full apparatus of sophisticated electromagnetic theory, so that in an elementary syllabus, the question must be left open.

The relationship between the acceleration of charges and the supply of energy to the waves that are generated as a result, is possibly open to misunderstanding. One of the authors of the new J.M.B. Teachers' Guide for Physical Science (Prinsloo et al 1968) has written: "The energy which the electromagnetic radiation carry [sic] away is derived from the energy supplied to the accelerating charge." In a sense, this is correct, but it should not be taken to mean that it is only when charges are accelerating that energy can be supplied to the fields composing the wave. Again, the simple example of transmission from a plane sheet makes this clear. In this idealized situation, energy is being supplied to the wave even while a steady current is maintained in the sheet, since work is being done against the constantly present induced electric field. This is admittedly a very special case, but it does emphasize that the existence of the induced field, rather than acceleration as such, is the fundamental requirement for the supply of energy to electromagnetic waves.

We would not regard radiation from a dipole antenna as a syllabus topic, but in view of its practical importance and the fact that it is illustrated in so many textbooks, the teacher would be well advised to acquire some background knowledge of it. As we have already indicated, the often-reproduced diagrams showing closed loop lines of force being pushed away from the vicinity of the antenna, while helpful, do not tell the full story of the basic physics. In Part Four of the Teachers' Resource Book and Guide for the P.S.S.C. course, an attempt is made to give some indication of the basic physics. Because we believe that the P.S.S.C. account contains

/the.....

the essence of the matter but does not say enough about it, we take it as the starting-point of our discussion.

A simple, two-rod dipole antenna is shown connected to a battery, the upper rod being negatively charged and the lower rod positively charged, and some of the field lines of the resulting electric field are drawn in on one side of the antenna. (These lines are not drawn right on to the antenna; this is perhaps unfortunate, as we shall see.). (See Fig. 7.21.4(a) page 829A). Then: "...let us suddenly switch the battery to the other polarity so that the top rod becomes positively charged and the lower one negatively." The diagram illustrating the new situation is drawn so as to show the field as it appears before enough time has elapsed for the static field of the new distribution of charges to become fully established. (See Fig. 7.21.4 (b), page 829A).

A few field lines nearest the antenna are drawn as pointing from the upper rod to the lower, while those further out are shown just as they were in the earlier diagram. In short, as far as the field lines are concerned, the two diagrams are identical, except that the field lines nearest the antenna have had the arrows on them turned round. In order to make our point properly, we quote the P.S.S.C. authors' comments on this diagram in full: "There was no mistake made in the drawing. The field beyond P is shown just as it was before, while only the field near the antenna has changed. The point is that the regions beyond P have not yet had time to 'learn about' the change of the polarity of the antenna. Why not? When the electric field was changed in direction, a magnetic field was generated perpendicular to the paper. But if a magnetic field appeared where none was before, it changed. This changing magnetic field generated an electric field which was just such as to compensate for the change in the original

/electric.....

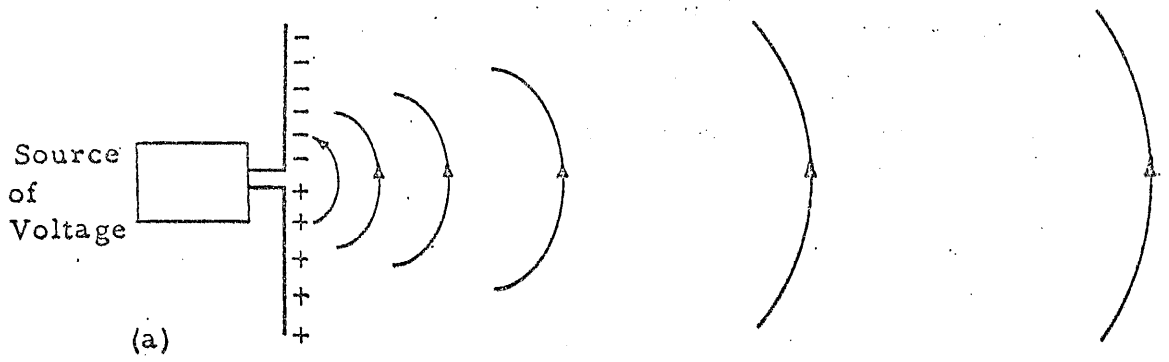

Fig. 7.21.4(a)

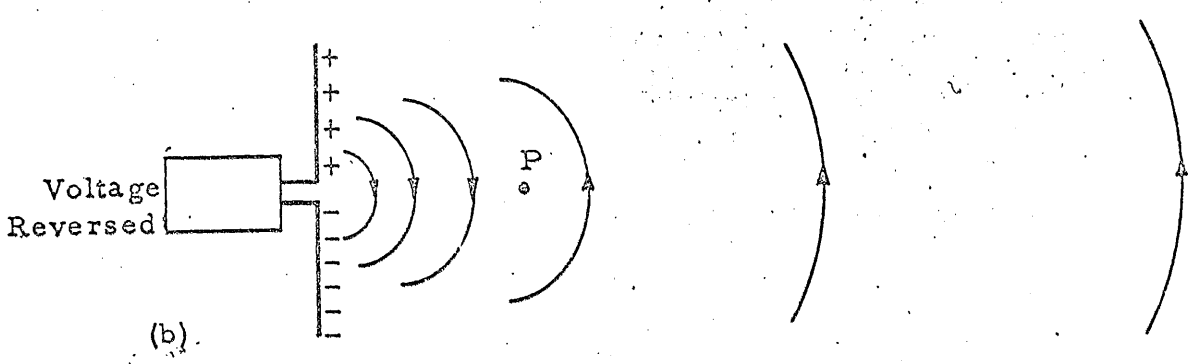

Fig. 7.21.4(b)

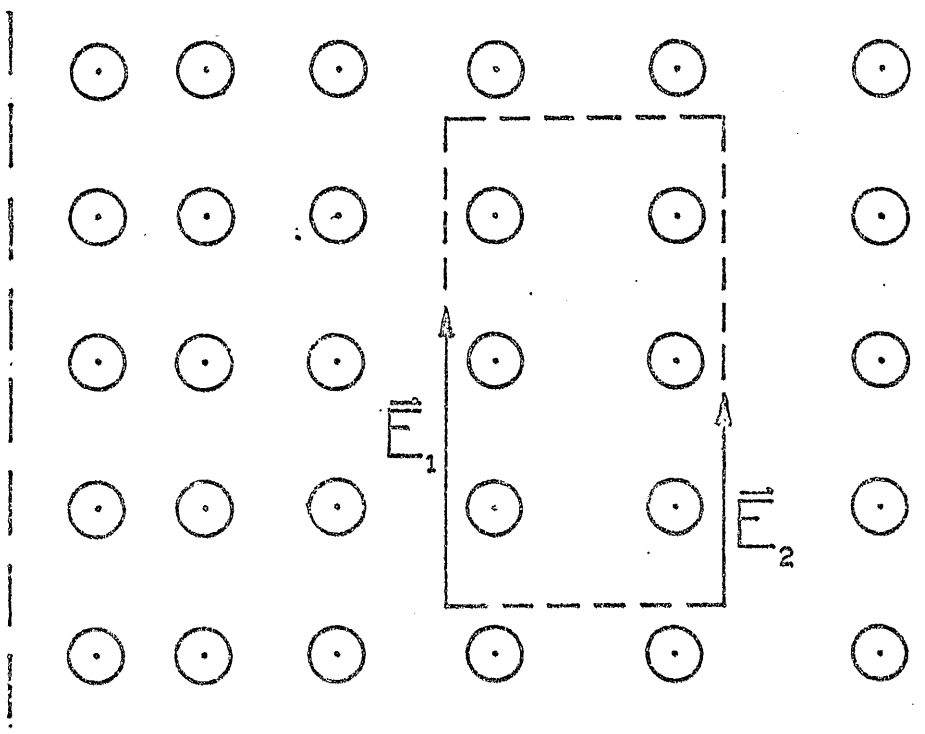

Fig. 7.21.4(c)

electric field for points beyond P. The result of all this is that points distant from the antenna can 'learn about' the change in polarity of the antenna only through a changing electric and perpendicular magnetic field which can travel only at the speed of light."

All the statements made in this quotation are correct, but we do not believe that the average teacher will regard all of them as being so obviously true as not to require further elaboration. Consider the central sentence of the quotation, in which it is asserted that an electric field is generated which just compensates for any change in the electric field at points beyond P. Teacher and pupil should by now accept readily enough that a changing electric field will "produce" a changing magnetic field which will in turn produce an electric field, but it will be far from obvious to them that in the present instance this process of mutual recreation of fields will have the stated effect. This is because prior to this they will only have studied the simplest case of electromagnetic wave propagation, in which the lines of force of the magnetic and electric fields were straight lines "closing up at infinity". Since the lines of force of the fields participating in radiation from a dipole do not form such simple patterns, a deeper insight into the induction of electric and magnetic fields is required for the understanding of wave propagation in this case. It might occur to the teacher from consideration of Lenz's Law that a sudden change in the electric field must paradoxically produce no change, but he would doubtless like to see how it works.

The root of the problem would be more clearly seen if in both P.S.S.C. diagrams, the lines of force were not abbreviated but shown in full. It would then be clear that in the second P.S.S.C. diagram, the lines of force of the (apparently) "unchanged" part of the field, unlike those of the newly-emerging "static" field of the reversed dipole, form

/part.....

part of closed loops. In fact, these lines must sweep round to connect up with certain of the outermost of the lines to the left of the point P in the second diagram. However, diagrams showing an extended two-dimensional picture of the field in this particular case do not seem to be given in textbooks - at all events, we have not been able to find any books in which they appear - and the P.S.S.C. may well have been reluctant to be the first to provide such diagrams.

To justify the assertion that the lines of force of the "unchanged" portion of the field must be part of closed loops, one may use the following argument. If we assume the "answer", viz. that the original field is to be "swept out of the way" by the new at the speed of light, then the lines of the original field that still remain in the picture can no longer be attached to the charges on the dipole, and if they are not so attached, then they must form closed loops. (From the general appearance of the pattern of field lines, it is also clear that the lines in question cannot close up at infinity.) Electric field lines of this character can only be the lines of an induced field, i.e. one that has been set up by a changing magnetic flux. So we see that for the original electric field to be "sustained" after it has become severed from the dipole, we need a change of magnetic flux.

Now we can return to the P.S.S.C. explanation and elaborate on it. Suppose that when the connections from the battery to the dipole are reversed, the original field tries to "flip over" all at once. In doing so, it would produce a very rapid increase in downward-directed electric flux through a plane perpendicularly bisecting the axis of the dipole, and this would set up a magnetic field whose lines of force would surround the axis of the dipole. This means a rapid increase in magnetic flux directed out of the paper on the right-hand side of the dipole, and into it on the

/other.....

other side. To see what kind of induced electric field results from this, one could consider the change in magnetic flux passing through a narrow rectangular loop with its long side parallel to the dipole axis. (See Fig.7.21.4(c), page 829A). The linear dimensions of the loop are supposed to be small compared with the distance of the loop from the dipole axis; for the sake of clarity, this condition is not observed in our diagram. Lenz's Law (and a commonsense realization of what would happen in a number of such rectangular and parallel loops placed side by side) would show one that the induced electric field must have an upwards component on each long side of the rectangular loop, and that the electric intensity is bigger on the side nearest the dipole. (The relative lengths of the vectors labelled $\vec{E}_1$ and $\vec{E}_2$ in the diagram are intended to show this.) This is precisely the field pattern that is wanted (at any rate as far as its variation along a perpendicular bisector of the dipole is concerned), viz. an upwards-directed electric field diminishing in strength with increasing distance from the dipole.

In accordance with expectations, the induced electric field has in this case been found to have the same direction as the original field; and perhaps one can now take on trust that its intensity is just so much as to compensate for the attempted change in electric field produced by the reversal in connection. This means that the message that the connections have been reversed will not travel out instantaneously but must be propagated at the speed of light by the process described earlier. A teacher who wants to gain experience in thinking about this topic would find it instructive to work through the process of propagation in detail, showing that the directions of the $\vec{E}$ and $\vec{B}$ vectors are always in accordance with expectation and that when combined they give the correct direction of propagation of the energy of

/the.....

the wave. To do the first, he should use the technique of investigating the change of magnetic or electric flux through a narrow rectangular loop, and remember that although the $\vec{E}$ or $\vec{B}$ vectors may point in the same direction along both long sides of the loop, there will be a net EMF (or "circulation of $\vec{B}$") round the loop if the field strength is bigger along the one side than on the other. To do the second, he can make use of the concept of the "Poynting vector", which points in the direction of the energy flow. The direction of this vector can be found from a corkscrew rule: if the corkscrew is turned in the appropriate sense from the $\vec{E}$ to the $\vec{B}$ vector through the smallest possible angle, it will advance in the direction of the Poynting vector.

The ideas developed above in analysing the simple switch-over of a dipole can be applied to a discussion of the "oscillating dipole", i.e. one whose field strength is made to vary sinusoidally with time. It is readily seen that near such a dipole, the variations in strength of the $\vec{E}$ and $\vec{B}$ vectors are out of phase. When the current into the dipole is zero, so that $\vec{B}$ near the dipole is zero, $\vec{E}$ is a maximum; when $\vec{E}$ is decreasing, $\vec{B}$ is increasing; when $\vec{E}$ is zero, $\vec{B}$ is a maximum; and so on. Use of the Poynting vector to investigate the energy flow in this case shows that near the dipole, the energy of the electric and magnetic fields merely surges alternately towards and away from the dipole, never really escaping from its vicinity. On the other hand, in those portions of the field which detach themselves from the dipole, the $\vec{E}$ and $\vec{B}$ vectors are in phase and the Poynting vector is consistently directed away from the dipole, as one might expect. (This is particularly well illustrated in Bitter (1956).)

We are now able to give a more convincing explanation of why the fields can "detach themselves" from the dipole than

/is.....

is possible in terms of "lines of force" alone. If the dipole is "oscillating" rapidly, then the resulting rates of change of electric and magnetic flux will be large enough to "sustain" some of the field created during each half-cycle of oscillation (cf what happens when the dipole field is "flipped over", when the rates of change of flux are so large that virtually the entire field is initially preserved) and so allow it to go on its way through space. The more rapidly the oscillations take place, the bigger will be the rates of change of flux and the more energy will be radiated away. On the other hand, if the oscillations are very slow, scarcely any of the new field created each half-cycle will be "preserved"; the fields simply grow and collapse, and hardly any energy escapes from the vicinity of the dipole.

To complete the background to radiation from a dipole antenna, we mention that the sophisticated theory puts the above ideas on a secure footing. It predicts that the expressions for the $\vec{E}$ and $\vec{B}$ vectors in the fields of a dipole will contain terms corresponding to those one obtains in the "steady" condition, and terms which depend on the inverse <u>first</u> power of the distance from the dipole. The first terms in these expressions are sometimes called the "induction" terms; the other terms, known as the "radiation" terms, represent the portions of the field associated with the waves radiated away from the dipole. As we might expect, as we go further away from the dipole, the radiation terms assume greater importance on account of their inverse first power dependence on the distance. The radiation terms also depend on the square of the frequency, confirming our expectation that the higher the frequency, the more efficiently the dipole will radiate. (Feynman et al (1964) have pointed out the interesting fact that for very small distances, the <u>two</u> terms in the formula for $\vec{B}$ yield a result which is identical

/with.....

with that which would be obtained if electromagnetic effects were propagated instantaneously.)

Teachers should be on their guard against faulty descriptions of dipole radiation which one occasionally finds in textbooks. In a recently-published textbook (Stollberg and Hill 1965), no mention is made of the phase difference between the $\vec{E}$ and $\vec{B}$ vectors close to the antenna; and in the diagram illustrating dipole radiation, the electric and magnetic field lines shown give a Poynting vector (not shown) which is directed towards the antenna!

7.21.5 Electromagnetic waves in the laboratory

It is important that the foregoing treatment of electromagnetic waves, which has been conducted exclusively in terms of the highly abstract concepts of the electric and magnetic field, should be supplemented by suitable demonstrations and experiments. It is not enough to point to the role played by "radio" in everyday life; the pupil needs to study some of the properties of electromagnetic waves at first hand.

First, he needs to be shown that if an electromagnetic disturbance is produced in one part of the laboratory, a corresponding disturbance can be "received" somewhere else in the laboratory. In other words, he needs to carry out Hertz's experiment (cf Magie 1935), or a suitable modern version of it. One such version has been described by Bailey (1967), in which the "transmitter" is a buzzer and battery connected in series with a tuned circuit. A portable transistor radio set, set to a silent point on its tuning dial, can be used as a receiver.

Next, the pupil should see that electromagnetic waves have some of the characteristics which he has observed waves on ripple tanks to possess. The most convenient apparatus for this purpose is the cheap 3cm microwave apparatus now available commercially from a number of sources. There is

/a.....

a danger here that the mystery attaching to these black boxes may distract the pupil's attention from the true purpose of the experiments. It would not be worth the trouble to explain the operation of a klystron transmitter to the pupil. Fortunately, this is not necessary; the pupil has merely to be convinced of the fact that the transmitter is a source of electromagnetic waves, and this can be done by showing that when it is switched on, electromagnetic disturbances can be detected in its vicinity. This focuses attention on the detector, whose "mystery" has now to be elucidated. For this purpose, we suggest the following experimental sequence.

(i) Set up a simple triode VHF oscillator/transmitter (cf Chaundy 1964), and show that this produces a "response" in a detector consisting of a dipole antenna of appropriate dimensions (1m overall length, if the frequency of the signal is 150 MHz) with an indicator bulb at its centre. There is little or no mystery about this detector: the lamp lights up when there is a current in it. The fact that this detector does respond to the V.H.F. set shows that the latter is a source of electromagnetic disturbance.

(ii) Replace the indicator bulb by a crystal diode with a milliammeter connected across it. The rectifying properties of the diode can be readily established with the aid of a battery and a milliammeter, and the functioning of the diode-ammeter combination as a replacement of the indicator bulb can be simply explained. The arrangement can now be used to detect signals from the V.H.F. transmitter.

It can now be pointed out that the detector in the microwave set is also a combination of diode and ammeter. This prepares the pupil for the conclusion that when the microwave detector gives a reading, it is indeed receiving electromagnetic waves.

/There.....

There is no need to discuss in detail the elementary experiments in"microwave optics" which the pupil should carry out or see performed, as they are straightforward in principle and fully discussed elsewhere (Chaundy 1964, Bailey 1967, Jardine 1967). We can simply mention here that with this cheap microwave apparatus it is possible to show a variety of wave effects, such as reflection, refraction, polarization, diffraction and interference, with the minimum of delicate adjustment. The wavelength of the microwave radiation can be obtained from measurements on a standing-wave pattern set up between the transmitter and a metal reflector (for this experiment a probe detector must be used, in place of the "box" containing a receiving horn, diode and microammeter). It is unfortunate that direct measurement of the frequency of the microwaves cannot be undertaken by the pupil at this level, as it would otherwise have been possible for him to calculate the velocity of the waves from measured values of the wavelength and frequency. There does not seem to be any method for measuring the velocity of electromagnetic waves in the "radio" range that is suitable for school use. In this connection, however, the teacher can tell his pupils of "radar" or microwave methods of distance measurement (including that based on the "tellurometer", an instrument designed in South Africa) which depend on the adoption of an agreed value for the velocity of electromagnetic waves.

CHAPTER 8: LIGHT AS WAVE MOTION, IN THE SENIOR SECONDARY SCHOOL

8.1 The wave nature of light

As we indicated in our discussion on light in Section 3.19, an important aim in the teaching of light is the promotion of knowledge and understanding of its nature. The extent to which light is studied in the junior secondary school does not make it possible for pupils to draw anything resembling firm conclusions about the nature of light. The emphasis there is on those aspects of light normally covered by the heading "geometrical optics". The course could have been designed so as to make it possible in the junior secondary school to compare "wave" and "particle" predictions for the speed of light in a transparent medium; we have argued in Section 3.19 that this is best done in the senior years. (In Section 8.2 we consider how this evidence against a naive "particle theory" of light can be presented, and to what extent Foucault's experiment on the speeds of light in air and in water can be considered as "crucial"). It is now necessary to allow the pupil to study those aspects of the behaviour of light that seem particularly "wavy" in character, such as interference and diffraction. Historically, it was the quantitative experimental work of Young and Fresnel that led the wave theory of light to be taken seriously once more, after it had been in a state of eclipse during the previous century. It seems appropriate that the "wavy" aspects of light should be brought to the attention of pupils in similar fashion.

It is perhaps only fair to point out that at least one modern authority has seriously put forward a theory of light in general (and of the diffraction of light in particular) according to which there are no waves travelling in space and

/light.....

light is treated "not as a thing but an interaction between things, consisting ..... of rapidly alternating force" (Burniston Brown 1963). In support of this theory its author points out, among other things, that "attention is concentrated on what is demonstrably present, namely the source and the screen, and not a tertium quid - the completely undemonstrable 'wavefront' (which owes its origin to the apparent similarity to sound, where the wavefront is demonstrable)". He also remarks that "various contradictions associated with the diffraction theories of Fresnel, Kirchhoff and Sommerfeld" are avoided in his theory. It may be suspected that Brown's view of electrodynamics is somewhat different from the usual: that this is so is apparent from earlier papers he has published (e.g. Burniston Brown 1956, 1958). While it is true that his theory is better able to account for the facts of diffraction than any naive application of Fresnel-Kirchhoff theory, it should be noted that rigorous diffraction theory in which the wave equation is solved by putting in the boundary conditions, is also able to account for these facts (a simple account of this approach is given by Mendoza (1953).). Thus it cannot be claimed by Brown that his unorthodox view of light is better supported by the facts than the orthodox one. Certainly it does not seem that a case can be made out for teaching Brown's rather than the orthodox view at an elementary level.

It is a feature of the Nuffield and P.S.S.C. courses and of the new Common Basic Syllabuses in South Africa that the pupil is led a considerable distance "up the garden path" by being shown some strong evidence for light being a form of wave motion, after which he is confronted with the photoelectric effect and its implications. This is parallel to the historical development: the wave theory of light had been firmly established and successfully fused with Maxwell's

/theory.....

theory of electromagnetic waves before quantitative investigation of the photoelectric effect had begun to raise doubts about the completeness of the theory. This sequence of presentation would seem to be pedagogically sound, inasmuch as it gives the pupil a chance to get used to one important idea at a time. Interestingly enough, the corresponding sequence of presentation for the nature of electrons has recently been criticized by Wichmann (1967), who is strongly opposed to the traditional method of indoctrinating the pupil at first with the corpuscular description of electrons, and only then giving him the experimental evidence for the "wavy" aspects of electronic behaviour. According to Wichmann, "We should properly first learn the most basic facts," among which he includes the fact that an electron of momentum p has wavelength $\lambda = \frac{h}{p}$ associated with it. Wichmann's suggestion might work for teaching pupils about electrons, but if it does it will surely be partly on account of pupils having previously learnt from their experience of light that Nature is not as simple as one might think. That experience, we suggest, is best gained in the traditional sequence of teaching. If pupils were to have evidence for the corpuscular behaviour of light presented to them in the same breath as the facts of optical interference and diffraction, one suspects that not only would they acquire the belief that Nature is not simple, but also that physics is a thoroughly bewildering and confusing subject.

In what follows, we consider what should be done to establish and reinforce the idea that light is a form of wave motion, and that light waves are to be identified with electromagnetic waves.

8.2 Refraction and the particle and wave models of light

In the P.S.S.C. physics course, pupils meet the laws of

/reflection.....

reflection and refraction and then study a particle model of light similar in its essentials to that proposed by Newton. It is shown that with this model it is possible to predict Snell's law, provided that it is assumed that the light particles travel faster in media of greater optical density. The results of Foucault's experiments on the velocity of light in air and water are quoted, and it is pointed out that "Not only have the experiments failed to show the ratios [of the speeds of light in air and in water] that we expected - they indicate that the ratios are inverted. Our particle model of light seems to fail" (P.S.S.C. 1960, 1965, 1968). In this connection, the P.S.S.C. authors quote the result of Foucault's 1862 determination of the velocity of light in water. It is of interest to note that as early as 1850, Foucault had already established to his own satisfaction that "the velocity of light is less in water than in air" (Foucault 1850, quoted in Taylor 1959).

In the Nuffield course (cf Nuffield Guide III) much the same line of reasoning is followed, except that the detailed quantitative treatment of the law of refraction and its implications for the velocity of light in different media are not given. In the P.S.S.C. course, the presentation of this material precedes the introduction of wave motion, which is introduced essentially as "something else which travels", i.e. as a possible alternative model for light. Since this follows the principle of having the pupil consider one model of light at a time, we feel it to be recommendable, despite the fact that this presentation does not run quite parallel to the historical development: the wave theory of light had gained many followers before Foucault had carried out his supposedly "crucial" experiment which exposed the limitations of Newton's particle model. (cf. Whittaker (1951): "By the genius of Young and Fresnel the wave theory of light was

/established.....

established in a position so strong that henceforth the corpuscular hypothesis was unable to recruit any adherents among the younger men".)

The question arises whether Newton's "wrong" model should be included at all in an elementary study of light. If its inclusion were only a matter of teaching the history of the subject, it would be difficult to defend. But there is a deeper motive: pupils are given an opportunity to experience something of the testing of a plausible theory and to note the grounds on which it is "rejected". We place the word "rejected" in quotation marks because the rejection is not absolute, which is another valuable point to be made in the teaching of this topic. One cannot hope to teach beginners everything about the assessment of a theory by the consideration of only a few examples, but something, nevertheless, can be taught and the present example is conveniently simple for the purpose.

Care needs to be exercised in the interpretation of Foucault's results as far as they affect particle models of light. We note that the P.S.S.C. textbook says "our particle model of light appears to fail," allowing for the possibility that other particle models might yet be found to succeed. Nineteenth-century adherents of the wave theory were not so cautious. There is no doubt that they regarded Foucault's experiment as "crucial" in this respect. As Foucault himself put it, "It should be observed...that the experiment which demonstrates a lesser speed [of light] in water than air is entirely decisive and that it passes a verdict between the two systems that is without appeal" (Foucault 1850, quoted by Burke 1966). With the hindsight now available to us, it is easy to see that this view is erroneous, but as Sparberg (1966) and others have argued, one ought to be suspicious of "crucial" experiments on principle. As far as Foucault's

/experiment.....

experiment is concerned, Sparberg remarks that it "disproved one of the assumptions of Newton's theory, which was not killed by the death of one assumption." In discussing the failure of "our" particle model with pupils, we ought to make clear that failure of a theory to agree with experiment on a particular point does not necessitate the complete rejection of the theory; there is always the possibility that the theory can be modified.

In the present instance, the teaching sequence requires that we do drop the particle model for the time being and turn to the wave model instead. One can provide a motivation for this: the pupil already knows from his ripple-tank studies that waves obey the law of refraction and travel more slowly in the medium in which their direction of travel is bent towards the normal to the interface. Therefore it should seem natural enough to the pupil that a wave model is now being considered as a serious alternative to the particle picture of light. The teacher should take care to warn the pupil against the idea that any model chosen is likely to be completely satisfactory. In this way he can prepare the pupil for the fact later to be disclosed that light does after all have particle-like as well as wave-like properties, and at the same time teach a valuable lesson on a sensible attitude towards theories.

To show how the naive particle model could predict Snell's law, the experiment described on page 21 of the P.S.S.C. Laboratory Guide could be done by pupils or demonstrated by the teacher. In this experiment steel balls are set rolling across a level surface at an oblique angle to an edge of the surface; on crossing the edge, the balls roll down a short slope and then on to a level surface again. Because the balls speed up in a direction perpendicular to the edge of the surface when they roll down the slope, they must change

/direction.....

direction in the process, so that "refraction" occurs. From the fact that the component of the velocity of the ball parallel to the edge of the surface does not change, it follows directly that the ratio of the sine of the "angle of incidence" to the sine of the "angle of refraction" is constant for a given height of slope, and that the steel balls are moving faster when their path is bent towards the normal.

8.3 Wave motion and refraction by lenses

It is desirable for pupils to know that the converging and diverging action of lenses is explicable in terms of a wave model. The ripple tank should be used to illustrate this point. Glass sheets cut in the shape of the cross-sections of converging and diverging lenses can be placed in the tank so that the layer of water above them forms a converging or diverging lens for the incident waves. The change of curvature of the wavefront is plainly visible in these demonstrations. It is neither necessary nor desirable to attempt to justify this behaviour in terms of Huygens's principle (cf. our remarks in Sections 3.19.4 and 6.2.4); the pupil sees that the wavefronts do change curvature so that the waves converge or diverge, and he sees the analogy with the action of lenses.

8.4 Diffraction of light, I

Pupils now know that a simple wave model accounts satisfactorily for the facts of refraction whereas a simple particle model does not. A natural continuation of the argument is to consider why other wave effects are not readily apparent with light whereas they are with sound; in other words, the attempt must be made to meet Newton's objection to a wave model of light. As the pupil has already studied the diffraction of waves in a ripple tank, he knows that if diffraction effects are to be readily apparent, the essential

/criterion.....

criterion to be satisfied is that the dimensions of the diffracting aperture or obstacle must be of the same order as or less than the wavelength of the waves. It should consequently not be difficult for him to accept that if light is correctly described by a wave model, the reasons why its diffraction effects are not ordinarily noticeable could be that the dimensions of the obstacles and apertures normally encountered are very large compared with the wavelength of light. This point should at once be put to the test, preferably by means of a large-scale demonstration of the diffraction of light by apertures and obstacles. The main requirements for such a demonstration have been fully set out in the Nuffield Guide (III and V). The essential point is that the experimental set-up should be as simple as possible. In particular, the arrangements for viewing the diffraction patterns should be simple, so that no confusion can arise in the pupil's mind concerning the possible role in the formation of these patterns of, say, lenses used for viewing the patterns. As suitable obstacles for casting shadows, the Nuffield Guides suggest the use of "familiar things, such as a pin and a needle, and perhaps a hair." For apertures, it is possible to follow the Nuffield recommendation of using a metal plate (or foil) punched with holes of various sizes. If the school laboratory includes a slit of adjustable width, it would be a good idea to use this to show the effect on the pattern of the gradual variation of slit width. The "Cornell Interference and Diffraction Slitfilm Demonstrator", consisting of sets of single and multiple slits of different widths drawn on film and mounted in glass, is available at a very low price (of the order of two Rand) from the National Press, Palo Alto, California. It is eminently suitable for elementary demonstrations and experiments in diffraction and interference, and could certainly be used to advantage in demonstrations of the type considered here.

/For.....

For comment on the use of the laser in this connection, see Section 8.9.

8.5 Two-slit interference

In the demonstrations discussed in the preceding section, pupils are shown that light can be diffracted and can deduce from the nature of the experimental set-up that the wavelength of the light must be very small. At this point in the presentation of the topic, it is not necessary to attempt to explain diffraction. What is more important is that pupils should now study in more detail a phenomenon which will enable them to make a passably accurate measurement of the wavelength of light and which will extend their experience of the wave behaviour of light. In the Nuffield Guide (V) it is mentioned that some teachers might wish to use the diffraction grating in place of Young's slits at this point, on the grounds that "the grating geometry is simpler" and also, one presumes, because it is easier to set up the grating experiment. We are at one with the Nuffield Guide in insisting on the use of Young's slits, as we feel that, whatever may be said about the geometry, the physics of the phenomenon of two-slit interference is somewhat simpler to grasp than that of multiple-slit diffraction.

The standard methods of carrying out two-slit interference experiments at university or A-level (e.g. use of Fresnel's biprism) are too expensive and complicated to be considered for school use. Fortunately, there are a number of simple ways of demonstrating two-slit interference, all of which are essentially just modern versions of Young's original experiment. The version adopted by the P.S.S.C. and described in their Laboratory Guide is the simplest to set up, but it is also the one which presents most difficulties of interpretation for the pupil. The experimenter looks at the

/light.....

light source (a long straight-filament lamp or a long, narrow, brightly illuminated slit) through a pair of slits scratched on a microscopic slide coated with suitable black material. Having "monochromatized" the incident light by placing a piece of coloured cellophane in front of the source, he makes measurements on the fringe pattern he sees by holding a ruler or metre stick just above the source and viewing the fringes against this background. (A "window" scratched on the slide makes possible simultaneous viewing of fringes and ruler.) What partly obscures the interpretation of this experiment for the pupil (scholar, that is!) is the fact that the interference fringes are formed only on the retina. The pupil has to call on his understanding of geometrical optics in order to appreciate that the relation between his measurements and the wavelength is correctly given by the formula for two-slit interference. He has to realize that the apparent fringe pattern he sees in the plane of the source is the same as would be produced there by a pair of actual (coherent) sources in the position of the two slits. One suspects that this is the kind of fact the pupil might be inclined to accept without really understanding it. Perhaps it is for this reason that this version of "Young's slits" interference is not discussed at all in the Nuffield Guides. The Nuffield proposals for demonstrating interference require the use of a similar experimental arrangement to that recommended for demonstrating diffraction (Nuffield Guide III, V). A pair of slits is illuminated by a very bright source, either directly (in which case the source should be a "line" source), with the source itself functioning as the first "slit" in Young's original arrangement, or through a lens which has been set up to focus an image of the source roughly in the position of the viewing screen. If the latter scheme is adopted, the

/double.....

double slit is placed immediately in front of the lens. The viewing screen is of the same type as that recommended for the diffraction demonstrations.

Several ways of making a double slit are discussed in the Nuffield Guides. Nearly all of them require ruling the slits with a suitable instrument on a microscope slide that has been coated black. Of these, they recommend ruling with a ball-point pen as the best. The present author's experience of making double slits is limited, but he can confirm that the method recommended by the P.S.S.C. and noted by the Nuffield Guides, viz. ruling the slits with a pair of thin razor blades held next to each other does produce satisfactory results - at any rate when these slits are used in conjunction with a laser. As is familiar from diffraction theory and is discussed in detail in the Nuffield Guide (V), the appearance of the two-slit interference pattern will depend not only on the separation of the slits but also on the relation between slit width and slit separation. Using a pair of double-edged razor blades of standard thickness to rule the slits results in a slit width and separation which can give rise to a pattern of several dozen clearly distinguishable fringes when the arrangement is illuminated by a laser. If a bright source of the type recommended by the Nuffield Guide is used, this too should be able to produce a satisfactory fringe pattern with such a pair of slits.

For preference, pupils should make their own pair of slits and take their own measurements on the distant screen when a fringe pattern is obtained with "monochromatized" light from the source. The separation of the slits can be deduced from the thickness of one of the razor blades as measured with a micrometer-screw gauge. The result the pupil obtains for the wavelength of the light he has been using will be accurate enough for him to know that the wave-

/length.....

length of light is of the order of hundreds rather than tens or thousands of nanometres. Thus he does not have to take this important fact on trust from anybody; when more accurate values of the wavelength of the light from particular sources are quoted in the literature or by the teacher, he will know from his personal experience that the order of magnitude of these values is correct.

Certain questions remain to be answered about the underlying principle of the experiment. First, how much attention should be directed to the coherence of the sources at this stage? The Nuffield Guide (V) suggests that this should not be emphasized - "probably not even mentioned unless a pupil asks - because we are discussing a basic demonstration of wave behaviour rather than examining the more advanced details". Perhaps this is too modest a view of what senior pupils can cope with in connection with interference. The pupil knows from his ripple tank investigations that there must be a constant phase difference between sources if there is to be a stable interference pattern. He can note the experimental fact that two independent light sources illuminating different slits will not produce an interference pattern, and infer from this that two independent light sources do not have a constant phase difference between them. Why conventional light sources behave in this way need not be discussed at this point. (Teachers might care to note that even if the two independent light sources are lasers, they will not be able to maintain a constant phase difference long enough for a stable interference pattern to be observed: cf P.S.S.C. 1965, 1968).

The second question to be answered in connection with the principle of the "Young's-slits" experiment is why one is justified in treating the slits as the effective sources of the light. The usual explanation given in elementary

/textbooks.....

textbooks relies on Huygens's Principle, although this fact is not always made explicit. Cf. Sears and Zemansky (1964): "According to Huygens's Principle, cylindrical wavelets spread out from slit S_0 and reach slits S_1 and S_2 at the same instant. A train of Huygens wavelets diverges from both of these slits, which therefore act as coherent sources." The idea that "cylindrical wavelets" spread out from source S_0 is itself erroneous, presupposing as it does that the phase of the disturbance is constant along S_0, but apart from this we have the doubtful status of Huygens's Principle itself to contend with. It would probably be ill-advised to have pupils think too much about this problem. The safest procedure is to point to the direct analogy between Young's experiment and a corresponding experiment carried out in the ripple tank. Pupils should be shown that waves diverging from a "point" source in the ripple tank and intercepted by a barrier with two holes in it, give rise to an interference pattern on the far side of the barrier that is just like that for two point sources. Then it can be suggested to them that <u>if</u> light is a kind of wave motion, it might behave in the same way as water waves are observed to do as far as two-slit interference is concerned. We do not think that this particular problem need be gone into more deeply than this. However, we do think that at least as much needs to be said about it as has been indicated above. It is not enough to pass from "two-source" to "two-slit" interference without some sort of comment.

8.6 Interference in thin films

This topic is briefly treated in the P.S.S.C. course and not mentioned at all in the Nuffield course; it is an "optional extension" to the South African Common Basic Syllabus for Physical Science. It is tempting to include some mention of it in an elementary syllabus - the colours shown

/by.....

by thin films are a very familiar part of everyone's experience, and the phenomena associated with thin-film interference are strikingly beautiful, have many interesting applications and are further impressive evidence for the "wavy" aspects of the behaviour of light. But as anyone with experience of teaching this topic at university will probably confirm, many difficulties arise when one tries to teach it properly, of which the derivation of the formula for the path difference between successive reflected or transmitted "rays" is by far the least. One has to explain why broad sources of light have to be used, in contrast to the use of narrow sources in two-slit interference; why a soap film appears dark by reflected light if it is very thin; why "wedge fringes" in thin films appear to be located in the vicinity of the film, while "parallel-plate" fringes are localized at "infinity"; why the latter fringes must be circular; why the films must be thin if fringes are to be seen and the light is not too well collimated; and so on. All this needs an appreciably higher level of understanding from the student than does two-slit interference. We do not think that consideration of these subtleties should or need form part of any treatment of the topic in a matriculation syllabus. Pupils should see demonstrations of the phenomenon, e.g. by viewing the light reflected from a pair of perspex sheets irradiated by a monochromatic source. The "wedge fringes" produced in the air film between the sheets are generally very obvious to the naked eye under these conditions. By way of explanation of these fringes, all that pupils need know is that interference is possible between light coming from the upper and lower surfaces of a thin film because these "sources" have a constant phase difference between them. The qualitative relationship between the spacing of the fringes and the angle between the surfaces of the film should be pointed out. Pupils wishing

/to.....

to know why the films have to be thin can be shown with the aid of ray diagrams that if the film is too thick, pairs of coherent rays cannot simultaneously enter the eye lens.

A very spectacular form of thin-film interference can be obtained with a very thin parallel-sided film illuminated by light from a very bright, small source. The resulting fringes are "non-localized fringes of equal inclination", to use Tolansky's (1949, 1954) terminology. This demonstration was popularized by Pohl, so that the fringes are sometimes known as "Pohl's fringes" (cf. Sears and Zemansky 1964). The simplicity of the arrangement is a recommendation for its use in elementary classes as an additional means of bringing home the wave character of light.

8.7 Colour, wavelength and frequency

In the context of interference, it is natural for the connection between colour and wavelength to be noticed by pupils. It is obvious even from their relatively crude measurements that the wavelength of red light is greater than that of blue light, so that the pupil can make the correlations: red light → longer wavelength, blue light → shorter wavelength. Following the usual approach, the teacher would then elaborate these correlations into a more detailed statement of the relationship between colour and wavelength, perhaps giving a table of wavelength bands corresponding to the different spectral colours. And there the matter would be left. In our opinion, it ought to be taken a little further. The wavelengths with which it is usual to correlate the colours of the spectrum are those in air or in vacuo. This fact reminds one that the wavelengths of light inside the vitreous humor (the "medium" for light in the neighbourhood of the retina) must be shorter by the appropriate factor, viz. the reciprocal of the refractive index, than those usually quoted. But then one is led to reflect that it

/would.....

would be odd indeed if the sensation of light were to depend on the wavelength of the incident light; how would the receptors in the retina "know" what this wavelength was? On the other hand, it would be entirely reasonable to expect the response of these receptors to depend on the frequency of the incident light, which is an unchanging feature of the light as it passes from medium to medium (cf Coffman 1950) (and which determines the energy of the photons reaching the retina). Hence it is better to correlate the colour of light with its frequency than its wavelength. The teacher should point this out, explaining that although it is customary to quote wavelengths in vacuo when indicating the physical stimulus corresponding to colour, this is essentially a matter of habit. At present it is not possible to measure the frequency of visible light directly, whereas its wavelength can be measured to considerable accuracy; hence it is an ingrained habit to refer to wavelength rather than frequency when referring to visible light.

8.7 Diffraction of light, II

There are certain aspects of the diffraction of light which the pupil should study besides those he has already encountered. First, he should note the dependence on wavelength of the optical diffraction pattern for a single slit of fixed width; observation of single-slit diffraction patterns for light of different colours would do for this purpose. If it is available, he should also be shown the excellent film-loop (Ealing A20-2066/1) which shows the effect on the diffraction pattern of continuous variation of wavelength from one end of the visible spectrum to the other. Time should not be spent in the classroom on the derivation of the well-known formula for single-slit Fraunhofer diffraction. The standard simple derivation depends on Fresnel's modification

/of....

of Huygens's construction, the use of which, as we have seen, is to be regarded as justified by the success of its predictions in particular conditions rather than by any sound underlying physical principle. The direct dependence on wavelength and the inverse dependence on slit width of the separation of the diffraction bands can be treated as an empirical result, deducible from measurements on the ripple tank analogue of the optical diffraction pattern. If one wants to go a little further, one can mention that the light emerging from the slit behaves to a good approximation as if different parts of the slit acted as coherent sources of waves. Then the diffraction pattern can be explained as arising from the constructive and destructive interference of these waves.

Pupils should have at least a nodding acquaintance with the diffraction grating. Cheap diffraction gratings are now available and greatly facilitate the study of spectra in class. Introduction to the conventional optical diffraction grating makes it possible for the pupil to meet the principle applied in three-dimensional X-ray diffraction studies to deduce the spacing of atoms in solids, and also to understand the experiments which show diffraction of "matter waves". The Nuffield Guide (V) comments: "Some pupils will take the grating as an extended form of Young's slits, and the pattern of spectra as a brighter form of Young's fringes." This is perhaps the most convenient way to introduce the grating in class. The pupil knows from his ripple tank and optical study of two-slit interference that two slits in a barrier or screen can act as effective sources of waves, so that the idea of a multiple array of slits acting as a set of multiple sources should not be too difficult for him to accept. It should be reinforced by a demonstration in a ripple tank in which a grating is simulated by an arrangement of multiple slits and, if it is available, by the P.S.S.C. film loop on

/multiple.....

multiple-slit diffraction (Ealing A80-2439/1). If the pupil has understood the geometry of Young's two-slit experiment, he should appreciate that at points far enough away from the grating, constructive interference from the wave contributions from pairs of adjacent "slits" should occur at particular angles of emergence θ from the grating given by $\sin\theta = \frac{n\lambda}{d}$, where the symbols have their usual meanings.

It is desirable for pupils to carry out measurements of wavelength with the aid of the diffraction grating. (A description of a class experiment of this type is given in the Nuffield Guide V). The grating spacing must of course be given to pupils. As the Nuffield Guide (V) points out, the maker of the grating deduces the spacing from the number of rulings made on the grating. Pupils could be invited to suggest how, in principle, measurements obtained with Young's slits and with the grating on the same kind of light could be used to check the maker's value.

8.9 Use of the laser in demonstrating interference and diffraction

Gas lasers suitable for use in elementary teaching are expensive, but not prohibitively so; the current South African price of such an instrument is between two and three hundred Rand. It is thus becoming feasible for a school to consider buying at least one laser for its laboratory, and teachers will be wondering whether it is sensible to buy one and how they should use it if they do.

We consider the second of these questions first. The commercial gas laser for educational use is a demonstrator's dream, requiring only to be switched on in order to be fully ready for use. To demonstrate interference and diffraction

/effects.....

effects with it, all that is necessary is to place the laser behind the single slit, double slit or multiple slit whose interference or diffraction pattern is to be shown - the light will do the rest. This makes it very tempting to use the laser in the initial demonstrations of these phenomena. Of the possible objections to this, we may dismiss at once any which are based on the fact that the laser is a "black box". Any light source is to some extent a black box, as far as pupils are concerned; even the simple incandescent lamp comes into this category, since the processes by which it radiates light will be far from fully understood by beginners. But the laser is a source of highly coherent light, so that it is "special" in a way conventional light sources are not. Even this can hardly be counted an objection to its use. The fact that the light from the laser is coherent introduces no new optical principles into the discussion; let us frankly admit that if certain more conventional light sources were coherent, we should be only too glad to use them in demonstrating interference and diffraction. A more serious objection to the use of the laser is the simple fact of its novelty; it is likely to attract what may be an undesirable amount of attention to itself on account of the glamour which at present surrounds it. There is a real danger that pupils may think that interference and diffraction phenomena are connected essentially with the laser, which is not the impression one wants them to receive. So a wiser strategy might be to carry out the initial demonstrations and experiments in this connection with conventional light sources, as described above, and to postpone use of the laser until pupils have appreciated the essentials of the phenomena under study. Then a laser can be used both to add a touch of glamour and excitement to the proceedings and to reinforce and extend pupils' knowledge and understanding of these phenomena.

/It.....

It could be used for a comparative study of interference patterns obtained with double slits of varying separations; for demonstration of what is meant by coherent light; for an exceedingly effective demonstration of the dependence of the spacing of single-slit diffraction fringes on slit width (if a slit of continuously variable width is available); for demonstration of the formation of diffraction patterns from the rulings on a metal ruler; and so on. Altogether, the laser would certainly be more than merely a valuable toy if it were part of the equipment of the school laboratory, and, other things being equal, one could recommend its purchase. However, if money were short, it is doubtful whether one could give it the highest priority; several oscilloscopes, say, might be of more value to a school laboratory than a single laser.

In the P.S.S.C. physics textbook, second edition (1965) and in the P.S.S.C. "College Physics" (1968) the following remarks appear: "Physicists hope that one day it will be possible to make laser sources of identical wavelengths, fully locked in phase. Then interference effects with light will be as easy to see and understand as those with water are in the ripple tank." When this happens it may be that lasers will have become much less expensive than they are now, and also that their glamour will have faded. In that case it might be wise to amend the above recommendations and insist that all school laboratories be equipped with at least one pair of gas lasers of "identidal wavelength". (Perhaps, in view of what has been said in Section 8.7, it would be better to say of such lasers that they have identical frequency.)

8.10 Polarization

This is a topic which is obviously felt by the planners of both the Nuffield and the P.S.S.C. syllabuses to be "non-essential", since it is mentioned in neither course. Indeed,

/it.....

it is not even referred to in the "College Physics" version of the P.S.S.C. course. On the other hand it has found its way into the Common Basic Syllabus for Physical Science in South Africa. This is reasonable in a syllabus which does include explicit treatment of electromagnetic waves. In discussing the latter, the teacher will have had to explain that the "vibrations" of the electric and magnetic field vectors in a plane electromagnetic wave take place at right angles to each other and to the direction of propagation of the wave. According to the outline given in the Common Basic Syllabuses, one should introduce polarization of waves by pointing out that confinement to one plane of the vibrations of the particles of the medium indicates "polarization" of the wave by definition. With this background, pupils should be able to appreciate that plane electromagnetic waves of the type hitherto discussed with them are, in fact, polarized. If one can now show that light exhibits polarization in addition to its other properties, i.e. that the intensity of the right kind of light beam on passing through the right kind of analyzer depends on the orientation of the analyzer, then this is

(a) valuable supporting evidence for the wave character of light,

(b) evidence for the transverse character of light waves,

(c) further evidence of the similarity between light and electromagnetic waves. The invention of "Polaroid" has made it an easy matter to carry out this experiment at elementary level; one literally needs no more than two pairs of "Polaroid" sunglasses and a source of light, although it would doubtless be more elegant to use mounted plane sheets of polaroid.

It is worth examining what this simple extinction experiment proves, and how it proves it. It is found that the

/transmission.....

transmission is a maximum for two orientations of the analyzer, these orientations being at 180° to one another; is reduced practically to zero at two other orientations of the analyzer which make angles of 90° with respect to those for maximum transmission; and has intermediate values between these orientations. This is precisely what one expects if the "vibrations" in the beam emerging from the "polarizer" are confined to one plane only and if the analyzer transmits only the component of these vibrations in a certain direction. To this extent it supports the view that the waves incident on the analyzer must be polarized. However, as Towne (1967) has stressed, the experiment does not rule out the possibility that these waves might be a mixture of longitudinal and transverse forms of waves. Longitudinal components of the incident waves might be absorbed by the polarizer and so not reach the analyzer; or longitudinal components might be transmitted by both polarizer and analyzer but be incapable of detection by the eye. In Towne's words, "All we can say is that we are successfully able to interpret the results of certain experiments without making use of a longitudinal component." It would help to sharpen the critical faculties of pupils if they were reminded of such points as these.

A more complete argument concerning the similarity between the observed behaviour of light waves and the expected behaviour of electromagnetic waves would take into account the detailed dependence of the intensity of the transmitted light from the analyzer on the orientation of the latter. The expected variation of intensity with angle is given by Malus's Law ($I = I_0 \cos^2\theta$); ideally, one should check whether the intensity of the transmitted light is described by this law. As a first approximation to this, we may accept the results of the simple two-polaroid experiment as being suggestive enough.

/A.....

A word needs to be said about the action of "Polaroid" itself (cf. Shurcliff 1964, Towne 1967, Crawford 1968.). "Type H" polaroid, now commonly used, consists of a stretched sheet of some suitable plastic material (usually polyvinyl alcohol) that has been impregnated with iodine after stretching. The polymeric molecules in the sheet are aligned by the stretching, and the iodine atoms that attach themselves to loose molecules in the process of impregnation thus form long, thin "chains" within the sheet. The iodine atoms "[provide] conduction electrons that can move along the chains but not perpendicular to them" (Crawford 1968). Therefore a sheet of "Type H" polaroid resembles a grid of metal wires in that it has a preferential direction of conduction and, like the metal grid, will transmit electromagnetic waves best if the electric field vector is at right angles to this preferential direction of conduction. If the teacher tells his pupils about the internal constitution of "Polaroid", he should be careful to make this point, lest the impression be gained that the chains of atoms in the polaroid sheet function as a kind of picket fence, permitting only the passage of waves which have their "vibrations" aligned parallel to the atomic chains (cf. Shurcliff 1964). If pupils have been shown only the analogy of transverse waves on strings incident on suitably aligned slits, they might well be inclined to gain this impression. On the other hand, if pupils have been shown the microwave analogue of the action of polaroid "H" sheets, they will be prepared for the correct view.

There is no need in this essentially brief treatment of the polarization of light to introduce any special new terms, such as "plane of polarization". This term, for which conflicting meanings have been given in the literature, seems now to be falling out of use, to be replaced by the use of the terms "the plane of the $\vec{E}$-vector" and "the plane of the

/B-vector."

$\vec{B}$-vector". At this point in the treatment, where the identification of light waves with electromagnetic waves has not yet been made, it is not yet appropriate to use this terminology; one should talk simply of "the plane of the vibrations" [of the $\vec{E}$-vector, that is to say] for the moment.

8.11 Intensity, illumination and the inverse square law

Up to this point the pupil will have been using the term "intensity" in connection with light by simple analogy with its use in connection with wave motion in general. The assumption tacitly made here is that light <u>does</u> transport energy. This assumption should be brought into the open, and pupils asked to suggest evidence in favour of it. Sophisticated pupils will probably recall the laser and its ability to "punch" holes in razor blades, which clearly requires energy; some may think of a light meter for a camera and the fact that it responds to the incident light by giving a reading; others will think of sunlight and its obvious heating effect, or of ultra-violet or ordinary incandescent lamps. Perhaps in the case of the last-mentioned examples pupils will accept that all incident radiation, visible or otherwise, contributes in some measure to the heating that is produced.

On the assumption that energy is transported by light waves, it is easy to show that the intensity of the light must fall off inversely as the square of the distance if the source is small and there are no complicating factors. The P.S.S.C. textbook (1960,1965) suggests an experiment with a light-meter to confirm this, though the experiment described seems more of a "thought" experiment than one which could conveniently be carried out with a class. The light-meter is calibrated by setting up one, two, three...similar "standard light sources" at a standard distance from the meter, noting the reading in each case and marking the scale as registering one, two, three...units of illumination at these points. The meter is then used to measure the illumination at known distances

/from.....

from one of the standard sources. We do not think it essential to deal with this experiment in the classroom; pupils should know that the inverse-square law <u>can</u> be verified, but the means whereby this can be accomplished can be left as a reading assignment.

The "illumination" used in the above experiment was not given formal definition. How exactly should one define it for the pupil's benefit? To define it rigorously as "luminous flux per unit area" raises too many difficulties which ought not at all to be the concern of the pupil at this stage. "Luminous flux" is a quantity which depends partly on the response of the eye to incident light, and whose unit is defined in terms of the luminous flux from a standard source into unit solid angle. It is not surprising that in recent years a number of introductory textbooks intended for use at first-year university level have omitted all mention of this aspect of the subject (cf Sears and Zemansky 1964, Halliday and Resnick 1966). We suggest as the answer to the above question that since under given conditions "illumination" is proportional to "intensity", the readings on the light-meter in the above experiment be taken as a measure of the latter quantity. The calibration procedure described in the previous paragraph must then be regarded as establishing a scale of <u>intensity</u> for the meter.

8.12 <u>Identification of light with electromagnetic waves</u>

There are a number of points that can be made by way of convincing pupils that "light" (meaning, for the moment, infra-red, visible and ultra-violet light in the usual sense of the term) is to be identified with electromagnetic waves. Obviously the most satisfactory way of establishing this identification is to provide direct evidence of the electric and magnetic fields associated with light, say with some kind of

/antenna.....

antenna capable of responding directly to these fields in a light beam. But there are obvious difficulties associated with trying to set up such an arrangement; the extremely high frequency and short wavelength of the radiation make direct observation of the fields impossible, because we cannot set up field-detecting equipment capable of following the extremely rapid variations in the fields, or of sufficiently small size to be able to record the fields over a region small enough for the spatial variation in field strength to be negligible. The atoms or molecules of matter in bulk have to serve as these antennas, and much has now been done in the way of checking theoretical predictions concerning the interaction between light and matter; with the advent of the very strong electric fields made available by highly intense laser beams, work in this direction has received added stimulus. In all cases, the experimental results are in accord with the theoretical predictions based on the (partly) electromagnetic character of the incident radiation. This is impressive evidence, as near direct as can be arranged, of the electromagnetic character of light, but it is not of a kind that will be of much help to the teacher in the present situation. He will be able to make his point best by direct comparison between the observed behaviour of waves that are obviously electromagnetic in character and that of light.

To introduce this comparison, he could start by reminding pupils of the predictions of Maxwell's theory of electromagnetic waves. The relevant predictions are the following:-

(a) Electromagnetic waves exist;

(b) they do not need a material medium;

(c) they are transverse;

(d) they show all the usual characteristics of waves, i.e. they can be reflected and refracted and obey the principle of

/superposition.....

superposition in appropriate circumstances;

(e) they travel through free space with a speed of 3×10^8 m/s, very nearly;

(f) they exert "radiation pressure".

Preferably, all this material (with the possible exception of (f)) should have been studied earlier by the pupil, perhaps with the aid of simple microwave equipment as far as (a), (c) and (d) are concerned (cf Section 7.21 5). Regarding (b), it is probably easiest to stress this point with reference to the now well-known "space-age" fact that radio signals have been successfully transmitted over many millions of kilometres of largely empty space. Regarding (e), the determination of the speed of waves of obviously electromagnetic origin by various techniques can be used as the basis for the comparison between the behaviour of light and electromagnetic waves in this respect. In other words, the teacher does not have to depend only on Maxwell's calculation of the speed of light in vacuo in order to establish this point. Ideally, he might perhaps wish to be able (i) to compare Maxwell's prediction with the observed speed of electromagnetic waves and (ii) to compare the measured speeds of electromagnetic waves and light; but to rely essentially on (ii) while simply asserting the facts of (i) would be quite adequate for present purposes. The main point to make is, after all, that light travels as fast as electromagnetic waves are known to do.

Concerning (f), it might at first be thought surprising that this should have been included, as it is rather more subtle than the others and the teacher may feel it is less important. However, the prediction that electromagnetic waves must exert a "radiation pressure" is interesting in more ways than one. It is of immediate relevance because it has been shown experimentally that light exerts a radiation pressure equal to that predicted by Maxwell for electromagnetic

/waves.

waves. Also, pupils should have heard of radiation pressure by the time they come to study the photon character of light, since the fact that light exerts such a pressure is consistent with the idea that photons transport momentum. In any event, it is not too difficult to explain why incident electromagnetic radiation should exert a force on a conducting sheet placed in its path. If such a sheet is placed perpendicular to the incident radiation, a conduction electron in the surface would acquire a drift velocity in the plane of the sheet under the action of the electric field of the wave. It would then experience a "magnetic force" in the direction of travel of the wave: the combined effect of such forces on the "free" electrons in the metal would constitute a pressure on the sheet. (For an account of how the forces on the electrons are transmitted to the sheet as a whole, see Section 7.15.7.) Detailed calculation shows that the "radiation pressure" exerted by a polarized beam of electromagnetic radiation is simply equal to the energy density of the radiation (cf Panofsky and Phillips 1962). This relation was verified for light waves by Nichols and Hull in 1903 (cf the account given by Halliday and Resnick (1966)) and by Lebedev at about the same time. Pollock (1963) has devised a beautifully simple method of measuring the radiation pressure of a light beam; a mirror suspended by a torsion fibre in air is driven at resonance by the force derived from the radiation pressure in a light beam which is interrupted at the appropriate frequency, and from appropriate measurements on this system the radiation pressure can be inferred. Unfortunately, the comparatively sophisticated theory of damped harmonic motion required to relate the measurements to the result puts the experiment beyond the reach of a South African school syllabus. The teacher will have to fall back on simple assertion and tell pupils that the predicted value of radiation pressure has been confirmed experimentally.

/It.....

(It is of some interest to note that what is sometimes called the Langevin radiation pressure of an acoustic wave (cf Beyer 1950) is also equal to the energy density of the radiation. However, the origin of the radiation pressure in acoustic waves is quite different from that in electromagnetic waves. In the latter case, the radiation pressure arises from the fact that there is a net transport of momentum in the forward direction of the wave, which in turn arises from the fact that the electric and magnetic field vectors of the wave are perpendicular to each other and to the direction of the wave. In the case of acoustic waves, the radiation pressure arises from the fact that the compressions occupy appreciably smaller volumes than the rarefactions, so that the average value of the pressure on the walls of a tube inside which sound waves are travelling is less than it would have been in the absence of a sound. We are thus not dealing here with a result which may be thought to be generally applicable to all forms of wave. Beyer (op.cit.) gives a full discussion of the radiation pressure exerted by sound waves.)

In sum, the case for the identification of "light" with electromagnetic waves is to be made to rest largely on the fact that, quantitatively as well as qualitatively, light seems to behave in essentially the same way as waves of obviously electromagnetic character. It was evidence of this kind which convinced Hertz, eighty years ago of the correctness of this identification: as he then put it, (see Magie 1935), "The experiments described appear to me, at any rate, eminently adapted to remove any doubt as to the identity of light, radiant heat and wave motion." One may confidently hope that, clearly presented, such evidence will prove similarly convincing to present-day pupils.

/8.13.....

8.13 Dispersion

The dispersion of light is worth some attention in a school syllabus, partly for its own sake (as the "Comments" on the Draft Common Basic Syllabus remark, "dispersion is fundamental physics") and partly because it underlies an elegantly simple method of obtaining spectra of light sources. The principal point to establish here is that the speed of light in a material medium depends on its frequency. This can be done in the following sequence:-

(i) Show that white light passed through a prism gives a pure spectrum.

(ii) Remind pupils of the association between the colour of light and its frequency (or wavelength in vacuo).

(iii) Show, without deriving any formulae, that the deviation produced by a prism depends on its refractive index.

(iv) Draw the conclusion from (i) - (iii) that the refractive index of the prism depends on the frequency of the light passing through it.

(v) Draw the conclusions from the relation between refractive index and wave velocity that (iv) implies that the speed of light waves in a transparent medium depends on their frequency. Compare this with the behaviour of surface waves on water in ripple tank.

This is as far as any syllabus for South African Matriculation requirements should take this topic. Brighter pupils might be interested in certain further implications of the above discussion, such as the concept of group velocity. However, as Mendoza (1965) observes, "... the group velocity finds no immediate application in school physics", so that we see no particular point in discussing it here.

/8.14.....

8.14 The electromagnetic spectrum

It is fitting to follow a discussion of the identification of electromagnetic waves and light with a discussion of the electromagnetic spectrum as a whole. The difficulties of teaching this topic as one would like to are well summed up in the Nuffield Guide (V), where it is remarked that "The things that we can demonstrate make a very patchy story" and "In practice, our teaching here must be largely by assertion, with some pieces of description to support it...We should give a quick survey and post up a large chart for the rest of the year".

Among "The things that we can demonstrate" are the existence of infra-red and ultra-violet electromagnetic radiation; details of suitable demonstrations are given in the Nuffield Guide (V). With regard to the longer-wavelength-than-infra-red or shorter-wavelength-than-ultra-violet sections of the spectrum, there would be little time to give pupils any detailed justification of assertions made concerning wavelength or frequency. On the longer wavelength side, use could be made of the microwave apparatus (if available) to show that the wavelength characteristic of the equipment was of the order of a few centimetres (3 cm, typically); for the rest of this section of the spectrum, pupils could be invited to look at the wavelengths and frequencies stated on the dial of a radio set. On the shorter-wavelength-than-ultraviolet side of the spectrum, it will have to be all "assertion"; the teacher has merely to take care that what he asserts are facts and not fantasies, such as the old story about cosmic rays being exclusively gamma-rays of very short wavelength.

8.15 Introduction to the study of spectra

The pupil is now ready to learn something of the association between particular kinds of spectra and their sources.

/It.....

It would not be appropriate at this stage to make too much of this - a closer study of some of the features of atomic spectra will have to be undertaken later when the pupil knows something of atomic physics - but it might be useful for the pupil to learn to appreciate that certain types of spectra originate typically from certain kinds of source. First, there is the classification of spectra into continuous, line and band spectra to be noted, accompanied at the very least by photographic illustration of these types, though naturally it would be better for pupils actually to see examples of all three. These spectra should first be studied as emission spectra, with the aid either of a simple grating spectrometer of the type devised by the P.S.S.C. (or P.S.S.C. Laboratory Guide) or with a conventional spectrometer with a prism or grating mounted on its table. (The P.S.S.C. spectrometer has the advantage that it can readily be replicated.) A continuous spectrum can be provided by an incandescent lamp; a neon indicator lamp is a cheap source of a line spectrum, and a conventional hydrogen discharge tube, which can be operated from a small induction coil, gives the line spectrum of atomic hydrogen with the band spectrum of molecular hydrogen as a fairly strong background.

For a reasonably complete presentation of absorption spectra, one should ideally show examples of continuous, line and band absorption spectra. Then nobody could be in any doubt that there really were three kinds of absorption spectra, and there could be no excuse for anybody adopting the ridiculous classification of spectra into the four categories of "continuous, line, band and absorption". (Warren (1965) remarks that "This extraordinary classification has been explicitly demanded by G.C.E. examiners"!) Perhaps, however, it is not necessary to go as far as this. If pupils see Fraunhofer lines in the solar spectrum and a demonstration of

/the.....

the absorption spectrum of sodium vapour, this should make the nature of an absorption spectrum clear to them. The teacher can point out that band and continuous absorption spectra can also be obtained with the appropriate absorbing material.

At this stage, little need be said regarding the sources of these different types of spectra beyond the facts that
(i) any solid, liquid, or gas at high pressure that is not at absolute zero will emit electromagnetic waves over a continuous range of frequencies;
(ii) line and band spectra are emitted by gases at comparatively low pressure.

CHAPTER 9: ELECTRONICS

9.1 Introduction: some general considerations

Following the lead of the "new" syllabus of the South African Joint Matriculation Board, the Common Basic Syllabus for Physical Science includes a "single-square-bracket" section on the fundamentals of electronics; i.e. this section is an "optional extension" to the official syllabus. In this respect, this syllabus differs from those of the P.S.S.C. and the Nuffield Physics Project, neither of which includes any treatment of electronics as such, although the P.S.S.C. course does treat thermionic emission in passing. The planners of the Common Basic Syllabus remark on the lack of "basic physics" in their proposed electronics section, but add that "...there is an almost irresistible demand for some understanding of electronics and this should be seriously considered." We feel that to lend serious consideration to the inclusion of electronics is no more than realistic at present; that even if syllabuses appear to be crowded, efforts should be made to find some room for electronics in them. We have indicated the reasons for this in Section 2.9.4, where we suggested that if electronics were to be included, its teaching should be based as far as possible on underlying physical principles. In this section we shall be principally concerned with how that suggestion might be applied.

Should semiconductor electronics be included in a section on elementary electronics? There are some to whom it will appear absurd that this question should even be put: surely, they will say, in view of the extent to which semiconductor devices have supplanted thermionic ones in everyday electronics, it is obvious that we should include some semiconductor electronics if we are not to give pupils a false view of the subject. But those who wish to proceed more

/cautiously.....

cautiously do have a point in their favour: whereas it is quite easy in an elementary treatment to relate "thermionic electronics" to the underlying physics, doing the same for semiconductor electronics presents problems which might at first be thought insuperable, if electronics is not to come right at the end of the physics course. In our view, the attempt should nevertheless be made, for several reasons. First, semiconductor electronics is "relevant" to the modern world; second, it is much less expensive to equip a laboratory for semiconductor instead of thermionic electronics, as costly power supplies are not needed; third, the use of low voltages obviates the risk of shocks. There is no mention of semiconductor diodes or transistors in the "electronics" square-bracketed section of the Common Basic Syllabus, but it could be claimed that this does not mean that these devices are intended to be excluded. Unfortunately, the writer of the corresponding section of the new Teachers' Guide to this syllabus (Prinsloo et al 1968) has taken a very cautious view of his responsibilities in this connection. The only concession he makes to semiconductor electronics is to say: "Although its [sic] action cannot be explained satisfactorily at this stage, pupils should be made aware of the existence of solid-state diodes. (Many pupils will probably be familiar with the cat-whisker-galena diode used in crystal receivers)". (The parenthetical remark here is surely more appropriate to pupils of, say, twenty years ago than to those of the present day!) The transistor is not mentioned at all. We shall proceed below on the assumption that any acceptable modern syllabus in school electronics must include semiconductor devices, and we shall show that there are ways of solving the problems which this inclusion raises.

One can agree that thermionic emission is still a good starting point for the study of electronics in general, since

/in.....

in this way one does begin by discussing a physical phenomenon and then leading on from it to a discussion of its application in electronics. Semiconductor diodes can be introduced as black boxes which are seen to have the same property of "rectification" as thermionic diodes. Thermionic triodes can be discussed next, again because they are devices the physical principles of which are quite easy to grasp; and after triode amplification has been dealt with, the transistor can be introduced as an amplifying device, and some attempt can be made to explain its action in terms of the behaviour of a combination of semiconductor diodes. If the overall syllabus for the course were so designed as to include a later section dealing with electronic energy levels in solids and other applications of atomic physics, further attention could be given there to discussion of the physical principles of semiconductor diode and transistor action.

9.2 Thermionic emission, rectification and semiconductor diodes

If the method of presentation of electricity advocated in the preceding pages (cf especially Section 7.14.2) has been followed, pupils will have encountered thermionic emission in their study of the electric current in metallic conductors. At that stage, the principal point of interest was the fact that the emitted particles must have a negative charge, which was deduced from the "one-way" behaviour of the current in a thermionic diode. For present purposes, the study of this phenomenon need not be taken much further. Pupils should have some idea of the shape of the voltage-current characteristic of a thermionic diode. It is not vitally necessary for them actually to observe the flattening-off of

/the.....

the characteristic at higher voltages: what is more important is that they should be aware of the non-linearity of the characteristic and of the fact that it is virtually confined to the region of the graph corresponding to positive anode-cathode voltages. To demonstrate these points, any low-voltage diode (e.g. the Philips EA 50 or EB 91) can be used. A rough explanation of the shape of the graph, or at any rate of why the current in the diode should rise at first with increasing anode-cathode potential difference, should be given to pupils. The existence of the space charge, its neutralizing effect in the vicinity of the cathode on the electric field set up by the anode-cathode p.d., and the effect of this applied electric field on the net number of electrons entering and leaving the space charge per unit time are points which could be briefly discussed.

It is fitting that pupils should meet semiconductor diodes, since they have largely supplanted thermionic diodes in conventional diode applications. These devices can be introduced as black boxes whose current-voltage characteristics can be investigated by the pupils themselves. The similarities between these characteristics and those of thermionic diodes should at once be apparent. The use of diodes as rectifiers can immediately follow this comparison, beginning with the simplest possible rectifying circuit, viz. a diode with a load resistor connected in series with its cathode, the two circuit elements being connected across the output from a source of low-voltage A.C. (Gordon and Neser (no date) show as their only example of a rectifying circuit that used to charge a battery; since they make no real attempt to explain how the circuit works, the reader could easily be left with the erroneous impression that a battery is somehow essential for the operation of a rectifying circuit.) The rectifying action of this circuit should be demonstrated with the aid of a cathode-ray oscilloscope, with pupils being asked to account for what they

/see.

see. The point should be made that the diode will behave most nearly like a simple on-and-off switch in this circuit if its forward resistance is negligibly small in comparison with the resistance of the load. Extension of these ideas to include smoothing (with a single capacitor) and full-wave rectification (with four semiconductor diodes forming a bridge rectifier) raises no great problems provided that the treatment remains qualitative.

9.3 Triodes and amplification

The fundamentals of triode action should be demonstrated, preferably with a triode that looks like a triode (i.e. one which has a clearly visible anode, grid and cathode). The reasons for the triode behaving like this should be discussed, with particular reference to the action of the grid. Gordon and Neser (op.cit) give a good example of how not to treat the latter: "If the grid D is given a negative charge, the electrons from C will be repelled so that they do not reach P. But if D is positively charged it will, like P, attract the electrons, many shooting through the openings to the plate." The clear implication of this passage is that the grid has only to be given a negative potential with respect to the cathode for the anode current to be completely cut off. That this is not the case will be obvious to any pupil who sees the action of the triode demonstrated. Rogers (1960) has given an excellent elementary account of the controlling action of the grid: provided that pupils have a qualitative understanding of electric fields and their lines of force, they should have no difficulty in following this.

The discussion of the amplifying action of the triode should be kept simple. This is evidently not the view of the author of the relevant portion of the Teachers' Guide to the Common Basic Syllabus, who says that "Although this section is

/to.....

to be treated quantitatively, illustration of the meaning of the following two quantities [amplifying factor (sic) and transconductance] with the use of typical triode characteristics will lead to a better understanding of amplification." He then proceeds to give verbal equivalents of the approved symbolical definitions of "amplifying factor" and transconductance, and comments that "Knowledge of these quantities should clarify the process of amplification considerably, and can best be illustrated when the changes in voltage and current which are used in the definitions are shown on typical triode characteristics." Such an approach would be admirable for students of electronics at a university or technical college, but is surely much too sophisticated for the average matriculation pupil. The teacher could reasonably complain that the time he would have to spend in clarifying the concepts of amplification factor and transconductance would be much better spent on other, more fundamental topics in the syllabus, especially when the essential points concerning triode amplification can be conveyed to pupils by simple demonstrations. For example, a suitably-chosen load resistor can be connected in series with the anode, and the variation of anode voltage with grid voltage can be demonstrated both with slowly varying voltages, monitored by meters, and with more rapidly varying voltages, monitored by a cathode-ray oscilloscope. The pupil does not really need to know more about amplification that what is conveyed by this demonstration.

9.4 Transistors and amplification

As was suggested in Section 9.1, the transistor can be introduced at this point. Some attempt should be made to explain the action of the transistor in simple terms, otherwise it will not be possible to justify to pupils why it should be

/connected.....

connected up in the way in which it is for measurements of its characteristics or use in an amplifying circuit. Unless the pupil already knows something about electrons and holes in semiconductors, the explanation given to him of transistor action should not make explicit use of the term "hole", except as an unexplained term used to describe the (majority) charge carriers in p-type semiconductors. For a "first-order" understanding of transistors, the pupil needs to know first of all that there are two types of semiconductor used in a transistor: p-type, in which there is a preponderance of positive charge carriers, the nature of which need not yet concern him, and n-type, in which there is a preponderance of electrons. Also he needs to know that a semiconductor junction diode consists of the junction between these two types of semiconductor, and that its cathode is the portion containing n-type semiconductor, i.e. the diode conducts readily when its p-type portion is made positive with respect to its n-type portion. With this background, pupils can be invited to consider the transistor, which can be represented to them as "a combination of two junctions very close together on the same cyrstal" (Osborn 1964). (There is surely no way in which the pupil can be led to "discover for himself" that it would be a good idea for a device of this sort to be constructed. After all, the discovery of the point contact and junction transistors was considered worthy of the award of a Nobel Prize.) That a transistor does in fact consist of such a combination of junctions should be made apparent to pupils by allowing them to inspect the construction of a transistor (e.g. a photo transistor, as suggested by Osborn (op.cit.)). The standard "simple explanation" of transistor action (cf Bennet 1965, 1968) can then be given, followed by appropriate demonstrations or experiments (cf. Osborn op.cit., Jardine 1967) using the transistor connected in the grounded-emitter configuration. At this stage, pupils need not obtain a full-blown set of collector-current

/versus.....

versus collector-voltage characteristics. Instead, as suggested by Jardine (op.cit.), they can plot the transfer characteristics of the transistor, i.e. the graph of collector current versus base current for constant collector voltage.

The next step in the argument is the comparison between the triode and the transistor; the essential point to make here is that whereas in the triode it is the grid-cathode voltage which controls the anode current, in a transistor it is the base-emitter current which controls the collector current. Then the pupil should proceed to the experimental study of transistor amplification: suitable circuits are given by Osborne (op.cit.) and Jardine (op.cit.). Certain details of these circuits, such as the use of a protective resistor to limit the collector current and so prevent "thermal runaway" and the use of a capacitor in parallel with this resistor to minimize "current feedback", ought not to be glossed over, as they are standard features of any transistor amplifying circuits the pupil is likely to use, and he ought to have some idea of their function. Simple qualitative explanations can be given in each case.

9.5 Oscillators, triode and transistor

Electrical oscillations are widely used and in a modern course in school physics will play their part in certain demonstrations. Pupils should therefore have some idea of how these oscillations can be generated and maintained. Fundamental to this section is some qualitative understanding of the behaviour of circuits containing inductance, capacitance and resistance in series; this would ordinarily have been covered in the study of electromagnetism (cf. Sections 7.20.10 and 7.20.12). From that preliminary study, it will have emerged that to maintain oscillations in any practical L-C circuit, energy will continually have to be supplied to it in order to make up for what is lost as "heat". Under the above heading,

/it.....

it should be made clear how this can be achieved "electronically". (Resistance-capacitance oscillators and multivibrators could not conveniently be incorporated in the syllabus, unless it were decided to allot "Electronics" a much more prominent role in the complete syllabus.)

Preferably, electronically-maintained oscillations should first be demonstrated at frequencies low enough to permit the changing voltages or currents to be recorded on meters. Chaundy (1964) gives a circuit of a tuned-plate oscillator using a very large inductance (500H) and capacitance (12-80μF) in the tank circuit. Higher frequency oscillations using a smaller inductance and capacitance in the tank circuit can be studied with the aid of an oscilloscope. There are several other fairly simple oscillator circuits which could be used to demonstrate electronically maintained oscillations, but we suggest that in a course like this it is necessary to show only one, and that this should be shown in both its triode and transistor versions. Jardine (1967) gives a circuit of a transistorized version of the tuned-plate oscillator, using an OC71 transistor and a 9:1 transistor-circuit output transformer to provide the tank coil and the necessary coupling to the emitter circuit.

9.6 The cathode-ray oscilloscope

The cathode-ray oscilloscope may now be said to occupy an established place in elementary physics teaching. Pupils should accordingly spend a little time in making closer acquaintance with the principles of operation of this instrument. By the time they come to study the instrument itself, they should be familiar with the principal features of the cathode-ray tube, viz. its electron gun and deflecting plates. Additional features that would require comment at this point are the grid for control of intensity, and the focusing electrodes. The function of the former presents no real problems

/of.....

of explanation, being similar to that of the grid in a triode. The latter would present difficulties if one tried to go into detail about the action of the focusing anodes. Fortunately this is not necessary. One need simply say that the electric fields inside these anodes are so shaped that they have the desired focusing effect on the electron beam passing through them. It should be mentioned that the walls of the cathode-ray tube are coated with a layer of conducting material which is connected to the second anode. This provides a constant potential enclosure for the electron beam. Also it should be pointed out that the electrons reaching the fluorescent screen form part of a circuit which should be closed in some way. The "return" portion of the circuit is provided by secondary emission of electrons from the screen following on its bombardment by electrons in the incident beam; these secondary electrons drift back to the anode.

Pupils should acquire a "block-diagram" degree of familiarity with the oscilloscope, so that they are aware of the functions fulfilled by the various circuits in the oscilloscope, in particular the horizontal and vertical amplifying circuits, the time-base circuit, the triggering or synchronizing circuits and the beam positioning circuits. Pupils should by this time know enough electronics to understand the principles of operation of amplifying and beam-positioning circuits. The principle of operation of the time-base circuit is worth some attention. First, it should be explained that the "waveform" output of a time-base circuit must be of the form of a voltage rising linearly with time, then falling back to its initial value - a sawtooth waveform. Next, it should be shown that a crude sawtooth waveform (one which is not linear) can be produced by a simple resistance -capacitance circuit in which a neon tube is connected in parallel with the capacitor.

/For.....

For the pupil to understand the operation of this circuit, he needs to know that the neon-tube "strikes" at a higher voltage than that at which it "cuts off". Ideally, the variation of voltage across the capacitor should first be demonstrated by a voltmeter connected across the latter and then by display on the oscilloscope screen, but this presents difficulties: the voltmeter must have a resistance of several megohms if its connection into the circuit is not to upset the operation of the latter (a large resistance and capacitance must be used in the circuit to provide a sufficiently large time constant). If a suitable high resistance voltmeter is not available, the omission of the preliminary demonstration does not matter very much. As we mentioned in Section 2.4, there is nothing really circular in using the oscilloscope to display waveforms which help to elucidate the operation of the oscilloscope. The pupil should have been shown much earlier in the course that the oscilloscope does indeed plot graphs of voltage against time (cf Sections 2.4, 3.16.4), so that he can accept the display of a sawtooth waveform as showing the voltage across the capacitor in a simple sweep circuit. Finally, this crude sweep circuit should be used to provide the time-base for the oscilloscope, to show that it causes the spot to sweep back and forth across the screen as expected, and to show that even with a simple time-base circuit of this type it is possible to obtain stationary patterns on the screen when a repetitive signal is applied to the Y-input terminals, although the pattern will appear "squashed-up" towards one side. This will provide opportunities for discussing the conditions under which stationary patterns can be obtained, and the desirability of having a linear time base.

Modern oscilloscopes, including those which the pupil is likely to use, employ "triggering", rather than synchronization of the repetition frequency of the time base with that of

/the.....

the input signal, in order to secure a stationary display on the screen. "Triggering" circuits are much too sophisticated to be suitable for treatment in a course such as the present one, but their function can be briefly indicated. It can be explained that in "triggered" operation, the sweep circuit only "fires" when it is triggered by the triggering circuit, and that the latter operates in such a way that it only "fires" the sweep circuit when the voltage of the input signal attains a certain set value on the rising part of the voltage curve (positive slope operation) or on the falling part of the curve (negative slope operation).

In the present context it is desirable for the pupil to see how the oscilloscope can be useful in the study of electrical and electronic circuits, both as a measuring instrument and as a portrayer of electrical "waveforms". Many possibilities will suggest themselves to the enterprising teacher. For help in this connection the teacher can turn to an article by Christiansen (1963), in which a number of waveform and other demonstrations of interest are described.

9.7 Modulation, transmission and reception of radio waves

The ubiquity of radio makes it sensible to include a section on it in the "electronics" section of the syllabus. The details of modulation, transmission and reception of radio waves would give rise to many difficulties of presentation at this level, but the general principles on which these are based should be within pupils' comprehension. Elsewhere (Helm 1963) the present writer has outlined a suitable teaching sequence.

CHAPTER 10: PHOTONS, ELECTRONS, ATOMS AND NUCLEI

10. 1 The particle-like behaviour of light

10.1.1 Introduction: teaching objectives

Besides being concerned with the topic that forms the above title, this section of the syllabus serves the important purpose of introducing the pupil to the quantum theory. Traditionally, the latter is often introduced along strictly historical lines, beginning with Planck's treatment of the perplexing data obtained by Lummer and Pringsheim in their experiments on black body radiation. But this is an area very remote from the pupils' experiance. Since the photoelectric effect can easily be made part of the pupils' experience and its quantitative aspects disclose the need for Planck's quantum hypothesis in a comparatively simple way, it is sensible to begin with it instead. In any event, discussion of the photoelectric effect follows naturally enough on a consideration of the wave theory of light.

The main difficulty in teaching this material is the traumatic shock it gives the pupil who, despite wernings he may have received, will have settled down quite happily with a wave picture of light and a particle picture of the electron. We do not think, however, that he can safely be spared this upsetting experience; the quantum theory has been with us for a very long time and we cannot now expect to ignore its implications even in an elementary physics course. If this leaves the pupil with the impression that nature is not so simple, after all, this will be all to the good. Certainly the pupil will worry about how the "wave" and "particle" aspects of photons or electrons can be reconciled, and equally certainly the teacher will have to guide him towards a solution of this problem. Only the most optimistic of teachers will believe that he can actually solve this problem for his

/pupils,.....

pupils, i.e. leave them all with the feeling that they "understand" the nature of photons and electrons; the subject is too new and bewildering for that. In our view, realistic teaching objectives in presenting this material would be the following:

(i) To give pupils a secure knowledge of the relevant experimental facts and the hypotheses introduced to explain them.

(ii) To make clear the circumstances in which it is appropriate to use the "wave" or "particle" aspects of photons or electrons in describing a given situation.

(iii) To give the pupils some insight into the "probability" description of photon and electron behaviour.

(iv) To make clear the differences as well as the similarities between photons and electrons.

(v) (An important negative objective) Not to mislead pupils by means of dogmatic statements concerning the nature of photons and electrons.

10.1.2 Introductory demonstrations

A simple and well-known demonstration of the photoelectric effect makes use of an electroscope, an insulated zinc plate whose surface has been carefully cleaned, and an ultraviolet or mercury vapour lamp (cf Lewis 1964, Nuffield Guide V). The zinc plate is connected to the electroscope, and the lamp is shone on to the plate after the electroscope has been (i) positively and (ii) negatively charged. No effect is observed when the electroscope is positively charged, but the electroscope is found to discharge if it has been negatively charged. This shows that the radiation from the lamp causes the plate to lose its excess negative charge. Again with the electroscope positively charged one can show that bringing the lamp closer to the plate produces no effect, whereas this operation results in an increased rate of

/discharge.....

discharge if the electroscope has been negatively charged. Therefore the greater the intensity of the incident radiation the more rapid is the loss of negative charge from the plate. (Robinson (1965) has described a variation of this demonstration in which the electroscope is not given an initial charge, but instead positively- and negatively-charged objects are held near the zinc plate while the latter is irradiated. If a positively-charged object is held nearby, the electroscope is found to acquire a positive charge, whereas the nearby presence of a negatively-charged object has no effect.)

Pupils will no doubt guess as a result of their previous experience of thermionic emission from metals that the incident radiation produces a loss of electrons from the zinc plate. They should be asked how this guess could be verified; the answer is "by measurement of the charge-to-mass ratio of the emitted particles and by measuring the total charge on a known number of particles emitted from the plate" (cf Section 7.14.3). Measurements have been made of the charge-to-mass ratio of particles emitted photoelectrically; in all cases the figure obtained has been that for the electron. The process by which electrons are ejected from a metallic surface by incident radiation can be explained in terms of the acquisition by the electrons of energy from the incident beam. Those electrons which have gained sufficient energy in this way are able to overcome the attractive electric forces binding them to the metal and so they leave its surface.

10. 1. 3 "Problematical" aspects of the photoelectric effect

The above features of the photoelectric effect can readily be reconciled with the view that the incident radiation is a form of wave motion. One can see in a general way how electromagnetic waves can transfer energy to the free electrons in the metal; the electric field component in the wave can

/exert.....

exert forces on the electrons and do work on them. The above experiment also shows that the bigger the rate of flow of energy to the metallic surface, the bigger is the rate at which it loses its charge, as we would expect in terms of this explanation.

The pupil now needs to be confronted with the "disturbing" aspects of the photoelectric effect, viz. (i) the absence of any correlation between the intensity of the incident light and the time elapsed between the turning-on of the beam and the emission of photoelectrons; (ii) the fact that if the frequency of the radiation is too low, no photoelectrons will appear no matter what the intensity may be; (iii) the fact that the energy of the photoelectrons depends on the frequency and not the intensity of the incident light. It is customary in elementary treatments to deal only with points (ii) and (iii), but point (i) is just as important, since from a detailed study of it there emerges a picture of a light beam as a randomly-arriving stream of "particles." In the P.S.S.C. treatment of the topic (1960, 1965), this aspect of it is considered first.

10.1.4 Randomness of absorption of photons

Unfortunately, those aspects of the photoelectric effect which run counter to the wave theory of light are much less easy to demonstrate convincingly than those which do not appear to conflict with it. Point (i) in Section 10.1.3 in the form in which we have stated it is impossible to demonstrate in an elementary course, although we can demonstrate the random arrival of photons from weak beams of ultra-violet light or X-rays by using Geiger-Müller counters to detect the arrival of individual photons from these beams (cf Nuffield Guide V). With points (ii) and (iii) we can fare somewhat better, although even here the demonstrator's task is far from easy, even when he has managed to acquire the right

/equipment.

equipment. Both the Nuffield and the P.S.S.C. authorities think it best to solve the problem by showing the P.S.S.C. films on "The Photoelectric Effect" and "Photons", with any demonstrations that might be done being allocated a secondary role. This recommendation is obviously sensible, and one can only hope that these films will be more readily available to South African schools than they have been up to the present. Since there may be some who would wish to explore the possibilities of teaching some of this material by direct demonstration, we give below a brief survey of some of the methods that can be used and the difficulties that attend them.

In principle, point (i) is very simple to show, and the P.S.S.C. textbook gives a detailed account of an elegant experiment that was carried out to investigate it. (Those responsible for the original experiments of this type are not named in the P.S.S.C. textbook, but it would appear that it is the work of Derieux (1918) and Kelly (1920) that is being referred to. Cf. Hughes and Du Bridge (1932).) A number of tiny specks of metal float in the space between a pair of parallel plates having equal and opposite charge. It is supposed that the specks all have the same mass and that they all carry the same charge. If a light beam is shone on to the specks, electrons will be liberated from them, their charge must change, the electric and gravitational forces on them no longer balance and they "jump", i.e. they suddenly change their velocity. In experiments of this type, it is found that there is no definite "waiting period" after which all the specks jump, as might be expected if the light were simply a form of wave motion, nor do all the specks jump at the same time. Some of the specks may jump immediately after the beam is turned on; afterwards others will follow suit, but not according to any particular pattern. No matter how weak the

/light.....

light beam is, there always seems to be some chance that a speck will jump immediately after the beam is turned on. On the average, more electrons are ejected per unit time by a strong than by a weak beam, but the behaviour of individual specks appears to be entirely random. Thus there appears to be no orderliness at all about the way in which the specks take up energy from the incident light. (Hughes and Du Bridge (1932) have pointed out that, strictly speaking, experiments of this type disclose the random emission of electrons and not the random absorption of photons. In fact a great many of the latter have to be absorbed, on average, before an electron is emitted.)

The above summary makes clear both the elegant simplicity of this type of experiment - no "black boxes" by way of fancy detecting equipment, no sophisticated methods of timing - and its utter impracticability as a demonstration experiment in an elementary class. The experiment shown in the P.S.S.C. film "Photons" (devised by John King of M.I.T.) is described and discussed in some detail by French and Hudson (1965). The design of this experiment is simple enough, but the equipment it needs is of a high order of sophistication. Light from a lamp is shone on to the photocathode of a photomultiplier tube. The incident beam is interrupted by a rotating shutter which allows one pulse of light per shutter revolution to pass through to the photomultiplier. A filter is interposed between the shutter and the photomultiplier to reduce the intensity of the light by a known factor. The output from the photomultiplier tube is displayed on the screen of an oscilloscope. For large light intensities, this output appears as a smooth variation of voltage with time. For considerably reduced intensities, the output is no longer smooth but "spiky", indicating the fact that finite numbers of electrons are being

/released.....

released from the photocathode of the tube. So far, there is nothing in the experiment that points to a "graininess" in the incident light. However, suppose the intensity is reduced to such an extent that the average energy arriving per pulse at the photocathode is less than the amount needed to eject a single electron. If light were simply a form of wave motion, then the energy arriving at the photocathode would be spread uniformly over the illuminated surface and there would not be enough energy reaching any one portion to liberate an electron. In fact, the experiment shows that electrons are emitted even in these circumstances. This suggests that the incident energy, instead of flowing in a continuous stream, may be arriving at the surface in a random succession of "packages", so that although on average there is insufficient energy in the beam to eject an electron, any one package may contain more than enough energy for this purpose. In addition, the experiment shows that even with very weak incident light, electrons may be emitted right at the beginning of the light pulse.

French and Hudson stress the need for caution in interpreting the results of this experiment; that although it is difficult to interpret the results of this experiment in terms of the wave theory of light, it is not impossible. For example, one could imagine that some sort of energy storage mechanism was operative in the metallic surface, having the effect that energy received during one incident pulse was only concentrated on particular electrons as kinetic energy during the next pulse. Therefore, although the experiment is "beautiful and suggestive" (French and Hudson), it should not be taken as establishing by itself the case against the wave theory of light. (Cf. also the point mentioned above concerning the fact that in experiments of this type we are

/not.....

not directly observing the random absorption of photons but the random emission of electrons consequent upon the absorption of a great many photons by the surface.) Teachers using the film or simply describing the experiment should encourage their pupils to look for "contrary" interpretations of this kind.

The "extra" equipment needed for King's experiment, such as a photomultiplier tube and high-quality oscilloscope, would not be available at any South African school nor, one suspects, at most schools in the world. Even if it were available, a dedicated enthusiast would probably have to spend many hours in getting it to "go" to his own satisfaction. Thus it, too, is not a practical proposition as a demonstration experiment.

We seem to be left with the demonstrations in which ultraviolet light or X-ray beams are directed on to Geiger-Müller tubes and cause the latter to "count". In their way, these are impressive demonstrations, since they show that an apparently steady beam of radiation can produce a disorderly succession of counts. The Nuffield Guide (V) recommends holding lighted matches in front of a G-M tube that is intended to count alpha particles, and using a G-M tube intended to count gamma-rays to detect the photons from an X-ray tube operated at low filament voltage ("barely warm") and less than its usual high tension voltage.. Of the first demonstration, it is implied that interposition of a sheet of glass between the G-M tube and the lighted match will completely stop the counting, so that visible light is presumably not responsible for the counts. Of the other demonstration, the Nuffield Guide (V) says: "A G-M tube at the other side of the room will count X-ray photons one by one, showing us a random stream." This implies that each photon is counted as it arrives, and nothing is said elsewhere to contradict this implication. Actually, G-M tubes designed to count gamma-rays do not count

/each.....

each photon. Whereas a "particle" entering a G-M tube will readily produce an ion pair and so "trigger" the tube, gamma-ray or X-ray photons rarely produce ion pairs. Their arrival is recorded by a "secondary" process: if a photon reaches the wall of the tube, it may eject an electron by photoelectric emission and this will trigger the tube. But by no means every incident photon will cause the emission of an electron; in fact, the "counting efficiency" of gamma-ray sensitive G-M tubes for incident photons is generally less than 1%, i.e. less than 10 counts for every 1 000 incident photons (cf Segrè 1964). This fact does not detract from the value of the demonstration, but teachers should be careful not to overstate the facts when interpreting it. (Were semiconductors readily available, it might be possible to obtain much greater counter efficiencies: "With some semiconductors one can get close to 100 per cent" (Frisch 1965).) For this demonstration, it is desirable that the G-M tube should not remain entirely a "black box"; some account of its principle of operation should be given to pupils.

These films, demonstrations or descriptions of experiments should be supplemented by the use of photographs. French and Hudson reproduce a very striking series of photographs taken from an article by Rose (1957) showing how the quality of a photograph improves as the amount of light used to produce the image is increased. There are six photographs in the sequence, and the overall increase in the "amount of light" (i.e. the energy used to produce the image) is by a factor of 10^4 from the first picture to the last. (A similar sequence of photographs is reproduced by Huggins (1968).) In the first picture, all that can be seen is a mass of randomly-distributed white dots; in the second, it becomes evident that there are more of these dots in some

/places.....

places than in others, and an overall pattern of distribution begins to emerge. This feature is confirmed by the third picture, which begins to look like a conventional photograph, and of course the quality of the image improves progressively in the remaining three pictures. As French and Hudson remark, the "random-graininess" (our word) that is so strongly evident in the images obtained with very weak light would present very great difficulties of interpretation for a wave theory of light, but is instantly understandable in terms of a corresponding "random-graininess" in the incident light. Pupils certainly ought to see photographs like this. Only the result of this experiment is "mysterious"; what could be simpler in principle than producing a series of photographic images of the same object with progressively increasing illumination? The teacher who is interested in photography might like to make his own set of suitable photographs. Photographs showing the result of the passage through a cloud-chamber of a beam of X-rays are also relevant here. The P.S.S.C. textbook (1960,1965) reproduces a photograph showing the tracks produced by photoelectrons ejected by X-rays from argon atoms in a cloud chamber. What is particularly significant is the similarity of the photoelectron tracks in length, indicating that each photoelectron must have received much the same energy from the incident beam.

10.1.5 The photoelectric effect: dependence on frequency

The above evidence is concerned to establish the fact that, as far as its interaction with matter is concerned, a beam of light behaves as though it were a randomly-fired volley of packets of radiant energy. Investigation of what we have called points (ii) and (iii) in Section 10.1.3 reveals some information concerning these "packets". The simple demonstration of the photoelectric effect with an electroscope

/described.....

described earlier can easily be extended to give an indication of point (ii). A glass plate placed between the mercury vapour lamp and the zinc plate arrests the discharge of the electroscope, if the latter has been given a negative charge. Ultra-violet "light" cannot penetrate the glass, so that the emission of photoelectrons from the zinc plate is presumably brought about by the ultra-violet component of the incident radiation and not by the longer wavelength visible light component. This suggests that there is a connection between the wavelength of the light and the production or otherwise of the photoelectric effect, a connection that would certainly be difficult to explain on the basis of a simple wave theory. For this connection to be plainer to the pupil, he should see performed, either on film or in the classroom, the classic experiment in which the dependence of the maximum kinetic energy of the photoelectrons on the frequency of the incident radiation is investigated. In principle, this experiment is also exceedingly simple. The essential apparatus comprises a photoelectric cell, a suitable light source, filters or some other means of restricting the incident radiation to one narrow band of frequencies at a time, a very sensitive ammeter for registering the current in the photocell, and a means of applying a suitable adverse potential difference between the electrodes of the cell.
If the photocathode of the photocell is irradiated with monochromatic light, and an adverse potential difference applied between anode and cathode is just large enough to cut off the photocell current, this potential difference is a measure of the maximum kinetic energy of the emitted electrons. (The reason why we stress that it is the maximum kinetic energy that is measured is that the "free" electrons in a metal will not all have the same energy; they are distributed over a range of energy levels up to - or, at room temperature,

/slightly.....

slightly beyond - the so-called "Fermi level". Two electrons occupying different energy levels will consequently emerge with different kinetic energies upon taking up the same energy from the incident radiation.) To carry out the experiment, one need simply show that the adverse potential difference required to produce "cut-off" depends on the frequency and not on the intensity of the incident radiation, and measure this "cut-off" potential difference for different frequencies. Alas, there are difficulties. Photoelectric cells specially made for use in this type of experiment are very expensive (they would cost R25 or more in South Africa). The currents they deliver are exceedingly small: the demonstration photocell marketed by Rank Audio Visual Limited has to be used with a galvanometer able to cope with currents of the order of 10^{-7} A, and that made by Leybold produces currents of as little as 10^{-10} or 10^{-11} A at some frequencies, which means that it cannot satisfactorily be used with conventional galvanometers: a D.C. amplifier, a costly item in itself, has to be used to obtain a measurable deflection for these currents. To carry out the experiment at several different frequencies, one needs some sort of "monochromator" to single out the frequencies required, which means using a set of filters or a spectrometer (cf Bolton 1966), although the latter alternative means working with a reduced light intensity. If the filters are to be any good they will cost several Rand apiece. One can use an ordinary commercial photocell in place of the special cells: Ahlgren (1963) gives details of an experiment using a 929 phototube, which costs "about $3" and Bolton (1966) suggests that Rank VA50 and Mullard 90AV cells can also be used (they cost about R6 each). Sensitive galvanometers must still be used, however; for example, Ahlgren suggests the use of a galvanometer having a sensitivity of at least 1 microampere per mm. There is,

/however,.....

however, at least one exception to this rule. Gottlieb (1965) has described a cheap version of the experiment using a 929 tube, a fluorescent tube as source, a plastic replica diffraction grating as monochromator, and a two-transistor amplifier to amplify the current. The total cost of this apparatus is around 25 dollars, and the amplified current can be read on a 100 microamp galvanometer.

The difficulties we have referred to so far concern only the acquisition of the apparatus. We still have to consider the difficulties attached to the actual performance of the experiment. First, there is the existence of a "contact potential difference" between the photocathode and the collecting anode, which prevents direct determination of the "work function" of the metallic surface (i.e. the minimum work necessary to remove electrons from the surface). The Nuffield Guide (V) lays some stress on the presence of contact potential differences in the experiment; since its discussion of this aspect of the matter seems to us rather puzzling in certain respects, we consider it here. The Guide describes the plotting of the graph of critical opposing voltage versus frequency, and says that "with very careful work and some apprenticeship to the apparatus, the plotted points of the overall graph will be likely to be close to a straight line", but "the best straight line will not pass through the origin because the opposing voltage we measure is not the full voltage that operates inside the tube to drive electrons back. There are other 'contact' potential differences which must be taken into account". In the first place, even in the absence of these contact p.d.'s, the graph would not pass through the origin, but would cut the frequency axis at the "threshold" frequency for the metallic surface in question, i.e. the minimum frequency of light required to eject photoelectrons from the surface. Thus the intercept on the frequency axis finds a very simple physical interpretation. The presence of

/contact.....

contact potential differences will alter the value of the intercept and so make it impossible to determine the value of the work function directly, as we have noted. But this will not affect the slope of the line, which yields by far the most important piece of information in this context, viz. the value of $\frac{h}{e}$ ("h" is Planck's constant, "e" is the electronic charge). The Nuffield Guide's objection that one would have to explain the phenomenon of contact p.d. and so "spoil the teaching of the simple, important ideas of the photoelectric effect" therefore appears to fall away.

More serious objections to the performance of this experiment in the school laboratory concern the difficulty of taking the actual voltage measurements. Even if contact p.d.'s were absent, it is not in fact true in general that the adverse p.d. which produces zero net current is the true "stopping potential". This is because of the "counter-current" produced by photoelectrons emitted from the anode (cf Hoag and Korff 1948, Ahlgren 1963, Iona 1966). Zero net current is obtained when the photocurrent from the cathode is exactly balanced by the counter-current from the anode. To determine the true stopping potential, one should determine the value of adverse p.d. at which the curve photocurrent versus adverse p.d. starts to flatten out. This makes "accurate" performance of this experiment a much longer and more troublesome affair than might have been deduced from our brief description of it above. Add to this the fact that the exceedingly small currents measured in the experiment make it particularly susceptible to malign and often untraceable influences, and we see that it is scarcely an obvious choice as a demonstration experiment. Gottlieb (1965) claims that his version of the experiment will provide "convincing demonstrations", and on account of its low cost, one is inclined to feel that it is worth trying. But the following statement from the

/Nuffield....

Nuffield Guide (V) seems fair as a general summing-up: "...we do not recommend this demonstration except where the teacher has a strong interest in carrying it out - in which case his explanation will gather force from the sense of personal experience that will invest his teaching".

Those teachers who might want to preface a demonstration of this experiment or the showing of the relevant film with an account of the appropriate theory will have history partly on their side, in that Einstein published his theory of the photoelectric effect eleven years before Millikan published the account of his brilliant confirmation of it. (Brigg (1966c) quotes substantial excerpts from Millikan's paper on the subject.) Einstein's theory could be introduced in this way: "We have seen some evidence for light behaving as though it were composed of 'particles' of some kind; let us call them 'photons'. We also know that if the frequency of the light is too low, it will not cause a photoelectric effect however intense we make it. [This supposes that this aspect of the photoelectric effect will already have been discussed.] Let us, following Einstein, try the guess that the 'photons' composing the radiation carry with them an amount of energy proportional to the frequency. A 'photon' associated with radiation of high frequency would then have a comparatively large amount of energy to hand over to an electron, enabling it to detach itself from the surface of the metal and perhaps move off with considerable kinetic energy. By elaborating on these ideas, Einstein was able to produce a very satisfactory theory of the frequency dependence of the photoelectric effect". The very simple derivation of the famous photoelectric equation can then follow and be used in due course to interpret the experimental results obtained in the school laboratory or on film.

/10.1.6.....

10.1.6 The nature of photons

The magic quantities "h" (Planck's constant) and "hν" (the energy of a photon) will have emerged from this discussion of the photoelectric effect. The teacher will now have to make some attempt to answer the questions "What is a photon?" and "How can we combine photons and waves in a single theory?" In a sense, the first of these questions is unanswerable. We know that a photon is characterized by a definite "quantum" of energy which is equal to the frequency of the radiation multiplied by Planck's constant. This leads to the familiar short way of describing a photon, which is to call it a "bundle of energy", a phrase (or one of its equivalents) now well entrenched in the physicist's vocabulary. How literally should one take this description? A sentence from the P.S.S.C. textbook may help us here: "Each photon carries (or perhaps the photon is) the energy hν". Let us note the use of the word "perhaps", and proceed with due caution; there is more to the photon than can be conveyed by a single slick phrase. At least one can add to this some other information concerning the physical properties of the photon; that it always travels with the speed of light, that it has zero rest mass and that it has momentum. (Some mention of the Compton effect could be made in order to support this last piece of information.) This shows that while a photon has some of the characteristics of what is ordinarily called a particle, the word "particle" is clearly not an adequate description of it. Consideration of the answer to the second of the questions posed above will reinforce this view.

In trying to explain how to combine photon and wave pictures of light into a single description, it is probably best to start by referring to a specific situation. Suppose we set up an arrangement of screens and slits and a monochromatic light source so as to produce an interference pattern from

/"Young's.....

"Young's slits", and that we allow this interference pattern to be formed on a screen. Wave theory tells us and experiment confirms that there will be a pattern of alternating dark and bright bands on the screen, whose separation is determined by the wavelength of the light, the separation of the slits and the distance between the slits and the screen. We fit photons into this picture by saying that the wave theory of light tells us, on average, how the photons from the source distribute themselves over the screen. If large numbers of photons pass through the apparatus, most of them will go to those parts of the screen we call the maxima of the interference pattern, none at all will arrive at the minima, and intermediate numbers will go to the regions of intermediate brightness. According to this idea, the pattern that we can see on the screen (or record on a photographic plate) is produced not by the effect of energy streaming continuously from the source to the appropriate points of the screen, but by the intermittent arrival of photons at these points.

We cannot give a detailed account of what individual photons do. Theory and experiment are powerless to tell us, for example, what path would be followed by a particular photon as it plays its part in the formation of the interference pattern; indeed, they suggest that the concept of a "photon path" is meaningless. In principle, we can record the arrival of a photon at a particular point in space by means of a suitable detector, though we may not always succeed in practice; but the moment we do succeed, we destroy the photon. What it does between the moment of its creation (i.e. its emission from the source) and its destruction, we cannot know about. Suppose the light source is so weak that, on average, only one photon can pass through the apparatus at a time. With such a source, what kind of pattern will be formed on the screen?

/We.....

We can find this out by replacing the screen with a photographic plate and using an exposure time long enough for the light reaching the plate to produce a developable image. (This kind of experiment has been done using exposure times of as long as three months; cf. Taylor (1909).) On completion of the exposure it is always found that the image on the plate is an interference pattern of exactly the same form as that obtained with strong light. A spectacular version of such an experiment was carried out in 1957 by Jánossy and Náray (cf Whiteman 1967), using a Michelson interferometer with very long light paths (about 14m) for the interfering beams. It was found that the "The placing of the [interference] bands (about 1.5 mm apart) was conspicuously the same for a low intensity (less than one photon in the apparatus at a time) as for an intensity 10,000 times as great" (Whiteman 1967).

These facts are very remarkable: they suggest that although a photon may start out from a well-defined point in the source, and end its life on an equally well-defined point on the screen or photographic plate, it somehow travels through both slits at once - it "interferes with itself". (Attempts to localize the photon near one or slit or the other and so show that it must go through one slit or the other, can easily be shown to be doomed to failure, since they inevitably lead to the destruction of the conditions under which interference can take place.) This is another indication of the special nature of the photon: no ordinary particle can behave in this way.

In the case of Taylor's experiment we can picture the photographic image being built up by the photons arriving at the plate one by one. The times of arrival of the photons will be irregular, and at first there will not seem to be any

/particular.....

particular orderliness in the way in which their points of arrival are distributed over the plate. But eventually, after enough photons have arrived, some orderliness does begin to appear; it definitely seems that more photons go to parts A, B and C of the plate, say, than to parts D, E,F.... Eventually we find that the overall pattern built up by the arrival of a great many photons is what we recognize as the familiar two-slit interference pattern. (We have already described some of the evidence that shows that the image of the pattern is built up in exactly this way.)

We can sum up the main features of the present view of the "wave-particle duality" of photons as follows:-

(a) Neither the wave theory nor any other can tell us where a particular photon will go. However, the electromagnetic wave theory does correctly tell us the probability of finding a photon at a given place. Where this theory predicts a maximum of intensity, the probability of finding a photon will be high; in other words, if large numbers of photons are emitted by a source, a great many of them will find their way to the place where the wave theory predicts an intensity maximum.

(b) Although we cannot predict the path of an individual photon, we can give an exact description of what happens when a photon interacts with matter. When a photon is absorbed by a metallic surface and a photoelectron is ejected, we can place an upper limit on the kinetic energy of the emitted electron by using Einstein's quantum theory of the photoelectric effect.

(c) Suppose that we arrange to form an interference pattern on a flat metal plate, and we use for the purpose light of sufficiently high frequency to eject photoelectrons from the plate. Then the wave theory will tell us at which regions of the plate most electrons will be ejected, while the quantum theory will tell us the maximum kinetic energy we can expect the ejected electrons to have.

/(d).....

(d) The photon has some particle-like attributes, but the collective behaviour of photons has to be described by a wave theory. So the entity that is called a photon is something whose attributes do not fit neatly into a category labelled "particle" or a category labelled "wave". It may be that sometimes we are more aware of the "wave" character of photons than of their "particle" character, or vice versa. In a very intense beam of light, for instance, so many photons are arriving per unit area per unit time that one cannot be directly aware of the arrival of individual photons. It is only when one allows the light to interact with matter, e.g. to be absorbed by atoms, molecules or electrons, that its "particle" aspects become evident. Nevertheless the results of countless experiments establish that all forms of electromagnetic radiation - from the lowest-frequency radio waves to the highest frequency gamma-rays - exhibit the kind of "wave-particle duality" we have been talking about. As the frequency of the radiation is increased, so its particle-like aspects become more obvious and its wave-like aspects correspondingly less so, since the energy of the photons is comparatively high and the wavelength of the radiation very short. But we can still obtain diffraction patterns with X-rays or lower-energy gamma-rays by using crystals as our diffraction gratings; and for higher-energy gamma-rays one can use the atomic nucleus as the means of obtaining diffraction with them (cf P.S.S.C. 1960, 1965). As the frequency of the radiation is lowered, so we observe the inverse behaviour. For radiation at microwave or radio frequencies the energy of a photon is very low, so that particle effects become very difficult or impossible to observe (there is no hope of producing a photoelectric effect with radio waves, for example); on the other hand the wavelength of the radiation becomes very long, so that wave effects are readily observable. But atoms and molecules do

/emit.....

emit or absorb photons having these low energies; the principle of operation of the newly-adopted caesium "standard clock" is based on this phenomenon.

(e) In trying to see the photon whole, we are compelled to the realization that no model of this entity that we may try to build using ideas familiar from the macroscopic world can fit all the facts. We have now been led to a description of nature that seems decidedly challenging to "common sense". But the "common sense" that is challenged by the "wave-particle" picture of light is founded on the familiar world of everyday experience - a world in which waves are waves (meaning waves on water or stretched springs or strings) and particles are particles (meaning objects that can always have a definite momentum and position ascribed to them, like marbles or cricket balls). In seeking to discover more about light, we have had to probe into a world some aspects of which are a good deal removed from everyday experience. Perhaps it is not so surprising after all that we should in the end have to be content with a description using ideas that cannot conveniently be accommodated within a framework of "common sense."

We have really only touched on the problems raised by the dualistic nature of the photon. Some of these problems have been entertainingly dramatized by Frisch (1965) in the form of a conversation between several characters. In his summing-up, Frisch remarks that he still feels "a bit uneasy" about certain aspects of the interpretation of quantum theory; this lends point to the "negative teaching objective" which was the fifth of the teaching objectives listed at the beginning of Section 10.1.1. In the next section we discuss some of the problematical aspects of quantum theory in some detail, and consider their implications for elementary teaching of the subject.

/10.2.....

10.2 The interpretation of quantum theory and its implications for elementary teaching

10.2.1 Wave-particle duality and quantum theory

The average South African teacher concerned with trying to teach the delicate topic of wave-particle duality is not likely to have an extensive acquaintance with quantum theory. Such an acquaintance is not essential for good teaching of this topic, but if it is to be taught with confidence and conviction, the teacher should have some familiarity with current views on the interpretation of quantum theory beyond what he might find in the official textbook or what he was told at university. This will help him to feel that what he has to say about the nature of the electron and the photon has some claim to acceptance and is free from self-contradiction (even though his pupils may not immediately be convinced of the latter) and will define for him the areas of the subject that are still controversial. For although quantum theory is generally accepted, its interpretation is still a source of lively debate among philosophers of science. It has been claimed that the so-called "Copenhagen interpretation" of the theory (see Section 10.2.2 below) "is still espoused today by the majority of theoreticians and practising physicists" (Jammer 1966; Peierls (1967), reviewing Jammer's book, concurs), but this interpretation is by no means the only one in the field; it has been strongly criticized by various workers (e.g. Margenau 1950, 1968, Landé 1960, 1961, 1965a, 1965b, Bunge 1967, Popper 1967 and Witmer 1967) who have put forward other interpretations. (Still others have even declared themselved dissatisfied with the existing theory and have sought to supplant it with theories of their own; since none of these theories has yet attained the status of a working alternative to the established quantum theory, they will not be considered here.)

/Fortunately.....

Fortunately it is possible, as we shall show, to extract certain ideas from the controversy which can serve as the basis for an adequate elementary treatment of wave-particle duality.

10.2.2 The Copenhagen interpretation

In view of the esteem in which this interpretation of the quantum theory is said to be held, it is fitting that we should begin our survey of approaches to wave-particle duality with an examination of what the Copenhagen interpretation has to say on the subject. At the heart of this interpretation stands the "Principle of Complementarity", associated with the name of Niels Bohr. According to Bohr, "the impossibility of any sharp separation between the behaviour of atomic objects and the interaction with the measuring instruments which serve to define the conditions under which the phenomena appear" and the fact that "however far the phenomena transcend the scope of classical physical explanation, the account of all experimental evidence must be expressed in classical terms" compel us to adopt the principle that "Evidence obtained under different experimental conditions cannot be comprehended within a single picture, but must be regarded as complementary in the sense that only the totality of the phenomena exhausts the possible information about the objects" (Bohr 1958). The "evidence" referred to here is experimental evidence concerning the nature of microsystems such as electrons and photons: Bohr is saying that "experimental arrangements in which behaviour can be visualized as predominantly wavelike exclude arrangements under which particle-like behaviour is dominant, and vice versa" (Barbour 1966). As an example, we may take the familiar thought experiment of the diffraction of electrons by a pair of slits (cf Feynman et al 1965 for a particularly full and clear

/discussion.....

discussion of this experiment). If the electrons from a suitable source pass unhindered through the slits, they produce an interference pattern on a screen beyond the slits. To this extent, they exhibit wave behaviour. Were the electrons only capable of behaving like waves, it would be meaningless to ask through which slit a particular electron passes. However, knowledge that the electron does possess particle-like attributes suggests that the question might be worth asking experimentally, say by allowing the electrons emerging from the slits to scatter photons from a suitably-placed source of "light" of sufficiently short wavelength. Then one might be able to establish that an electron apparently has gone through a particular slit, and so revive confidence in the view that the electron is really a particle. But one can only achieve this at the expense of destroying the interference pattern and with it, the macroscopic, "classical" evidence of wave-like behaviour. When a photon from the light source is scattered by an electron, the latter is also deflected, and the nature of the interaction between photon and electron (and of these entities themselves) ensures that the conditions under which interference was originally observed no longer apply. According to Bohr, one cannot accommodate the results of these two experiments "within a single picture" (in the sense that the electron or photon cannot be a wave and a particle at the same time), but neither experiment tells the whole story about the "behaviour of the atomic object" (Bohr 1958); that can only be given by listing the information obtained in both experiments. The "wave-like" and "particle-like" parts of the story are complementary, in the sense that both are needed for a description of nature.

In view of the stress that is laid in this approach to complementarity on the role of the experimental arrangement in bringing to the fore the wave-like or particle-like

/characteristics.....

characteristics of electrons and photons, it is worth pointing out that what Jammer (1966) calls the "sharp exhibition of such complementary notions" may well not occur in any practical experimental arrangement, contrary to what he seems to imply. For example, in an experiment in which light is diffracted, the diffraction pattern is still built up by the random arrival of photons at different points on the screen; and if the screen is made of suitable metal, and enclosed in a vacuum, the maximum energy of the ejected photoelectrons will still be determined by the energy of the incident photons. Although one of the "complementary notions" may predominate, it does not seem possible in principle completely to suppress the other.

To complete our picture of complementarity, we may mention that a number of authors follow Pauli (cf Jammer 1966) rather than Bohr in their statement of the Principle of Complementarity, preferring to express it in such terms as the following: "At the quantum level, the most general physical properties of any system must be expressed in terms of complementary pairs of variables, each of which can be better defined only at the expense of a corresponding loss in the degree of definition of the other" (Bohm 1951). Here the "complementarity" refers specifically to pairs of "variables" (e.g. position and momentum); in Bohr's exposition, it is assumed that complementarity may refer to more general concepts (or "notions"), such as those of wave and particle. We shall continue to interpret the principle in the latter sense, mainly because this makes it easier in an exposition at this level to relate it to the issue under discussion, viz. the nature of "microsystems" such as the electron and the photon.

/We.....

We must now consider what answer, if any, is given by adherents of the Copenhagen interpretation to the direct question, "What is an electron?" (or "What is a photon?"). One type of answer is the following: "...the question: 'Does light consist of particles?' is neither true nor false. The point is that the answer to this question depends on the experiment one has in mind. If we try to find the nature of light by doing a diffraction experiment we find that it possesses wave character, but if we perform a photoelectric experiment we find its corpuscular nature" (ter Haar 1959). This kind of answer has been dramatized by the memorable dialogue: "Daddy, is [the electron] a particle or a wave?" "Yes" (Hoffmann 1959). As we have implied above, it is going too far to suggest that we can completely isolate the wave or particle properties of a photon or electron in an experimental arrangement, so that it is simply not true to suggest, as ter Haar seems to be doing, that we can give a clear-cut answer, "The electron (photon) is a wave", or "The electron (photon) is a particle", depending on the kind of interrogatory experiment we choose to carry out. Ter Haar's further conclusion, that "one needs a logic system which is no longer based on the excluded middle or tertium non datur to deal with this problem" is thus unjustified. It is not clear whether ter Haar is actually adopting the view characterized by Whiteman (1967) as "One-level Duality, i.e. the view that there is just one level of physical actuality ...and that there will exist, at that one level of actuality, either particles (Euclidean points?) or ideally continuous waves, according to the experimental set-up which we choose to arrange." Whiteman remarks that "This [view] has merely to be stated with some precision of language...to be seen as self-contradictory and also incompatible with the essential features of the Copenhagen Interpretation". It seems to us to be not only self-contradictory but incompatible with what can actually be revealed experimentally.

/Another.....

Another attitude to our fundamental question shown by adherents to the "Principle of Complementarity" has been neatly summarized by Bunge (1956), who does not share it: "Microsystems... simply are not in themselves, since they do not exist apart from the experimental set-ups designed to measure observables. The words particle and wave designate neither material objects nor properties of material objects, but concepts, images or models that we use in the complementary descriptions of certain possible experiments. Outside the descriptions of experiments (which descriptions are equivalent, whether they are couched in the corpuscular or undulatory language), there are neither particles nor waves nor wavicles. It is meaningless to ask about the properties of something which, like matter, is only a fiction". Some physicists would no doubt prefer this positivistic approach to the problem - it absolves them from the duty of having to answer awkward questions; others would perhaps not go as far as to assert that the "microsystems" are only fictions, but would merely claim that they are concerned with answering questions about the behaviour of these microsystems and not about their nature. The latter would appear to be Bohr's own standpoint, since he is content to use such terms as "atomic object" and "atomic corpuscle" in such a way as to imply that these entities exist, but does not commit himself to any overall view of these microsystems.

In sum, it appears to be common practice in interpreting the Principle of Complementarity to stand aloof from answering direct questions concerning the nature of microsystems, and instead to "accept a set of mutually exclusive representations appropriate under different circumstances" (Barbour 1966), on the understanding that these representations are simply models of limited applicability, or to assert that these microsystems are waves or particles, depending on the experimental

/circumstances.

circumstances. We have indicated above why we believe the latter view to be untenable. As for the fence-sitting nature of the former view, this is unlikely to satisfy many of those who hold to the belief that experiments like those of Millikan (the "oil drop" experiment), Davisson and Germer, and G.P. Thomson and the various determinations of the charge-to-mass ratio of the electron were after all measuring the physical properties of something. It is natural that they should inquire what this "something" is and that they should hope for a straight answer. This seems to be denied them within the framework of the orthodox Copenhagen interpretation. We consider below some attempts at providing "straight answers".

10.2.3 Other interpretations

Among the most trenchant critics of the Copenhagen interpretation is Landé, who in a series of publications (Landé 1960, 1961, 1965a, 1965b) has upheld a "unitary-particle" view of quantum theory, as opposed to the "dualistic" view enshrined in the Copenhagen interpretation. According to Landé, "Matter consists of particles which by their statistical quantum performance give the appearance of waves. Radiation consists of 'field oscillators' which by their quantum actions produce the appearance of photonic particles dashing around" (Landé 1961). To explain how the "statistical quantum performance" of particles can "give the appearance of waves" in the case of diffraction of electrons by a crystal lattice, Landé resorts to the "mechanical particle" theory of diffraction put forward by Duane in 1923 (cf Bohm 1951). In terms of this theory, the crystal can change its momentum perpendicular to its lattice planes in amounts of magnitude $\frac{nh}{D}$, where D is the spacing between the lattice planes, h is Planck's constant and n = 1, 2, 3... . Transfer of these amounts of momentum to electrons would result in their being deflected

/in.....

in directions which are exactly those predicted on the assumption that the electron has a wavelength given by the de Broglie relation $\lambda = \frac{h}{p}$. Single- or double-slit diffraction of electrons is explained by pointing out that the "matter distribution curve [of screens containing a single slit or a pair of slits] can be decomposed into sinusoidal components of various periodic lengths ℓ occurring with various amplitudes" (Landé 1965b), each periodic space component reaching from one end to the other of the screen. Incident electrons react "to the screen as a whole solid body by way of its harmonic components ℓ" (Landé op. cit.). Once again, application of the rule of quantized momentum transfer leads to the diffraction pattern predicted on the assumption that the electron has appropriate wave properties. To the objection usually raised against a particle explanation of "two-slit" diffraction, viz. "How does the particle-like electron passing through one slit know there is another slit further along so that it can make its due contribution to a two-slit diffraction pattern?" Landé answers that "it does not need to know because it reacts to the diffractor as a whole mechanical unit via its periodic space components ℓ" (Landé 1965b).

The difficulties with this explanation of the wave-like behaviour of matter have been pointed out by Bohm (1951) and Whiteman (1967): attempts to make it work in cases other than the special ones chosen by Duane and Landé require the use of special ad hoc rules to suit the new situations and, in certain awkward cases, the adoption of "complicated, artificial and implausible assumptions" which are unlikely to be either self-consistent or "consistent with all known data concerning the properties of matter" (Bohm 1951).

A subtler form of "unitary particle" interpretation has been proposed by Popper (1962, 1967). Laying stress on the view that quantum mechanics is a statistical theory, he

/attributes.....

attributes what he calls the "great quantum muddle" to "taking a distribution function, i.e. a statistical measure function characterizing some sample space (or perhaps some 'population' of events), and treating it as a physical property of the elements of the population" (Popper 1967). According to Popper, "the elements" in quantum theory (e.g. electrons) have the properties of a particle, and the "wave shape (in configuration space) of the ψ function", which is associated with the distribution function, "has next to nothing to do with the physical properties of the particles" (Popper 1967). To justify his interpretation of the ψ function, Popper appeals to his "propensity interpretation" of probability. This is an attempt to provide an objective interpretation of probability other than the purely operational one in which "probability is the relative frequency in a long run" of repetitions of an experiment. Pointing out that such relative frequencies will depend on the experimental arrangement, he suggests that probabilities be looked upon as properties of this arrangement, and so arrives at his "propensity interpretation", according to which probabilities "characterize...the propensity of the experimental arrangement to give rise to certain characteristic frequencies when the experiment is often repeated" (Popper 1962). On this view, the ψ function of an electron "determines the propensities of the states of an electron. ...The electron is a particle, but its wave theory is a propensity theory which attributes weights to the electron's possible states" (Popper 1962). In sum, "Propensities are properties of neither particles nor photons nor electrons nor pennies. They are properties of the repeatable experimental arrangement" (Popper 1967).

There is an appealing simplicity about a view of matter in which "particles" like electrons are really particles in the classical sense, while their allegedly wave-like behaviour,

/which.....

which is bound up with the ψ function, is to be regarded as a property not of themselves alone but of the entire experimental arrangement. At first, this interpretation seems to be supported by the fact that the potential function and the boundary conditions, set by the experimental arrangement, play a crucial role in determining the form of the ψ function. But this is far from sufficient reason for regarding the function as a property of the experimental arrangement rather than of the "particle". The natural modes of vibration of an air column in a pipe are determined by the boundary conditions at the ends of the pipe, but this does not mean that the sound waves in the pipe are a "property of the pipe". The fundamentally significant fact here is that the ψ function (or its equivalent in some other "representation" of quantum mechanics) is essential to the description of the behaviour of a quantum-mechanical microsystem, and that this ψ function has always to satisfy a particular partial differential equation, viz. Schrödinger's (or Dirac's in relativistic situations). We shall only be able to uphold Popper's view that the ψ function is a property of the experimental arrangement rather than of the microsystem if we can somehow show independently of the conventional quantum-mechanical formalism that the behaviour of a microsystem can always be treated as the interaction between a classical particle and the whole experimental arrangement. This is precisely what has been attempted by Landé (see above), and it is significant that Popper (1967) appears to accept Landé's treatment. We have seen that this treatment is not free from objection, and must conclude that Popper's view that we should speak of a particle (in the classical sense) and its associated propensity fields rests on insecure foundations.

The name of Max Born has sometimes been associated with a "unitary-particle" interpretation of quantum theory (cf

/Bunge.....

Bunge 1967). However, the account Born gives of quantum theory in his book "Atomic Physics" (Born 1946) seems to be written from the viewpoint of an orthodox adherent of the Copenhagen interpretation; and in a rebuttal of Schrödinger's "unitary-wave" interpretation, Born (1956) has the following to say of "particles" in quantum mechanics: "[The concept of a particle] must satisfy two conditions: First it must share some (not in the least all) properties of the primitive idea of particle (to be part of matter in bulk, of which it can be regarded as composed), and secondly this primitive idea must be a special, or better, limiting case." One cannot conclude from these statements that Born believes that the "particles" of quantum mechanics are classical particles "which, due to their random motion, behave collectively as if they were waves" (Bunge 1967). This view has been advocated by Bopp (1962), who has backed it up with a detailed theory. The latter is not of the kind that can be summarized in simple terms; it has not achieved a widespread following, and it is not clear how it meets the difficulty faced by all "unitary-particle" interpretations, namely to explain how a particle can interfere with itself.

The above alternatives to the Copenhagen interpretation have all been "unitary-particle" ones. It is not surprising that "unitary-wave" interpretations have also been proposed. The foremost proponent of such an interpretation was Schrödinger. In the early days of the development of his own "wave mechanics", Schrödinger suggested that the absolute square of the ψ function representing an electron should be interpreted as the weight function of a charge distribution in space, and that "the particles of corpuscular physics are essentially only wave groups composed of numerous, strictly speaking, infinitely many, wave functions" (Jammer 1966).

/Having.....

Having shown that a wave packet constructed from the eigenfunctions of the linear harmonic oscillator did not spread out in the course of time, and that the maximum of such a wave packet executed simple harmonic motion, he "concluded that the particle picture of the oscillator [could] be consistently interpreted as resulting from the existence of such a wave packet" (Jammer op. cit.) and believed that it was possible in the same way to retain the picture of electrons orbiting in hydrogen-like atoms. However, it was soon shown by Heisenberg that such "matter-wave packets" must, in general, spread out (the harmonic oscillator is an exception in this respect), and this "concrete" picture of matter waves had therefore to be abandoned. But to the end of his life, Schrödinger held to the view that waves of some kind were the ultimate reality underlying matter. In a lecture given in 1952, he remarked that "... once one deprives the waves of reality and assigns them only a kind of informative role, it becomes very difficult to understand the phenomena of interference and diffraction on the basis of combined action of discrete single particles." The "waves of particles" have an individuality which the microscopic particles do not have; the atoms of a given mass of gas like helium must be treated as indistinguishable in statistical-mechanical computations, but "every one of the [associated system wave functions] has a well-defined structure which is different from that of any other" (Schrödinger 1953). Still, one cannot readily abandon the concept of "corpuscles", and one is bound to ask how "corpuscles" fit into this picture of matter ultimately being a kind of wave motion. To this question Schrödinger would not give a precise answer. "At the most", he remarked, it may be permissible to say that one can think of particles as more or less temporary entities within the wave field whose form and general behaviour are nevertheless so clearly and

/sharply.....

sharply determined by the laws of waves that many processes take place as if these temporary entities were substantially permanent beings. The mass and the charge of particles, defined with such precision, must then be counted among the structural elements determined by wave laws. The conservation of charge and mass in the large must be considered as a statistical effect, based on the 'law of large numbers' " (Schrödinger 1955).

Concerning this interpretation, one might observe that its claim to acceptance would be strong if the ideas expressed in the above quotation were incorporated in a satisfactory theory. But in quantum mechanics in its present form, there appears to be no way of accounting for the manner in which the "mass and charge of ...particles" are "determined by the wave laws". Schrödinger's "unitary-wave" picture of matter should be regarded not as an interpretation of orthodox quantum theory, but rather as an interpretation in search of a theory. One should not despise it on that account; it is always possible that Schrödinger's faith in the ultimate reality of "matter waves" might turn out to be justified. But we are concerned here with interpreting orthodox quantum theory, and in this context it seems that we are not entitled to press the claims of Schrödinger's interpretation.

An obvious way of circumventing the difficulties encountered by the Copenhagen interpretation and by the "unitary-particle" and "unitary-wave" interpretations is to adopt the view that electrons (and photons and other "microsystems") are neither "particles" in the classical sense nor waves, but entities which have some of the attributes of both and which retain these attributes in all circumstances. (To alter appropriately the dialogue given by Hoffmann (1959) and quoted above: "Daddy, is the electron a particle or a wave?" "No".) It is obviously impossible to visualize entities of this sort, but

/this.....

this is no good reason for refusing to countenance them; it may be that this is the price we shall have to pay for a self-consistent interpretation of quantum theory. A number of contemporary authorities have shown their willingness to embrace this view of microsystems. Cf. Feynman et al (1965): "Things on a very small scale behave like nothing that you have any direct experience about. They do not behave like waves, they do not behave like particles, they do not behave like clouds, or billiard balls, or weights, or springs, or like anything that you have ever seen... Historically, the electron, for example, was thought to behave like a particle, and then it was found that in many respects it behaved like a wave. So it really behaves like neither. Now we have given up. We say: 'It is like neither' ."

Wichmann (1967): "The conclusion to be drawn from [the] experimental facts is that an electron is not a particle! It is not a particle in the classical sense of the word... Electrons have wave properties but they most certainly are not classical waves: the electron wave-packet cannot be split like a classical wave-packet". (Throughout his book, Wichmann is concerned to stress that electrons, protons, neutrons, photons, etc. are really what he first of all calls "rapticles", i.e. objects which behave neither like classical particles nor classical waves.)

Bunge (1967): "One of the goals of this study will be to justify our claim that quantons [Bunge's term for the entities with which quantum mechanics is concerned] are neither particles nor waves... at the quantum level there seem to be no corpuscular and no undulatory properties but rather particle-like and wave-like ones - in short the 'particles' and 'waves' of quantum mechanics are metaphorical not literal, and we had better stop trying to picture the new in old terms."

Margenau (1950): "...the obvious properties of waves and particles are incompatible... Furthermore, wave and particle

/aspects.....

aspects do not exhaust the character of an electron...Electrons and photons are neither particles nor waves. They are no more one or the other than they are hot and cold, red and blue. And in saying this one should point out with particular emphasis that this is not an admission of ignorance. We mean to claim it as a positive fact that an electron is neither particle nor wave, and we deny that we don't know what it is." (Contrast this attitude with that of Wilkinson (1959): "Those who ask what an electron is are guilty of a sort of rank anthropomorphism.")

It will be noted that, properly speaking, this is also a unitary rather than a dualistic view of quantum-mechanical microsystems. The latter may be "dualistic" in the sense that they partake of both wave and particle properties, but it is not claimed that these microsystems are sometimes the one, sometimes the other, depending on the experimental arrangement; they are always what they are, viz. something special of which we can, alas, form no overall model or picture but which we are not entitled to reject on that account. As Margenau (1968) remarks, "Things too small to be seen need not partake of the simple picturable attributes of visual objects." One may claim with some justice that this particular unitary interpretation of quantum mechanics is the only one free from self-contradiction and not in need of any special pleading. By asserting that "quantons" are neither particles nor waves but possess particle-like and wave-like properties, one is saying nothing that conflicts either with experiment or the theory in its present form. Consequently there is no need to resort to any of the ingenious arguments of the type that have been put forward to reconcile the facts with "unitary-particle" and "unitary-wave" interpretations.

/We.....

We have implied above that there is no difficulty in reconciling the concept of a "quanton" with existing quantum theory. The entities (microsystems) such as electrons with which quantum mechanics deals are postulated to have particle-like properties such as mass and charge, and the wave-like properties arise from the kind of probability statement made by the theory about the positions of the microsystems: the probability of a "quanton" being in an element dv of configuration space is given by $|\Psi|^2$ dv, where Ψ , the state function of the particle, satisfies an equation which is analogous to but by no means identical with the classical wave equation. (For simplicity, we have referred here to a system consisting of a single "quanton" only.) It is arguable that insistence on the "quanton" concept can lead to an improved appreciation of what the theory really says. We have referred above to the importance attached by adherents of the Copenhagen interpretation to the interaction between measuring instruments and atomic objects, with particular reference to the role this interaction supposedly plays in bringing wave-like or particle-like characteristics of microsystems to the fore. It seems to be a tenet of this theory that this observer-object interaction is of fundamental significance for the understanding of what quantum theory is all about. Cf Bohr (1958): "...The elucidation of the paradoxes of atomic physics has disclosed the fact that the unavoidable interaction between the objects and the measuring instruments sets an absolute limit to the possibility of speaking of a behaviour of atomic objects which is independent of the means of observation". Bohr (op.cit.) also refers to "phenomena like individual atomic processes which, due to their very nature, are essentially determined by the interaction between the objects in question and the measuring instruments necessary for the definition of the experimental arrangements." Bohm (1951)

/claims.....

claims that "... at the quantum level of accuracy, an object does not have any 'intrinsic' properties (for instance, wave or particle) belonging to itself alone; instead it shares all its properties mutually and indivisibly with the systems with which it interacts." Someone unfamiliar with quantum theory might assume from these statements that one of its features is the prominence it accords to interaction between the observer (and his measuring apparatus) and the observed object. The actual situation is very different. Let us take the quantum-theoretical treatment of the hydrogen atom as an example, with particular reference to the allowed energy levels of this system. These energy levels can be found (to a good first approximation, at any rate) by substituting the appropriate form of the potential function into Schrödinger's amplitude equation and finding the values of energy which correspond to physically acceptable (i.e. mathematically well-behaved) solutions of this equation. No account is taken in this procedure of any observer-atom interaction.

Admittedly, the relevant postulate of quantum mechanics is sometimes stated in the form: "The only possible values which a measurement of the observable p can yield are the eigenvalues p_λ of the equation $p\psi_\lambda = p_\lambda \psi_\lambda$" (Lindsay and Margenau 1936), thereby introducing an operationalist note into the formulation of the theory. But this bald statement says nothing about the details of any observer-atom interaction in any actual measurement of, say, the energy levels of the hydrogen atom; nor does it offer any reason for the assertion that these particular values of energy will be obtained no matter what mode of measurement is used. If observer-object interaction really were fundamental to the quantum theory, these omissions would seem surprising. However, quantum mechanics is actually used in connection with such

/problems.....

problems as the hydrogen atom as though the solutions of the Schrödinger equation and the associated energy values refer to the hydrogen atom by itself and not to atom-plus-measuring-apparatus. In accordance with this usage, it would be better to word the above-mentioned postulate in the form: "The only possible values of the observable p are the eigenvalues p_λ of the equation $P\psi_\lambda = p_\lambda\psi_\lambda$". (Ziock (1969) gives a statement in this form applicable to energy eigenvalues.)

But, it may be asked, what about the testing of the theoretical prediction of the energy eigenvalues? Surely we are concerned here with observer-atomic object interaction. This is true, but the "interaction" is not of the kind that is neatly described by any "quantum theory of measurement". As Bunge (1967) points out, "the observations and experiments that have corroborated [quantum mechanics] have macrophysical upshots." We can check quantum-mechanical predictions about the energy levels of the hydrogen atom by "macrophysical" observation of the spectrum of the light emitted by hydrogen gas in a discharge tube, or by "macrophysical" observation of the ionization current in a Franck-Hertz type of experiment. In the former case we relate the observations to the properties of the hydrogen atom by tracing with the aid of the quantum theory the connections between the light emitted and transitions between the atomic energy levels. The theory does not predict any disturbance of the energy levels by the action of emission (except in so far as it takes into account the recoil kinetic energy of the atom); it does predict a disturbance of these energy levels as a result of various effects consequent upon the presence of the atoms in a discharge tube, and does so in a manner which enables this disturbance either to be allowed for or to be virtually eliminated in a suitably-designed experiment. This analysis suggests that in an experiment of this type, we are able to gain information about

/"intrinsic.....

"intrinsic properties" of atomic objects: the energy differences deduced from measurements of the wavelengths of spectral lines refer to the atoms themselves, not to atoms-plus-discharge-tube-plus-spectrometer-plus-observer.

Evidently the way the quantum theory is used to make predictions, the way it is tested, and the way it is used in the analysis of such tests all presuppose that the theory is not fundamentally concerned with indivisible system-"observer" combinations but with the "intrinsic properties" of systems. Adherence to the Copenhagen interpretation, or at any rate to its orthodox variety, tends to obscure this fact, if not to conceal it altogether. On the other hand, adherence to the "quanton" concept of matter is not bound up with an insistence on observer-object interaction; instead, one holds to the view that quantons are forever quantons, and it is a natural extension of this view that microsystems should have their own intrinsic properties.

On balance, we incline to the view that of the various concepts of the nature of matter considered above, the "quanton" concept - i.e. the view that the electron is neither particle nor wave, but is a sui generis entity possessing particle-like and wave-like attributes - has the most to be said for it. We think that it could be used as a satisfactory basis for the elementary teaching of fundamental ideas concerning electrons and photons, on the understanding that neither teacher nor pupil need pledge himself to the whole-hearted acceptance of this (or any other) view of matter and radiation. We elaborate on this idea in Section 10.2.6.

10.2.4 The Ψ-function, "matter waves" and reality

When pupils study the "wave-like" aspects of the behaviour of matter, one of the points that will have to be discussed is the nature of these so-called "matter waves". How do they differ from "classical" waves? Are they really "waves"?

/To.....

To what kind of reality, if any, do "matter waves" correspond? (In other words, does the Ψ-function represent something that is physically real, or is it a mere computational device?)

As a first step towards answering these questions, we must pay some attention to the theory which has been built around "matter waves", viz. that form of quantum mechanics known as "wave mechanics". (For simplicity, we confine ourselves largely to non-relativistic wave mechanics.) Some aspects of this theory have been referred to in the preceding section; here it will be necessary to consider it in some more detail. In this theory, the state of a given system is represented by a quantity Ψ, which is a function of the co-ordinates of the system and of the time. In principle, if Ψ for a system is known, then all significant physical information concerning the system can be deduced from Ψ by carrying out certain mathematical operations on it; for example, we can predict the values of the energy, the linear momentum, or the angular momentum of the system by carrying out the right mathematical operations on Ψ. It may be that for a given Ψ, one or more of these quantities cannot be known precisely: however, appropriate mathematical operations on Ψ will in that case tell us the "spread" or "uncertainty" in the values of these quantities. Typically, when we start to solve a problem in wave mechanics, we do not know the appropriate form of Ψ for the system of interest. To find it out, we have to solve the equation which relates the dependence of the Ψ function on the space co-ordinates to its dependence on time. This equation, which applies to any Ψ whatsoever, is known as "Schrödinger's equation including the time", after its inventor. It resembles, but is not identical with, the equation that is used in classical wave theory and is

/known.....

known as the "wave equation". (The "wave equation" for waves on a string tells us how the sideways displacement of any point on the string depends on the position of the point and on the time.) Two features of the Schrödinger equation are worth special mention. First, it is a linear partial differential equation. This means that any linear combination of solutions of the equation is itself a solution; in other words, the Ψ-functions obey a "Principle of Superposition". Second, the coefficient of one of the terms in it contains the quantity $\sqrt{-1}$. This fact is important when we come to consider the nature of Ψ.

Solution of the Schrödinger equation for a system consisting of an isolated, free "particle" results in a Ψ of particularly simple mathematical form. We shall not quote this expression, but simply note two things about it, viz. that it resembles, but is not identical with, the expression for a very long train of sinusoidal waves travelling down a stretched string, and that it is "complex" in the mathematical sense; that is, that it includes the quantity $\sqrt{-1}$. (This is a consequence of the presence of this term in Schrödinger's equation.) Leaving aside the second point for the moment, we remark that the close resemblance between the Ψ for a free particle and a classical wave function and the fact that Ψ is subject to a Principle of Superposition lead one to speak of Ψ also as a "wave function". Moreover, the resemblance between the two expressions is close enough for us to be able to look at the expression for Ψ and from it deduce an expression for the "wavelength" for the free particle; this is none other than de Broglie's formula. (What this means is that wave mechanics has been so formulated as to produce the de Broglie formula where we would expect it.)

Applied to more complex situations, such as a particle bound in some way, e.g. by being located between a pair of

/rigid.....

rigid walls or being subjected to a restoring force proportional to its displacement, the theory shows that it is possible in these cases to obtain expressions for Ψ which can be written as the product of two factors, a space-coordinate-dependent factor and a time-dependent factor. This happens in the solution of classical wave problems too; the "standing waves" formed on stretched springs or strings or in air columns correspond to solutions of the classical wave equation of just this type. In the classical case the "space-dependent" factor in the wave-function gives the variation of the amplitude of the wave with distance along the wave; so in wave mechanics we speak of the space-dependent part of the "wave function" Ψ as being the amplitude of Ψ. It is usually denoted by ψ. Manipulation of Schrödinger's equation including the time so as to give "standing-wave" solutions shows that the amplitude ψ of the wave function Ψ has to obey a special equation. This equation is called Schrödinger's "time independent" or "amplitude" equation. Standing-wave solutions of Schrödinger's time-dependent equation are very important, because it turns out that such solutions correspond to stable energy states of the system in question. If the system has a Ψ which is one of the standing-wave solutions of Schrödinger's time-dependent equation, then it will be in a stable energy state from which, however, it can undergo spontaneous or induced transitions to other stable energy states. In many problems of interest in molecular, atomic and nuclear physics we wish to know the values of the energies corresponding to the stable energy states of the system, i.e. we wish to know its energy levels. So we solve the Schrödinger time-dependent equation for standing-wave Ψ's. In practice, this means solving the Schrödinger amplitude equation for the appropriate ψ's: when we do this and put in the right "boundary conditions" for the problem, we are in fact solving for the energy levels at the same time. Again

/there.....

there is a simple parallel with classical wave theory. When finding the possible wave functions for standing waves on a string stretched between two points, one has, in principle, to solve an "amplitude" equation, putting in the boundary conditions that the displacements of the end-points of the string are zero at all times. One finds, of course, that standing wave solutions are only possible for certain frequencies of the wave. This corresponds to solving Schrödinger's amplitude equation, where one finds that standing wave solutions are only possible for certain energies of the system.

As we have described it, there seems to be a great deal in common between "wave mechanics" and classical wave theory - hence the name. So far, we seem amply justified in using a "wave" language to describe the behaviour of electrons and other microscopic "particles", particularly as wave mechanics has been so fruitful a theory. But we have come no nearer to answering the fundamental question posed at the beginning of this discussion - what is the nature of these "matter waves"? In classical wave theory, the quantity corresponding to Ψ is some familiar physical quantity - the displacement of a particle on a string from its mean position, the "sound pressure" in a gas, the electric or magnetic field strength at a point, to give only some examples. What does Ψ stand for? In the very early days of the development of wave mechanics, nobody really had any clear idea. Nor did it seem to matter so much, since the theory did give the right energy levels, and this was of more pressing importance. However, the question <u>was</u> important and one which was bound to exercise the minds of workers in the field. As we saw in the previous section, Schrödinger initially proposed that $|\Psi|^2$ be interpreted as a weight function of a charge distribution in space, but this was found to be untenable. In 1926, Max Born put

/forward.....

forward a "probabilistic interpretation" of the wave function, in the course of a paper dealing with the wave-mechanical treatment of collisions between microscopic particles. When reading through the galley proofs of this paper, Born added a statement making it clear that this probability interpretation referred to Ψ^2 (strictly $|\Psi|^2$), and not to Ψ itself. According to Born, the probability of finding, say, an electron in a given small region of space was proportional to the value of the absolute square of Ψ for the electron in that region. Not only did Born's interpretation of $|\Psi|^2$ find general acceptance, but it was soon realized that it could be satisfactorily incorporated in the formalism of wave mechanics as a special case of what is sometimes known as the "expectation value" postulate.

Clearly we have still failed to give a direct statement concerning the meaning of Ψ itself, although we now have an interpretation for the absolute square of Ψ. We can dispose of the question by saying that since Born's interpretation shows us that the so-called "matter waves" are in fact "probability" waves, the "intensity" of which is proportional to the square of Ψ, Ψ itself can be regarded as a "probability amplitude". This works especially neatly for stationary states of systems, since here $|\Psi|^2$ is equal to $|\psi|^2$, i.e. the square of Ψ is literally equal to the square of ψ, the "amplitude" of the wave function Ψ. But we are still left wondering about the physical reality of Ψ, if any. In this connection we have first to face the fact that Ψ is <u>complex</u>, mathematically speaking; the quantity it represents is partly "real", partly "imaginary". There are situations in classical physics where the square root of minus one appears but raises no conceptual difficulties, because it is recognized that it is merely a convenient shorthand way of stating certain physical facts of the "reality" of which there can be no

/doubt.

doubt. (The use of $\sqrt{-1}$ in analysing alternating current circuits is a case in point.) We cannot abolish $\sqrt{-1}$ from the expression for Ψ but, as Bunge (1956) has pointed out, Ψ can be put equal to the expression $R(x,t)e^{-iS/\hbar}(x,t)$ where R and S are real functions, and substitution of this expression into Schrödinger's equation splits the latter into two real equations. This result is not in any way <u>useful</u> - the resulting equations are non-linear and do not separate the R and S parts of the original function - but it is significant in suggesting that although Ψ is complex, this does not necessarily imply that there is no underlying reality attached to it. Still, the complexity of Ψ is not to be taken lightly. The presence of $\sqrt{-1}$ in the expression for Ψ shows at once that there is an essential difference between the form of Ψ and that of a classical wave function, and it puts one on one's guard in forming too concrete a picture of "matter waves."

Reputable authors are sometimes careless about the essential "complexity" of Ψ. Thus Atkins (1965) says "...the wave function ψ... has an oscillating value at any point in space,

$$\psi = \psi_0 \cos 2\pi\nu t \text{ ",}$$

and Halliday and Resnick (1966) say "the matter wave is described ... by $\Psi = \left[\Psi_n \sin \frac{n\pi x}{\ell}\right] \cos \omega t$ ",
The cases referred to are different, but in both the error is made of writing the time-dependent part of the wave function as a real instead of a complex number. As they stand, neither of these expressions could satisfy Schrödinger's (time-dependent) equation. Orear (1961, 1967) correctly remarks in regard to "matter waves" that "even though the wave intensity is always a real positive number, sometimes the wave amplitude must be expressed as a complex number containing $\sqrt{-1}$", where by "wave amplitude" he is referring to ψ, but he fails to

/point.....

point out that Ψ as a whole must be complex.

There are other aspects of "matter waves" which reveal their very different character from "classical" waves. Although in referring to the probability interpretation of $|\Psi|^2$ above we have compared this quantity to the intensity of a classical wave, the quantity associated with matter waves which is in fact more truly analogous to "intensity" is the so-called "probability flux" - rate of flow of probability past a given cross-section - which is, in general, given by a complicated expression involving Ψ and its complex conjugate. $|\Psi|^2$ is better thought of as "probability density", being as such analogous to "energy density" in classical wave theory. These very analogies make us realize that in "matter waves" it is not energy (and momentum) that is being transported, as in classical waves, but probability.

We have noted above that for a single-"particle" system, Ψ is a function of the three space coordinates and time, as is a classical wave function. But for a two-particle system, Ψ is a function of six space coordinates (three per particle) and the time; and for a system of N particles, Ψ is a function of 3N space coordinates and the time. This alone sets Ψ dramatically apart from classical wave functions. It is important to be clear about the meaning of $|\Psi|^2$ for systems of more than one particle. We illustrate this for a two-particle system. Suppose that we are dealing with a "standing wave" state of the system. Then $|\Psi|^2 = |\psi|^2$, and the latter will be a function of the six space coordinates x_1, y_1, z_1, x_2, y_2 and z_2. If we calculate $|\psi|^2$ for a chosen set of values of these co-ordinates, representing a particular pair of points in space, then this value of $|\psi|^2$ tells us the probability of simultaneously finding one of the particles in a small region surrounding one of the points and the other particle in a similar-sized region surrounding the other point.

/We.....

We note here that there is no question of solving a two-particle problem in wave mechanics by treating the "Ψ" for the system as being a superposition of "Ψ"s for the individual particles; i.e. we cannot think of the Ψ's associated with the individual particles as interfering with each other. If the two particles interact only slightly or not at all with each other (e.g. if they are two widely-separated electrons), then there is indeed a simple relation between the "system" Ψ and the "Ψ's" for the individual particles. But the system "Ψ" in this case is not the sum but the product of the individual-particle Ψ's.*

As an interesting example of a two-particle system, we may take the hydrogen atom, consisting of a positively charged nucleus and a single electron outside the nucleus. The standard procedure in solving this problem by the methods of wave mechanics is to choose a special set of six space coordinates, three of them for the centre of mass of the system, and the other three specifying the position of the electron relative to the nucleus (or vice versa). Then the mathematics of the problem leads to the convenient solution that Ψ for the entire system can be written as the product of two "sub-Ψ's", one of them referring to the motion of the centre-of-mass of the system through space, and the other to the "internal" motion of the atom. Thus we can think of the overall $|\Psi|^2$ as telling us the probability that the centre of mass of atom will be at a particular point in space and that the electron will simultaneously be at a certain distance from the nucleus. (Hoyt (1968) has pointed out that the "sub-Ψ" which refers to the internal motion can be thought of as determining the probability of electron and nucleus simultaneously being at known distances from the centre of mass of the atom.)

/The.....

* More generally, a linear combination of such products.

The points raised above establish beyond doubt that "the ontological status of $[\Psi]$ has to be considered as something intermediate", as Jammer (1966) puts it. One could go further and discuss the nature of Ψ in relativistic wave mechanics where, instead of being merely a wave function, it becomes a "functional" (Schiff 1955), i.e. a 4-"component" column vector; this would strengthen the argument, but in the present context it is scarcely necessary to pursue the point further. But it is as well to note that the quantum theory does not have to use the mathematical language of the equations that resemble those of classical wave theory. As was guessed by Heisenberg and others even before Schrödinger invented "wave mechanics", one could use the language of matrices instead to build a "matrix mechanics" which also makes successful predictions about the behaviour of molecules, atoms, electrons and nuclei. (It was Schrödinger who first demonstrated the complete equivalence of these very different-seeming theories; "they have a common abstract skeleton, and this skeleton is the abstract vector space theory" (Wichmann 1967).) So it is possible to describe or to teach quantum mechanics without ever mentioning the phrase "matter waves" or even the word "wave". At least one textbook written in recent years adopts this approach (Kaempffer 1965); another, aimed at a somewhat lower level (second-year students at the California Institute of Technology), does not introduce the term "wave function" or the Schrödinger equation until the sixteenth chapter out of twenty-one (Feynman et al 1965).

We are now in a position to attempt some kind of answer to the questions raised at the beginning of this section. First, there are a number of correspondences between the waves familiar from classical physics and "matter waves". The latter are represented by functions of the space co-ordinates and the time; these functions satisfy a linear partial differential equation; "free particle" wave functions resemble those

/for.....

for travelling, plane "classical" waves and can have a wavelength associated with them; and "standing-wave" solutions of the matter-wave equation can be obtained. On the other hand, there are a number of significant differences between classical waves and "matter waves". The Schrödinger equation is not identical with the classical wave equation; because the Schrödinger equation includes a term containing $\sqrt{-1}$ in it, the "matter wave" function Ψ is necessarily complex; for systems of N particles, Ψ is a function of 3N space coordinates in addition to the time; and in "matter waves" it is probability and not energy which is transmitted.

To the question of whether "matter waves" are really waves, we must answer that it depends on what one means by the latter. If under the heading of "waves" we are prepared to admit "disturbances", whatever they are, which conform to Schrödinger's equation and are propagated in 3N-dimensional space where N can be greater than unity, then "matter waves" are waves! Perhaps, however, it is more sensible simply to note that "There exists only a limited formal analogy between Ψ and a classical wave field" (Bunge 1967), and, instead of trying to widen the concept of wave to embrace the Ψ function, to stress the limitations of the classical wave concept as applied to Ψ. The fact that Ψ can take on a different guise in other representations of quantum mechanics lends support to the attitude that the phrase "matter wave" is only a manner of speaking and should not be taken too literally.

Finally, what of the "physical reality" of Ψ ? $|\Psi|^2$ is directly measurable, at least in principle, and so one need not hesitate to confer physical reality on it. Enough has been said to show that Ψ is much more elusive. In our view, the key to its "reality" lies in the central role it plays in quantum mechanics, in whatever representation. In this theory Ψ represents the state of a system and is linked to measurable properties of the system by means of various

/mathematical.....

mathematical relations that form part of a theory. Fundamentally, it is Ψ which is "subject to the laws of nature" (Margenau 1950). At the very least, it seems reasonable to accord some kind of reality to a concept which has this kind of status within a physical theory. Ψ cannot be measured and is difficult to pin down in words other than in such phrases as "state function" or "state vector", so that one can agree with Jammer (1966) that its "ontological status" is "something intermediate". But the conclusion seems unavoidable that "the $[\Psi]$ states of modern physics must be accepted as a valid part of physical reality" (Margenau 1950).

10.2.5 The "Uncertainty Principle"

The teacher who is moved by interest and necessity to find out more about the quantum theory will inevitably encounter the Heisenberg Uncertainty Principle in his reading. As the Principle is of considerable importance for the light it sheds on certain aspects of the quantum theory, and as the account of it that the teacher may read may be misleading, we consider it here.

As applied to the position and momentum of a "particle", the Uncertainty Principle can be stated symbolically in the form $\Delta p \, \Delta q \geqslant \frac{\hbar}{2}$, where Δp and Δq represent the "uncertainties" in the momentum and in the position co-ordinate respectively, and $\hbar$ is Planck's constant divided by 2π. The meaning of the term "uncertainty" is crucial to the interpretation of the Principle and requires precise definition. In the words of Margenau and Cohen (1967): "[Δp and Δq respectively] represent the standard deviations, or the square roots of the variances of sets of repeated observations of p and q when the system (loosely speaking, the particle) is in an arbitrary but definite quantum state ψ". To test the Principle, we should have repeatedly to "prepare" the system (or a

/sequence.....

sequence of identical systems) so that it is in the specified state ψ, and to carry out the required measurements of p or q after each preparation of the state. Sometimes the "preparation" is explicitly mentioned in the statement of the principle: "There exists no preparation capable of producing an ensemble [i.e. a set of systems of the same kind prepared in the same way] with both an x-component of position (for any x') within an uncertainty Δx and an x component of momentum (for any p'_x) within an uncertainty Δp_x if the product $\Delta x \Delta p_x$ is less than a quantity of the order of Planck's constant, h" (Eisenbud 1966).

The Principle can be deduced from the fundamental postulates of quantum mechanics; the derivation is straightforward and can be found in many standard texts (e.g. Lindsay and Margenau 1936, Bohm 1951, Schiff 1955). In simple terms, we may say that the Principle is a consequence of the fact that the statistical behaviour of microsystems is describable in "wave-like" terms. An electron with a well-defined momentum and constrained to move along the x-axis only, has a well-defined de Broglie wavelength. But a wavetrain of absolutely definite wavelength must have an infinite spatial extent, i.e. the Ψ-function extends throughout all space with the same amplitude. Therefore the value of $|\Psi|^2$ is the same at all points along the x-axis, and according to the probability interpretation of $|\Psi|^2$, this means that the probability of finding the electron in any one small segment of the x-axis is the same as that of finding it in any other similarly-sized segment. This in turn means that if the electron were repeatedly prepared to be in this state, successive measurements of its position would yield an infinitely large spread, while successive measurements of its momentum

/would always .....

would always yield the same value. By altering the state of the electron to one in which the Ψ function corresponds to a sharply-peaked wave packet, one attains a situation in which the position of the electron can be quite sharply specified, since now the probability of finding it within a small segment of the x-axis only has a nonzero value over a limited region of the x-axis. But a wave packet can only be obtained by superposition of a set of wavetrains corresponding to a range of wavelengths, so that the momentum of the electron is no longer well-defined. Hence an increase in sharpness of the one variable or "observable" is obtained at the expense of reduced sharpness in the other.

The foregoing account shows that the Uncertainty Principle is built into the foundations of the quantum theory. We do not have to appeal to particular experimental situations or to an apparatus-object interaction in order to derive the Principle. In popular or semi-popular accounts the matter is often represented differently, not to say misleadingly. (Cf Peierls (1964): "This principle expresses the fact that any observation on a physical system must necessarily involve its interaction with a measuring apparatus, and this interaction may cause a change in the state of the system under observation. Whereas in classical physics.... one can in principle... reduce the disturbance of the system so as to make it unimportant, the occurrence of the quantum of action in atomic physics limits the extent to which the interference can be reduced.") This approach to the principle is often supported by an account of Heisenberg's celebrated "thought experiment" in which a gamma-ray microscope is used to make simultaneous measurements of the position and momentum of the electron. To make the resolving power of the microscope large and so minimize the uncertainty in the position of the electron, short-wavelength gamma-rays are used. Suppose one photon is used to detect the position of the electron.

/When.....

When it is scattered by the electron, it imparts some recoil momentum to it; the more energetic it is, the larger the recoil momentum. According to the fundamental ideas of the theory, one cannot say anything definite about the "path" followed by the photon from the electron to the screen on which the image of the electron is to be formed. Therefore one cannot know how much recoil momentum the electron receives, because the angle through which the photon is scattered is uncertain by the amount of the aperture angle of the lens. Calculation of the uncertainty in position resulting from the limitation on the resolving power of the microscope set by the wavelength of the radiation used, and of the uncertainty in momentum set by the momentum of the photon and the aperture angle of the microscope, yields a pair of uncertainties whose product is of the order of magnitude of h.

The analysis of this thought experiment is useful in illustrating the difficulties that would be encountered in carrying out measurements at the quantum level, and in showing the association of Planck's constant with the "errors" in simultaneous measurements of position and momentum. Unfortunately it is seriously misleading in a number of respects, as various critics, notably Bunge (1967) and Margenau and Cohen (1967) have pointed out. The criticisms made by the latter pair of authors are worth quoting in some detail. First, it is pointed out that while the "Δx" inferred from the limitation of the resolving power of the microscope refers to the position of the electron before the measurement, the "Δp" which results from the disturbance of the electron by an incident photon refers to the state of the electron after the measurement. It terms of the fundamental form of the Uncertainty Principle, the Δp and Δx are both associated with the same state of the system in question. Second, nearly all of the analysis of the gamma-ray microscope experiment is conducted in strictly classical terms. For example, the calculation of the uncertainty in position is based entirely

/on

on classical wave theory, and the interaction between the incident photon and the electron is also regarded in essentially classical terms. It is only in regard to the passage of the scattered photon from electron to screen that classical reasoning is abandoned. But in this step, one is making use of the very ideas from which the Principle can be derived in its fundamental form. Thus it cannot be held that the Principle is a consequence of the "unpredictable and uncontrollable" effect of the interaction between the measuring instrument and the system with which it interacts: the situation is really the other way round. "... The uncertainty principle entails that the results of the thought experiment are relatively uncertain, but is not entailed by it" (Hesse 1961). Third, the "disturbance theory" (as Margenau and Cohen call it) in its naive form, in which largely classical reasoning is supposedly used to "derive" the Uncertainty Principle, violates logic: "Quantum mechanics is a wider discipline than classical physics... Hence [classical physics] cannot possibly contain [quantum mechanics] since its range is smaller, and one cannot use classical reasoning to derive the uncertainty principle" (Margenau and Cohen 1967). Finally, the Δp and Δx have no meaning with respect to a single measurement. "... Thus unless the gamma-ray microscope story is extended to embrace a great number of interactions between a particle and a photon and Δ is understood to refer to the statistical deviation of the results, the disturbance hypothesis is without relevance" (Margenau and Cohen 1967).

To these criticisms we wish to add a comment of our own. Virtually all commentators on the quantum theory, including those who evidently favour the "disturbance" explanation of the Uncertainty Principle, are concerned to stress that the Principle does not mean that the electron really has a simultaneously well-defined position and momentum, only one is

/unable.....

unable to measure them, but rather that the electron "simply does not have an exact simultaneous position and momentum" (Eisenbud 1966). The student who is introduced to the Uncertainty Principle via the "disturbance theory" and the gamma-ray microscope experiment may well have difficulty in accepting this. In order to imagine the measurement process "disturbing" an electron on which the measurement is being made, he is likely to picture the electron as simultaneously having a precise position and a precise momentum. He can justify this by referring to the gamma-ray microscope experiment. In the usual analysis of this experiment, the electron seems to be represented throughout as having a definite position and momentum; the position cannot be measured exactly because of the limitation on the resolving power of the microscope, and the momentum cannot be measured exactly because the incident photon gives the electron an impulse which is uncertain because one cannot know exactly where the photon goes afterwards. The positivist-operationalist retort to the point we are making here would probably be that "..... since position and momentum cannot be measured simultaneously and exactly, we cannot prove that the electron can have both a precise position and a precise momentum" (Eisenbud 1966). Since we cannot prove this, it is operationally meaningless to associate a precise position and momentum with an electron, and no more need be said. In our view, this line of argument misses the point. Undeniably, it follows from the fundamental assumptions of the quantum theory that entities like electrons cannot simultaneously have precisely defined positions and momenta. In other words, this is a consequence of the nature of these entities and not of the measurement operations we may or may not carry out. Whether or not physicists expound the Uncertainty Principle in this way, they certainly

/use.....

use it in this way. For instance, it is common (and unexceptionable) practice to use the Principle in arguments for the existence of a zero-point energy in bound systems, and to provide rough estimates of the lowest permitted energy of such systems as the particle in the box and the hydrogen atom (cf Feynman et al 1965). In such arguments or calculations, the "uncertainties" in position and momentum are never related to any real or imaginary measurement operations which disturb the system in question: the systems are "undisturbed", and the "uncertainties" are treated as part of their nature. To present the Uncertainty Principle in a way which virtually compels the beginner first to think of the electron classically and then requires him to banish this classical image on essentially operationalist grounds is to put the emphasis in altogether the wrong place, since it obscures the "built-in" character of the "uncertainties."

We think that the meaning and origin of the Uncertainty Principle are well summed up in the following statement by Wichmann (1967): "The uncertainty relations reflect experimentally discovered facts about nature [e.g. the possession by an electron of both wave- and particle-like properties]. The particles occurring in nature are not like classical particles; they behave quite differently, and that is why certain kinds of measurement cannot be made or even imagined."

We trust that the discussion above has made clear that nowever tempting it might be to present certain consequences of quantum theory by way of thought experiments leading to the Uncertainty Principle, this material is best kept out of the elementary class. Nor would it be possible in such a class to try to derive the Principle in a more fundamental way, say by using the "wave-packet" approach mentioned above; the pupil's knowledge of classical wave theory is nothing like sophisticated enough to enable him to understand this kind of derivation. Fortunately, there is no desperate need for the

/Principle.....

Principle in an introductory course. The only purpose that its introduction would serve would be to point up the fact that a quantum-mechanical microsystem does not have a simultaneously well-defined position and momentum, so that the concept of an "orbit" or "trajectory" cannot apply to it; and the last-mentioned point can be brought out sufficiently well by referring to the wave-particle "duality" of matter and radiation and especially to the fact that the "wave intensity" gives only the (relative) probability of finding an electron or photon in a given region of space. (Significantly, the P.S.S.C. course, which devotes considerable attention to elementary quantum theory, makes no mention of the Uncertainty Principle.)

10.2.6 Implications of the foregoing discussion for the elementary teaching of quantum mechanics

Before de Broglie published his famous paper in which he suggested that "phase waves" should be associated with matter, W.L. Bragg remarked to G.P. Thomson that he thought that the electron was not so simple as it looked (Thomson 1968). In due course, Thomson was to play his part in proving that the electron was indeed "not so simple as it looked." In the forty years that have since elapsed, not only has this idea become generally accepted, but it has also become accepted in certain quarters that physics teachers should convey to their pupils the essence of the new picture of the electron and of matter in general. The P.S.S.C. physics course includes a substantial introductory treatment of this topic, Year V of the Nuffield Physics Project concludes with a brief qualitative treatment of it, and we may confidently look forward to its inclusion in more and more senior school syllabuses. One's attitude toward teaching it is inevitably linked to one's attitudes towards the issues discussed above in Sections

/10.....

10.2.1-5, so that the implications of that discussion for elementary teaching must now be considered.

What should the teacher say to pupils about the so-called "wave-particle duality"? According to the Nuffield Guide (V), "We certainly should not burden our young pupils with a puzzling description of the concept of complementarity. Yet that is a very powerful new idea that emerged from our having to accept both wave behaviour and particle behaviour - mutually exclusive according to our choice of experiment - and teachers may find it worth their while to read an account of it for their own interest." We would agree with the first sentence of this quotation, and also with the recommendation contained in the second sentence, provided that teachers were not encouraged to read accounts of only the "complementarity" view of quantum theory. As we have seen, this "very powerful new idea" is not immune to criticism. As to positive advice, the Nuffield Guide authors evidently feel that one's teaching should be directed towards making pupils "learn to live with" the idea that both wave and particle models of light and matter have to be used in appropriate circumstances: "We should stop trying to decide which, and start trying to learn to use both, or rather either according to the kind of experiment we are doing... To help pupils to understand that more than one model may be kept in fruitful use, to persuade them not to ask too strongly 'which model is true; which is right?' we show them two maps of London. One is the usual stylized Underground map in several colours. The other is a map of streets with the Underground lines marked on it too. We ask which is the right one; and we are likely to receive 'It depends on what you want it for'."

Several points call for comment here. First, it seems to us appropriate that in teaching this topic, emphasis should

/be.....

be placed on the use of models and on recognition of the circumstances in which they are best used. Pupils cannot turn to the full-scale quantum theory for a prediction of how light or matter is to behave in a given experimental situation. They have instead to depend on a drastically simplified version of that theory, in which use is made of wave and particle models in order to make the underlying ideas intelligible in macroscopic terms. There can be no objection to this provided that these models are treated as no more than aids to vizualization and simple quantitative prediction, and to what Bunge (1967) calls "descriptions of macroevents - e.g. measurements." Consider an experiment in which light passes through a pair of slits and then falls on a metallic surface in a vacuum. To predict the intensity distribution on the metal surface, one can visualize light as a wave motion in the process of travelling from the source to the metal surface by way of the slits. To predict the maximum kinetic energy of the photoelectrons ejected from the metal surface, one can visualize light as sharply localized particles, each carrying energy hν. To take the wave and particle models literally would mean that light is a wave motion while on its way from source to metal surface and changes over to being a collection of particles when it reaches the surface. Not even the most devoted "complementarist" would have one regard the situation in this way: For this reason we think it dangerous to say as the Nuffield Guide (V) does of electrons and photons, that "Each has particle properties in some circumstances, and each has wave properties in other circumstances." This could easily be misinterpreted by the teacher unfamiliar with quantum theory to mean that in the experiment just described, light and matter do carry out an actual transformation from one character to the other. Such a teacher might be able to convince himself that this kind of thing could happen;

/he.....

he would have great difficulty in convincing his pupils!

Hesse (1962), in defending the use of wave and particle models in thinking about quantum theory, has argued that professional physicists still take these models seriously, whatever they may say about them, and that these models "are still essential, supplemented by the hunches that physicists have acquired about when to argue in terms of one and when the other." They are still essential because "if we were forbidden to talk in terms of models at all, we should have no expectations at all [about the behaviour of matter in new situations], and we should then be imprisoned forever inside the range of our existing experiments." This is putting it a little too strongly. Physicists do after all have an established theory to guide them in this connection, so that their "expectations" do not have to be guided by thinking in terms of models, although this may prove helpful.

The remaining comment we wish to make on the Nuffield Guide's (V) advice on the teaching of wave- particle duality concerns the use of two different kinds of map of London to illustrate the fruitful use of different models of matter and radiation. If used with due care, this is a helpful analogy, not only for its obvious suitability for its intended purpose, but because the maps are both maps of London: neither gives full information about London, but each conveys some correct and valuable information about something that is there all the time. The analogy should not be pressed too far: we cannot synthesize wave and particle properties into a single, neat, visualizable model as we can synthesize a map of London's streets and a map of its railway lines, rivers, canals and Underground into a "complete" map of London, and teachers should take care to point this out. If the attitude we have been adopting towards wave-particle duality is justified, then teachers should also not encourage their pupils to

/believe.....

believe that the wave (or particle) model of light (or matter) is "right" and the other "wrong" in a given set of circumstances; neither model is ever wholly "right", but one or the other may sometimes be more helpful.

The P.S.S.C. course gives an introductory treatment of wave-particle duality which seems to us to exemplify the correct approach at this level. On the one hand, there is as complete a survey of the experimental background as is possible in the circumstances; on the other hand, there is no attempt to suggest that the wave/particle properties of light and matter are somehow "produced" by the experimental set-up or depend essentially upon the interaction between the "microsystem" in question and the experimental set-up. Concerning photons, it is pointed out that as the frequency of the radiation goes up, its wave aspects become more difficult to demonstrate because of the shorter wavelength, while the particle aspects become more readily apprehensible, because of the greater energy associated with an individual photon; and concerning electrons, a corresponding discussion is given of the conditions under which the wave-like aspects of electron, proton and neutron behaviour are most likely to be apparent. The implication behind these discussions seems to be that light and matter are always of the same character whatever the experimental circumstances are, but that in a particular experimental arrangement it might be easier to show up one aspect of this fundamental behaviour than another. The nature of light is summed up like this: "...the real nature of light is more subtle. It shows us both wave and photon behaviour. We have therefore been forced to make a new picture of light, one which combines some aspects of wave benavio(u)r as we previously understood it with some aspects of photon or particle behaviour... Newton's corpuscular picture of light is by no means restored by the photon. It is

/no.....

no more true than the pure electromagnetic picture, with its waves of continuously distributed energy. Light is to be understood fully only by a novel scheme of the kind we have tried to outline". Similar statements are made concerning the nature of the electron: "Once again we have found that the atomic world cannot be fully described by the simple models of the older physics.... The waves determine the probable location of the particles, and the motion of wave trains determines the probable motion of the particles...Inside the atom we must be prepared to use a new version of mechanics which combines parts of Newton's particle mechanics with the effects of the wave nature of matter." These statements represent how far it is necessary or indeed possible to go in trying to summarize and synthesize the "duality" of light and matter for the inquiring pupil. Presumably hard-core "complementarists" would deplore the failure of these statements to draw attention to the alleged indivisibility of atomic object and apparatus. But even they might be prepared to agree that a description of complementarity might be found "puzzling", as the Nuffield Guide puts it; that a beginning pupil might find it easier to accept that "quantons" had intrinsic properties or attributes, and that introduction to quantum theory from this particular standpoint does not commit the pupil to continued acceptance of it in the years to come.

It might seem that what has been said above about modern quantum theory constitutes an argument against the teaching of the "duality" of electrons in an elementary course; if the matter is as subtle as this, surely one cannot hope to convey its essence successfully to beginners. This takes us back to our set of "realistic teaching objectives" (Section 10.1.1), which we might summarize by saying that the general aim of teaching pupils about electrons is not fully to explain their

/nature.....

nature - that is probably beyond anybody at any level - but to acquaint pupils with some of the basic experimental facts of electron behaviour, and to give some indication of how these facts can be accommodated in a modern theory which, in turn, can give some explanation of the behaviour of atoms, molecules and nuclei. Since this is 1969 and not 1899, it would be dishonest to present any survey of the facts about the electron which only mentioned particle-like aspects. The electron is "not simple", and it is no use pretending otherwise. The sooner one is confronted with this fact, the sooner one can come to terms with it.

10.2.7 A possible teaching approach to the nature of the electron

Historically, study of the wave-like attributes of matter began on the theoretical side, in particular with de Broglie's "matter wave" hypothesis; the observation of electron diffraction by Davisson and Germer in the U.S.A. and G.P. Thomson in Great Britain did not follow until some three years later (Jammer 1966, Trigg 1966c and Thomson 1968). In the meantime, prompted by the publication of de Broglie's ideas, Schrödinger had developed his theory of "wave mechanics", which he applied successfully to the solution of such problems of interest to "microscopic" physics as the energy levels of the hydrogen atom, the linear harmonic oscillator and the rigid rotator (Jammer 1966). This is a satisfactory teaching sequence to follow (though at an elementary level it must necessarily omit Schrödinger's work), i.e. to begin with de Broglie's hypothesis and then to consider the experimental evidence in support of it. Surprisingly, the Common Basic Syllabuses omit mention of de Broglie's hypothesis (the Comments issued with the drafts make it plain that the omission was intentional); we think that even though one might not mention de Broglie's formula, the idea behind it should

/certainly.....

certainly form the springboard for the teaching of this topic. The formula itself is so simple that we can think of no good reason for its exclusion.

De Broglie arrived at his celebrated formula by way of a subtle relativistic argument which could not be given in an elementary course and which is today of only historical interest; in modern treatments of quantum mechanics de Broglie's formula is derived from the fundamental postulates, rather than being regarded as one of the fundamental postulates. In any case, de Broglie's formula was essentially the result of an inspired guess, and any attempts (including de Broglie's own) to "derive" it from bases other than the fundamental postulates of quantum mechanics must be regarded as no more than plausibility arguments. Therefore teachers or syllabus planners need go to no great trouble to try to find a simple "proof" of de Broglie's formula, since no such thing exists.

We suggest that de Broglie's ideas can be introduced by saying that in 1924 he put forward the suggestion that just as photons exhibit both wave-like and particle-like behaviour, so might "material" particles, such as electrons. To make this idea quantitative, de Broglie proposed that an electron (say) having momentum p has associated with it a wavelength given by the formula $\lambda = \frac{h}{p}$. This formula should be "justified" by showing that it has the same form as the corresponding formula for the photon. To do this, one needs the formula for the photon momentum, viz. $p = \frac{h\nu}{c}$. This can be stated to be an experimentally-derived relationship (e.g. from measurements on the Compton effect or on the "radiation pressure" of light - cf. Section 8.12). Since the wavelength associated with a photon is given by $\frac{c}{\nu}$, it follows directly from the expression for the momentum of a photon that we may write $p = \frac{h}{\lambda}$, which is the desired formula.

/For....

For the moment, the pupil is not concerned with the use of the de Broglie formula to build a theory capable of predicting the energy levels of atoms and molecules. Instead, he must be given some account of how direct evidence for de Broglie's ideas has been obtained. The P.S.S.C. film on "Matter Waves" provides the best solution to this problem. If this is not available, or even if it is, pupils should see photographs of diffraction and interference patterns obtained with electrons and other material particles and where possible compare the appearance of these patterns with corresponding optical or X-ray diffraction or interference patterns. In the P.S.S.C. textbook (both first and second editions) photographs of the optical interference pattern obtained with Fresnel's biprism and a corresponding electron interference pattern are shown; a similar comparison is made between the optical and electron diffraction patterns of a straight-edge. The photographs of these electron diffraction and interference patterns were taken with an electron microscope, that is, with a device whose principle of operation is based on the wave-like aspects of electron behaviour. To this extent it can be claimed that the use of these photographs is "cheating". However, in this instance the "cheating" can readily be condoned. We can regard the electron microscope in this context as a "black box" which, in the realm of operation of powerful optical microscopes, is known to give us the same pattern of images as the latter. So we trust it to behave like a microscope when it is used to examine objects in which the detail is too fine to be resolved by optical microscopes. Certainly it should be mentioned to pupils in due course that the principle of the electron microscope is based on the wave-like properties of the electron: since this instrument is nowadays an important research tool in industry and in the physical

/and.....

and biological sciences, it is an important example of the practical use to which "matter waves" can be put.

Besides seeing these photographs and, if possible, seeing the recommended film, pupils should also learn, perhaps through a reading assignment, of the classic early experiments on electron diffraction (cf Trigg 1966c, Thomson 1968). This will bring home to them not merely that electrons can form diffraction and interference patterns, but that the quantitative aspects of these patterns can be predicted from de Broglie's relation $\lambda = \frac{h}{p}$. Perhaps also as part of a reading assignment, pupils should find out that corresponding experiments have been successfully carried out with other microscopic "particles" such as neutrons, protons, atoms and molecules.

The cost of commercially-available electron diffraction apparatus for elementary teaching puts it beyond the budgets of South African schools at the moment. Since demonstrations of this kind can have only qualitative value at this level, it does not seem that a strong case can be made out for asking the authorities to spend a great deal of money on supplying schools with the appropriate apparatus.

A question which must be taken up with pupils is our awareness of the "waviness" of particles. One can start by referring to a specific (and in this case familiar) example, such as the electron beam in a cathode-ray tube of whatever kind. Such a beam, we may well say, does not behave much like a wave: it behaves remarkably like a stream of particles, as earlier experiments were indeed intended to show. Calculation of the wavelength of the electron in a cathode-ray tube shows that it is of the orderof 0.1nm (for an electron accelerated through a p.d. of 100 volts). This is many orders of magnitude smaller than the dimensions of the cathode-ray tube or its components, so that the electron beam encounters nothing inside the cathode-ray tube which could possibly

/show.....

show up its wavelike aspects. To reveal these we should have to arrange for the electron beam to impinge upon a structure which does have the right special features, such as a "periodicity" of the same order of magnitude as the electron wavelength. In general, whether or not we can notice a particle's "waviness" will depend on its wavelength, and that is determined by its mass and its velocity. Teachers should discuss with pupils numerical examples of the wavelengths of particles (from electrons to cricket balls) having known masses and velocities.

Lest there should be any possibility of confusion, the question of the differences as well as the similarities between photons and electrons must also be discussed (cf our teaching objective (iv) in Section 10.1.1.). For the pupil's benefit, most of the points that have to be made are quite simple. The electron has a nonzero rest mass; the photon has zero rest mass. The electron carries a negative electric charge; the photon carries none. (This provides us with a ready experimental means of distinguishing between electromagnetic and other forms of radiation; the former is undeflected by electric and magnetic fields although, as Einstein was the first to point out, it may be and possibly is deflected by a gravitational field of sufficient strength.) In a sense, this is all within the pupils' experience. Another point that he should be capable of understanding, although it will not be possible to offer him direct proof of it, refers to the ease with photons can be created (cf. P.S.S.C. 1960, 1965 and Teachers' Resource Book and Guide, Part Four). Photons can be created by striking a match or switching on a light, whereas the creation of new electrons, protons, neutrons or other material "particles" of the microscopic world needs large amounts of energy and rather special circumstances.

/Further,.....

Further, in "ordinary situations" (the P.S.S.C.'s phrase) the total number of electrons or protons is conserved, whereas the number of photons in an enclosure containing radiant energy is not necessarily constant - heating the walls changes the enclosed number of photons.

(A further point for the teacher rather than his pupils to note is that there is a difference between the intrinsic angular momentum of the electron and that of the photon. The quantum number characterizing this property has the value of one half for the electron and unity for the photon.)

Finally, the teacher will have to try to combine the particle-like and wave-like aspects of the electron in a single description. To point the way to this, he can refer to a hypothetical experiment in which electrons emitted from an electron gun pass through a pair of parallel slits in a screen and strike another "screen" on the other side. The electrons reaching the further screen distribute themselves on it in an interference pattern of bright and dark "fringes"; the separation of these fringes is correctly calculated using de Broglie's formula for the wavelength of a particle. There is no theory that can predict where any given electron will go, but the "electron wave" theory correctly predicts the probability of finding the electron in a given region of the screen; electrons are most likely to go to the centres of the bright fringes, and will not go at all to the centres of the dark ones.

Sometimes an experiment like this is explained with the aid of such statements as "... wave patterns guide the particles to interference patterns" (Nuffield Guide V). We would not say that the statement is wholly incorrect - many would know what is really meant by it - but it is easily misinterpreted. To the unsophisticated it would suggest that electrons and other microscopic entities are really

/particles.....

particles (in the ordinary sense) and retain their identity as such throughout the process of formation of interference patterns. As we saw in Section 10.2.3, this interpretation of quantum theory is difficult to sustain. The "guiding wave" idea also gives the misleading impression that there really are <u>waves</u> which are somehow separate from electrons or other microscopic entities but which direct these entities where to go. More correctly, the "waves" represent an <u>aspect</u> of these entities, in that our description of these entities is incomplete without them. Picturesque language such as is used in the above-quoted statement may help in the description of this experiment, but the teacher should if he uses it hasten to add that there is really "something about" each electron that goes through both slits at once. On the other hand, when the electron is detected, it is detected <u>whole</u>, so that this particle-like aspect of the electron is always there too.

The teacher can next suggest that since electrons and other microscopic entities have both "particle-like" and "wave-like" properties, they can be neither particles nor waves. But in that case one must be careful not to attribute properties to the electron (say) which one cannot be sure that it really has. An electron, because of its nature, does not simultaneously have a well-defined position and momentum; therefore one is not entitled to think of it as executing a definite orbit or path. Pictures showing electrons executing circular or elliptical orbits around atomic nuclei are entirely fictitious - there is no experimental evidence which supports the idea of electronic orbits. Pupils may be prompted to comment here, "What about the electron paths in a fine-beam tube?" In reply, one can point out that the path of an electron beam in a fine-beam tube is very coarse and ill-defined,

/microscopically.....

microscopically speaking, whereas if we suppose an electron to be executing a well-defined orbit in an atom, we are assuming that at any instant both its position and its momentum are determined to a very high order of precision, many orders of magnitude higher than that of these quantities for the electrons in the "beam" of the fine-beam tube.

In an elementary account of wave mechanics that is heavily influenced by the "disturbance theory" (cf. Section 10.2.5), Bennet (1965, 1968) refers to "the inevitable uncertainty about the path followed by an individual photon" and remarks that "For a single electron all we can state is the probability of its following a given path." Inasmuch as they suggest that photons and electrons do follow paths, these statements are misleading; the term "path" would appear to have no meaning for electrons and photons. That such statements can appear in a recently- (and carefully-) written elementary text is an indication of how easy it is to lapse into error when trying to express the ideas of quantum mechanics in simple terms, and it must be accepted that errors like this are going to be made in the elementary teaching of the subject, at least to begin with. But adherence to the principles and facts outlined above can prevent the teacher from falling into grosser errors, and perseverance and willingness to grapple with these challenging ideas will surely bring their reward in the end.

10.3 The atom and its nucleus

10.3.1 Introduction It is fitting that the final section of a school syllabus in physics should be devoted to a study of the properties of the atom and its nucleus. Up to this point, pupils will not have been officially encouraged to "believe in" the existence of atoms and molecules. Some evidence, most of it qualitative, has been presented in favour of the

/existence.....

existence of these entities, and use has been made of them in building models and theories to explain various features of the behaviour of matter. But from here onwards, the existence of atoms and molecules must be taken for granted, and the emphasis is to be on the investigation of the structure of the atom and its nucleus with the use of the tools of the physicist.

When we say that the existence of atoms is now to be taken for granted, we do not mean that pupils should now be asked to accept this as a dogma. This would indeed be a sad lapse from the approach that has characterized the preceding sections. The point is that the quantitative evidence for the existence of atoms comes mainly from chemistry; long before physicists measured atomic (or rather ionic) masses directly, chemists had been successfully interpreting the results of their measurements in terms of the atomic hypothesis, using the additive property of mass, and had been able to draw up a table of so-called "atomic weights" (relative atomic masses). Although this evidence was not absolutely convincing - it did not, for example, convince Mach and Ostwald - it was strongly suggestive, and subsequent evidence obtained by physicists was able to put the issue beyond doubt. We do not think it is the business of the physics teacher to teach his pupils chemistry as part of a physics course, but we do feel that if the pupil has not had the opportunity of studying the chemical evidence elsewhere, it should be made available to him, perhaps in the form of a summary which could be duplicated, distributed and made required reading.

The following order of presentation seems a reasonable one:-

(i) The mass of an atom;

(ii) the size of an atom (preliminary discussion);

(iii) the unstable atom - radioactivity, the nature of

/radiations.....

radiations from unstable atoms and methods for studying these;
(iv) the discovery of the atomic nucleus;
(v) the structure of the atom;
(vi) the structure of the nucleus (including the discovery of the neutron);
(vii) the bombardment of nuclei - nuclear reactions;
(viii) miscellaneous related topics, e.g. radioisotopes, fundamental particles and cosmic rays.
This is not the only possible teaching order, but it has the merit of being a logical development, while at the same time it is likely to prove at least as "exciting" to teacher and pupil as any other order.

10.3.2 The mass of an atom

The direct determination of atomic masses is of interest not only in itself but also because of what is revealed by the data obtained in this way. While average values of relative atomic masses can be determined with great accuracy by chemical methods, it is the mass-spectrographic determinations of atomic masses which unmistakably reveal the existence of isotopes and which show that the bulk of an atom's mass is associated with its positively-charged component.

There is no need for a teacher to dwell on the historical aspects of the subject. The pupil is already familiar with the principles of the measurement of the ratio of charge to mass for the electron, and precisely the same principles apply to the operation of modern mass spectrographs or spectrometers. Accordingly J.J. Thomson's method of obtaining positive-ray parabolas need not be discussed, nor is it necessary to describe Aston's original mass spectrograph. Warren (1965) has complained that "Except in very advanced books it is rare to find any mention of modern instruments such as the mass spectrometers pioneered by Nier with their wide range of applications

/in.....

in physics and chemistry, and yet these instruments are no more difficult to understand than the older ones." While we might agree that this is a fair criticism of the presentation of this topic in an A-level syllabus, for which there might be a case for the inclusion of up-to-date techniques of mass spectrometry, we do not think that one need go as far as this in the course under consideration here. It is enough for pupils to know the basic principles of operation of the mass spectrometer, and these can be conveniently explained with reference to the instruments of Dempster or Bainbridge. The pupil interested in more modern instruments can be given suitable material for reference.

Since in all mass spectrometers or spectrographs one is actually measuring the charge-to-mass ratio of positive ions, the pupil should have some understanding of why particular values of the charge are chosen when the ionic masses are calculated. How do we know, for example, that when the charge-to-mass ratio for hydrogen ions is 9.58×10^7 C/kg, we must assume that the charge on each ion is one electronic charge and not two or more? What enables us to make similar decisions with equal confidence in regard to other positive ions? An essential element in the background to such decisions is the belief that electrons are constituents of all matter, or at any rate of matter as we commonly encounter it. At some stage the evidence for this belief must be reviewed for the pupil: the value of $\frac{q}{m}$ is the same for the "particles" emitted photoelectrically from all sorts of metals. It is with some confidence, then, that one can assert that when an atom or a molecule is turned into a positive ion by losing some of its negative charge, it loses one or more electrons. Millikan's experiment supports the view that the process of ionization amounts to a loss or gain of electrons.

/Turning.....

Turning to the hydrogen ion, we can say that its charge should be an integral multiple of the electronic charge. In Thomson's original experiments, he found that he obtained two values for $\frac{q}{m}$ for hydrogen ions - the one quoted above and one equal to exactly half this value. He interpreted the former value as corresponding to a hydrogen atom that had lost an electron, and the latter as corresponding to a "singly-ionized" hydrogen molecule. Support for these interpretations comes from other evidence. $\frac{q}{m}$ for the hydrogen ions which take part in electrolysis is also 9.58×10^7 C/kg, and the context in which this result is obtained suggests that the hydrogen ions which take part in electrolysis are ionized hydrogen atoms. Thomson did not find any indication of values of $\frac{q}{m}$ for hydrogen equal to twice 9.58×10^7 C/kg, which would have corresponded to a hydrogen atom carrying double the charge of the ions observed in electrolysis. On these grounds it was reasonable to suppose that an ionized hydrogen atom had lost one electron, and might indeed have only one electron to lose. (These facts are lucidly discussed by Rogers (1960).) The value of the mass of an ionized hydrogen atom could then be calculated on the assumption that it carried a charge equal to the electronic charge; this yields a value of the mass equal to 1.67×10^{-27} kg. (There is independent evidence that this result is correct; if one uses a value of Avogadro's number obtained from "kinetic theory" measurements, one can combine this with the mass of a gram-atom of hydrogen to find the mass of a hydrogen atom. The answer agrees with that obtained from mass spectrometry to within the limits of experimental error.)

The result for hydrogen immediately establishes that much the greater part of the mass of an atom is associated with its positively-charged portion. It also makes the task much easier of interpreting the results for the ions of heavier elements, since the average relative atomic masses of the

/elements.....

elements are known from chemistry.

Finally, the mass-spectrometric evidence for the existence of isotopes should be discussed, using neon and/or chlorine as examples, and the pupils' attention drawn to the fact that when we allow for the existence of isotopes and compute relative atomic masses accordingly, all atoms have relative masses that are very nearly but never exactly (with the exception of ^{12}C!) equal to integral multiples of the "atomic mass unit". The teacher can, in discussing this latter fact with pupils, consider its implications for atomic structure and hint that more light will be shed on it later.

10.3.3 The size of an atom

Talking of the "size" of an atom raises an important question: how exactly does one define the size of an atom? It is prudent to warn pupils about this difficulty and suggest to them that they will discover later that atoms do not really have well-defined boundaries. But for the present it will be useful for them to see that with data already at their disposal, they can make a tolerably precise estimate of the volume occupied by an atom when it is part of an element in the solid state. For example, the average volume occupied by a gold atom can be calculated from the mass of a hydrogen atom (known from mass-spectrograph measurements), the relative atomic masses of hydrogen and gold (" atomic weights") and the density of gold in the solid state. If one then imagines for the sake of simplicity that each atom is sitting in a cubical box of this volume, and that the atoms are like little balls just fitting into these boxes, one can easily calculate the "diameter" of an atom. Pupils can be referred to their earlier measurements on the estimate of molecular size, and to the fact that the spacing of atoms in crystals can be measured by using X-rays of known wavelength; the latter spacing is in accordance with the calculated figure of the atomic diameter.

/Can.....

Can one <u>see</u> atoms? Photographs of atoms taken with a field-ion microscope should be shown to pupils to illustrate that in a sense, atoms can be "seen", and that the featureless little blobs that appear in these photographs tell us nothing about the <u>structure</u> of the atom. To discover more about the latter, other, less direct methods will have to be resorted to.

10.3.4 <u>The unstable atom</u>

There are two main reasons for including some account of radioactivity in a course of this kind - the intrinsic interest of the subject and the fact that certain radioactive atoms emit radiations which can be used to probe atoms. The intrinsic interest of the subject is considerable, but there would not be time to develop it in a thorough, systematic fashion in this course. The standard way of demonstrating that there are three categories of radiation from naturally radioactive substances is to plot absorption curves using some suitable detector for the radiation. In principle, there can be no objection to introducing radioactive radiations in this way, but, as the Nuffield Guide (V) urges, "we should not labour [this] study". A carefully-conducted experiment showing the discontinuities in the absorption curve would take some time for teacher and pupils to carry out; this would be acceptable if the teacher had a particular liking for the experiment and could communicate his enthusiasm to his pupils, but not otherwise. What is required is to show the pupils <u>as quickly as possible</u> that there are three distinct types of radioactive radiations and to acquaint them with the main characteristics of each. A way to do this is to use pure alpha-, beta- and gamma-ray sources to show some of the distinguishing features of each. Alpha particles can be "detected" by a "spark counter" (see below) and by a

/G-M.....

G-M tube with a suitably thin end window; they ionize heavily, producing readily visible tracks in a cloud chamber, and are able to produce sufficient ions in the vicinity of an electroscope to discharge it if the source is held sufficiently near. They are easily stopped, having a range of only a few cm in air at S.T.P., and being unable to penetrate a piece of paper of ordinary thickness. All this can be demonstrated with comparative ease. One would like also to be able to demonstrate directly that alpha-rays carry a charge or indeed that they have a measurable charge-to-mass ratio. Unfortunately, very strong magnetic fields are required to produce appreciable deflection of alpha-"rays", which makes such demonstrations forbiddingly difficult in a school laboratory. It is easily shown that the radiation from a pure beta source is at once more penetrating and less strongly ionizing than alpha-"rays". (It is not invariably the case that beta-"rays" are more penetrating; as Warren (1965) points out, the least energetic beta-"rays" are less penetrating even than alpha-"rays".) Beta-"rays" are readily deflected by magnetic fields of moderate strength (e.g. the field of a strong, portable permanent magnet), as can be shown by fixing a pure beta source at one end of a glass tube, placing the window of a G-M tube at the other end and holding a strong magnet at right angles to the beam (cf Lewis 1964). The count rate of the G-M tube is reduced almost to zero, suggesting that the beta-rays are in fact charged particles. Repeating this experiment with a gamma-ray source will show that the application of the magnetic field does not affect the counting rate in the slightest: the teacher could assure pupils that increasing the strength of the magnetic field would make no difference to the result, so that whatever gamma- radiation might be, it is not composed of charged particles. The fact that gamma-rays are, in general, more penetrating than beta-rays is easy to demonstrate.

/There.....

There is a celebrated diagram used in nearly all elementary textbooks to illustrate the nature of alpha-, beta- and gamma-rays. It shows the three kinds of radiation emerging from a hole in a lead vessel containing radium; a uniform magnetic field is applied at right angles to the plane of the paper, and the gamma-rays are shown undeflected, the beta-rays strongly deflected to one side and alpha-rays slightly deflected to the other. This diagram was first used in 1903 by Madame Curie and shortly afterwards reproduced by Rutherford in his early book on radioactivity (Rutherford 1905). His example has been followed by countless writers on the subject, including the authors of a textbook on physical science published in South Africa in 1965 (Prinsloo et al 1965). Unfortunately, many users of this diagram have failed to take proper note of a cautionary remark by Rutherford: "The amount of the deviation of the alpha-rays compared with that of the beta-rays is much exaggerated in the figure." If this diagram is shown to pupils, this caution should be borne in mind. There is another respect in which this diagram can prove misleading. As Warren (1965) has pointed out, "It certainly succeeds in impressing many students with the wrong idea that all radioactive materials emit these three kinds of rays. This wrong idea is in fact sometimes explicitly taught." (Warren's (1965) book gives a convenient summary of the relevant "right ideas".)

For the purposes of the present course, pupils need know little more about the properties of alpha-, beta- and gamma-radiation than will already have been established in the demonstrations mentioned above. Of course, they must know the identity of the three types of radiation, and this is something that experiments in a school laboratory cannot establish with finality. Historically, the identity of alpha-rays

/was.....

was established by a sequence of experiments in which the charge-to-mass ratio of alpha particles was shown to be that of doubly-ionized helium atoms, the charge on a known number of alpha particles was measured and found to yield an average value of the charge on the alpha particle of about twice the electronic charge, and finally, in Rutherford's and Royds's famous experiment, it was shown by spectroscopic examination of the gas formed by accumulated alpha particles, that these particles must indeed be doubly-ionized helium atoms. There is a film of the last-mentioned experiment which Lewis (1964) recommends showing; we would suggest, however, that this need not receive high priority. The Nuffield Guide (V) recommends instead that pupils study a photograph of the cloud-chamber tracks produced by alpha particles in wet helium (e.g. the famous set of photographs obtained by Blackett and reproduced in Rogers (1960) and elsewhere), since if the right photograph is used it will show the result of a collision between an alpha-particle and a helium atom. The tracks of the particles after collision are similar in appearance and are at right angles to each other. It should be easy for pupils to convince themselves that this means that the colliding particles have equal mass; the elastic collision between a moving particle and a stationary one of equal mass is easily analysed in terms of conservation of energy and momentum to show that the velocities of the particles after collision are at right angles to each other, and this conclusion can be shown to be true for dry ice pucks on a table or for equal steel balls suspended from very long pendulums (cf Rogers 1960). We agree with the Nuffield Guide (V) that "it is better to show the right-angle fork and relegate the spectrum [Rutherford's and Royds's] experiment to historical or A-level studies." The interested pupil can always be referred to textbook accounts of the latter experiment, which is entirely straightforward in principle.

/The.....

The identity of beta- and gamma-rays is not a matter of major consequence for what remains of the course, so that it does not matter if the establishment of this identity is done largely by assertion. The pupil knows already that beta-rays are readily deflected by magnetic fields, and that gamma-rays are not affected by such fields. To complete the picture he can be told that beta-rays are fast electrons and gamma-rays are high-frequency electromagnetic radiation; as evidence of the former statement, measurements of the charge-to-mass ratio of beta-particles can be mentioned - at this stage it is not necessary to dwell on the variation of "mass" with velocity - and as evidence of the latter, it can be pointed out that gamma-rays exhibit essentially the same properties as we associate with electromagnetic radiation: for lower frequency gamma-rays it is possible to measure their wavelengths directly by the techniques of X-ray diffraction, the speed of travel of gamma-rays has been measured and found to be equal to that of light, gamma-rays are polarized and can give rise to the photoelectric effect, etc. The fact that all these radiations stem from the nucleus should not be emphasized as yet, since the existence of the atomic nucleus has yet to be demonstrated.

The pupil should not be left without some understanding of the principles of operation of the detecting devices of which so much use has to be made in this section. He has already met the cloud chamber in the junior years, but the ionization chamber and the G-M tube are new to him and some explanation should be given of how they work. The ionization chamber raises no difficulties of understanding; the G-M tube is much more complex, but a simple account can be given of the initiation of an avalanche discharge in it by the entry of an ionizing particle into the tube. The action of a scaler attached to the G-M tube to count its output pulses need not be explained; the scaler is simply a convenient form

/of.....

of pulse counter, and the fact that it is genuinely counting the pulses from the G-M tube can be made plausible if the output from the tube is also connected to a loudspeaker. Then the occurrence of "clicks" can be seen to be associated with successive counts of the scaler.

Prof. Eric Rogers has devised a simple but highly effective teaching sequence for introducing the action of devices like G-M tubes which can record the passage of individual particles. In this, pupils are successively shown a "salt counter", a "match counter" and a "spark counter". Details are given in the Nuffield Guide (V).

The exponential decay of the radiation from a given source should, in our opinion, be regarded as a possible "optional extension" of the course rather than as essential to it. Certainly the story of radioactivity is not complete without an investigation of this aspect, but the pupil who has followed that story up to this point is ready to grasp the use of alpha-particles to probe atoms, and this, as we have suggested, is the primary aim of his study of radioactivity. We would agree with the recommendation of the Nuffield course that "pupils should, if possible, watch a counter recording the decay of a real substance" (Nuffield Guide V); the Nuffield authors recommend the use of protoactinium (half life of about a minute) for this purpose, as it can easily be separated from uranium by extraction with an organic solvent.

10.3.5 The discovery of the atomic nucleus

The present section of the syllabus is one of the least amenable to presentation by means of selected demonstrations and class experiments. At present, the cost of the necessary apparatus makes it impossible to arrange for the classic experiments on alpha-particle scattering to be performed in the

/school.....

school laboratory. Lewis (1964) has expressed the hope that the development of silicon diode detectors will change this situation, but at the time of writing this hope has not yet been fulfilled. The present writer once asked a class of university students, "How do you know the atom has a nucleus?" The first answer was promptly given: "Because you told us." Strictly speaking, the answer was incorrect - the lecturer had not previously asserted anything about the atom to these students - but the student had made his point. There is a serious danger of one's teaching of atomic and nuclear physics taking the form of "Believe it or not", or rather "Believe this or else...". The teacher faces the problem of trying to make the astonishing facts of the subject sound convincing without being able to bolster up his arguments with demonstrations of the crucial experiments. He can solve this problem at least partly by constantly appealing to physical principles and by liberally illustrating his lessons on the topic with reproductions of cloud and bubble chamber photographs, photographs of apparatus and other suitable material. In this way he can at least give the pupil the feeling that if he were asked to prove that the atom had a nucleus and given the facilities with which to do it, he would know which experiments to perform and how to reason from the data obtained.

At this stage in his investigation of the atom, the pupil has seen no evidence for any particular kind of atomic structure, i.e. any particular arrangement of electrons and positively-charged matter within the atom. Pupils can be invited to suggest possible models of the atom that are consistent with what they know about it. This might strike them as somewhat artificial - "everyone knows the atom has a nucleus" - but it will do them no harm to be reminded of the need always to ask where the evidence is. We do not suggest that much time need be spent on such preliminaries; what is more important is that pupils should rapidly be introduced to the methods by which the atom can be probed, and in particular

/to.....

to the alpha-particle scattering experiments initiated by Rutherford. There is certainly much to be said for sticking close to history here. Rutherford's (or, strictly speaking, Geiger's and Marsden's) experiments were of a classic simplicity and elegance in principle and design, and are correspondingly simple to describe and discuss. They are also well-documented (excellent historical accounts by Romer (1961) and Andrade (1935) are available), and something like a "standard analogy" for alpha-particle scattering has been evolved and is now widely used in teaching (see below).

One can introduce the alpha-scattering experiments by drawing attention to the fact that alpha and beta particles are able to penetrate very thin foils, showing that the atoms themselves must be "transparent". The question now arises as to whether this "transparency" results from the atoms consisting largely of "soft" matter or of empty space. This was the question that was decisively answered by the alpha-scattering experiments carried out under Rutherford's direction at the University of Manchester. The principle of these experiments can be explained with the aid of one or other simple analogy. The "standard analogy" mentioned above (cf. P.S.S.C. 1960, 1965, Rogers 1960, Nuffield Guide V) is of the use of bullets fired into a bale of hay in order to find the position and size of lumps of metal concealed in the hay. Those bullets which scored direct hits on the hidden lumps would bounce off and emerge from the bale at various angles to their direction of incidence. Measurements on the paths of these bullets could be used to deduce the position of the hidden targets. If sufficiently many bullets were fired into the hay in the neighbourhood of any one of these deduced positions, the pattern formed by the scattered bullets would give information about the shape of the hidden target. Another instructive analogy is that suggested by Ashford

/(1960,.....

(1960, 1967). A contractor is given a certain amount of metal from which to make a wall of given dimensions. He can make the wall in the form of either a flat homogeneous sheet of metal or of a wire fence. If the finished article is covered in wallpaper, how could one find out what the contractor has done without removing the paper? One could fire rifle bullets at the wall and note what happened. If the wall consisted of thin sheet metal, all bullets would pass through undeviated and with slight loss of speed. If the wall were in the form of a fence, however, a small fraction of the bullets would be reflected off the thick wires and undergo deflections in consequence.

In 1908, Rutherford suggested to a junior colleague, Hans Geiger, that a young research student, Ernest Marsden, should see whether some of the alpha-particles used to bombard thin foils could be turned round through large angles. (Rutherford did so without apparently having any real conviction that any would be.) The alpha-particles were to be used as the "bullets" intended to probe the atomic "hay". The selection of alpha-particles from polonium as the bullets was prompted by their high kinetic energy, their comparative ease of observability and the fact that while they penetrated matter to some extent, they did not penetrate it so well that they simply went straight through under all circumstances.

Before proceeding to a description of the results of these experiments, the teacher should draw attention to the results already obtained by Geiger in experiments on the passage of alpha-particles through thin metal foils. Geiger had shown that in general these particles could be deviated through small angles, the deviation being greater the thicker the foil. This could be explained on the assumption that an alpha-particle passing through the foil undergoes a large number of very small chance deflections in random directions.

/The.....

The foil is a large number of atomic diameters thick, as can be shown by a simple calculation. Thus Geiger's experiments had established that alpha-particles could pass through hundreds or even thousands of atoms without appreciable deflection of their paths.

It is important also for the teacher to know of the sequence of events following Rutherford's above-mentioned suggestion to Geiger and Marsden, since this can guide him in his approach to his pupils. In their initial experiments, Geiger and Marsden (1909) found that a few of the incident alpha-particles were deflected through large angles; about 1 in 20,000 were turned through an average angle of 90°, and a very few of the particles were turned through angles of 150° or even more. Rutherford has recorded his utter amazement at this news: "It was quite the most incredible event that had ever happened to me in my life. It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you" (Andrade 1965). Since the average present-day pupil already "knows the answer", it will be impossible for the teacher to create in him the sense of wonder that Rutherford and his colleagues experienced; but the teacher should be able to convince such a pupil that if the atom did consist of a pudding of positive charge in which electrons were embedded, then the wide-angle scattering of alpha-particles is not merely difficult but impossible to explain.

Having had one or other of the appropriate analogies for alpha-scattering described to him, the pupil will be ready to interpret the results of Geiger and Marsden (and the appearance of the tracks in suitable cloud-chamber photographs) as indicating that atoms do not consist of "soft matter" but have hard cores instead. This is excellent as far as it goes, but the teacher should make it clear that the matter is subtler than this. It was not until early in 1911 (i.e.

/more.....

more than a year after the performance of the initial experiments of Geiger and Marsden) that Rutherford was able to tell Geiger that he "knew what the atom looked like". He had then worked out the theory of the scattering of alpha-particles by concentrations of positive charge in the atom. This was published in due course (Rutherford 1911). Geiger and Marsden then set about verifying this theory in detail. They were able to show that Rutherford's formula relating the proportion of alpha-particles scattered in a given direction to the angle of scattering fitted their results within the limits of experimental error; in effect, they had confirmed the correctness of Rutherford's hypothesis of a nuclear atom (Geiger and Marsden 1913). Because of the complexity of the derivation of Rutherford's formula (and indeed of the formula itself), pupils cannot be conducted along all the steps of this argument. But it is necessary to point out to them that the deduction of the existence of the atomic nucleus does not depend only on the qualitative observation that some alpha-particles incident on metal foils are scattered through very large angles. To guide them along Rutherford's line of thought, they can be shown a simple qualitative experiment in electrostatics (cf Lewis 1964). A small pith ball coated with aquadag hangs from the top of a retort stand by an insulating thread. A small metal ball mounted on an insulating stand is placed vertically below the point of support of the thread, metal ball and pith ball being at the same height. The balls are given a like electrical charge, and the pith ball is drawn to one side and released. It will be found that the trajectory of the pith ball about the stationary metal ball is approximately hyperbolic. The apparatus can thus be used to give a qualitative idea of what happens when a moving charged particle passes near a stationary one; in particular, it shows the influence on the trajectory of the moving ball of the value of the "impact parameter", i.e. the

/perpendicular.....

perpendicular distance between the initial line of motion of the centre of the ball and the centre of the stationary ball. The smaller the impact parameter, the bigger the angle of deflection of the direction of the moving ball.

As we have already mentioned, the derivation of Rutherford's formula is complex; it is impossible to give it in a school course. On the other hand, the basis of the derivation is very simple; the necessary physical principles are just the laws of particle mechanics and Coulomb's Law. This should be stressed by the teacher; and by way of achieving an acceptable compromise between no theory at all and the full theory, he can present pupils with a table showing how Geiger's and Marsden's results fitted in with Rutherford's formula. This course is recommended by the Nuffield Guide (V), in which the appropriate table is reproduced. We think the recommendation well worth following, since it reduces by a considerable factor the amount of material the pupil has to take on trust. He sees some data and can check for himself the calculations which test whether or not the data fit the formula. The P.S.S.C. have made a film on "The Rutherford Atom" which could be used with great benefit in the teaching of this section.

In his derivation of his scattering formula, Rutherford (1911) assumed that each incident alpha particle would be scattered once only. This gives rise to the prediction that the number of particles scattered per unit time at a given angle and arriving on a given small area of the detector must be proportional to the thickness of the scattering foil. Geiger and Marsden (1913) investigated this prediction in detail and showed it to be correct, provided the thickness of the scattering foil was not too large. In their earlier research, Geiger and Marsden (1909) had shown that the number of alpha particles scattered backwards from the target increased with increasing thickness of target and then became

/steady,.....

steady, thereby establishing the important result that this "reflection [was] not a surface but a volume effect". Elementary accounts of Rutherford scattering often omit reference to this matter, but it should definitely be mentioned. Pupils should be made aware of the possibility that the effect could have been a "surface" one, and of how this possibility could be eliminated.

It was a further prediction of Rutherford's formula that the number of particles scattered must be proportional to the charge on the nucleus. The inclusion of this point in the teaching sequence will be dealt with in the next section.

(The conscientious teacher may wonder why it is permissible in dealing with this particular microscopic problem to use completely classical reasoning. Incident alpha-particle and target nucleus are both treated as classical particles, i.e. as tiny billiard balls, and only classical laws are used to derive the scattering formula. Yet the formula works. Strictly speaking, we should use the principles of quantum mechanics to solve this problem, but if we did, we should find that by a singularly fortunate coincidence we are led to exactly the same formula as was derived by Rutherford along purely classical lines. This "coincidence" is a consequence of the unique property of the Coulomb force; if a scattering formula is derived for cases of scattering of microscopic "particles" in which the law of force is not Coulombic, "classical" and "quantum-mechanical" derivations would not lead to the same result, and it would be the quantum mechanical and not the classical formula which would be correct. As far as we know, there is no simple way of showing that a "wave theory" of scattering must lead to the results obtained in the Manchester scattering experiments. Therefore the only feasible way to teach this topic at this level is to use the

/classical....

classical approach as outlined above. Certainly we are to some extent "cheating" by doing this, but if we make a clean breast of this we shall do no harm. It can be argued that we might create fewer difficulties for pupils if we taught the discovery of the atomic nucleus before teaching pupils about the singular behaviour of microscopic entities. Then we could approach alpha-particle scattering classically with a clear conscience, because at this stage the pupils would have no reason to expect any other approach. Significantly both the Nuffield and the P.S.S.C. courses follow this order of presentation, and it may well be that the South African Common Basic Syllabuses are in error in apparently advocating that one teach pupils about photons and matter waves before going on to talk about the existence of the atomic nucleus. There is room for some divergence of opinion here, because the "photons and matter waves first" approach has its advantages too: after one has discussed the "waviness" of light, it seems reasonable to go on to talk about its "particle-ness", and then to do the inverse for electrons and other microscopic entities. Perhaps the teacher is the best judge of the more successful order of presentation, though his judgment in this case may well be arrived at as a result of having first tried out both methods.)

10.3.6 The size of the nucleus and the size of the atom

The question of the relative sizes of the atom and its nucleus is obviously important because of its relationship to atomic structure. An essential preliminary to the discussion of this question is to make the point that the magnitude of the positive charge on the nucleus can be deduced from alpha-scattering measurements with the aid of Rutherford's formula. It was Chadwick who first measured the nuclear charge in this way and found that within the limits of

/experimental.....

experimental error it was equal to the "atomic number" of the element times the electronic charge. At that time, the concept of "atomic number" was already in use in chemistry as the indicator of the place of an element in the periodic sequence of elements, and just prior to Chadwick's work Moseley had carried out an investigation of the characteristic X-ray spectra of elements, that established the connection between nuclear charge and atomic number with greater accuracy than was possible from Chadwick's direct measurements. The pupil will not have the background to understand Moseley's work, and it is not necessary to mention it. All that need be said here is that accurate measurements on the scattering of alpha-particles by nuclei confirm that the charge on the nucleus is an integral multiple of the electronic charge. There is no harm in simply designating this integral multiple as the atomic number, although if the pupil has some chemical background it is sensible to point out the connection between this "atomic number" and that defined from chemical considerations.

For present purposes the important fact is that the charge on the nucleus is known from alpha-scattering measurements, so that its value can be used in an approximate calculation of the upper limit to the nuclear radius. Since Rutherford's alpha-scattering formula is derived on the assumption that the interaction between the alpha-particles and the nucleus is purely electrostatic in origin, its success in predicting the observed facts of alpha-particle scattering may be taken as proof that in the initial experiments these particles never actually "hit" any of the target nuclei (cf P.S.S.C. 1960, 1965). If they did, forces of other than purely electrostatic origin would act between the nucleus and the particles, and the observed scattering would not be in strict accordance with Rutherford's formula. One must therefore

/conclude.....

conclude that even when one of the incident alpha-particles came as close as possible to the centre of a target nucleus, its distance of closest approach still exceeded the nuclear radius.

The distance of closest approach can easily be calculated from energy considerations. An alpha-particle of given initial kinetic energy will approach closest to a nucleus if it is aimed straight at its centre. It will advance towards the nucleus until all its kinetic energy has been converted into mutual electrostatic potential energy of nucleus and alpha-particle; the process of energy conversion will then take place in the opposite direction and the alpha-particle will retrace its path backward. To calculate the distance of closest approach, one need simply write down the equation expressing the fact that when the alpha-particle is at a distance d from the centre of the nucleus, its initial kinetic energy has all been converted into potential energy. There is a minor difficulty about writing down this equation, viz. that the pupil has not had derived for him the expression for the mutual potential energy of a pair of charged particles separated by a known distance. Nor can this formula be properly derived for him, since he lacks the mathematical equipment to understand the derivation. Fortunately this is not a serious problem. The calculation under discussion is only an estimate of the nuclear radius, intended to establish the difference in orders of magnitude between the radii of the atom and its nucleus, so that the pupil need not be unduly concerned about not being able to derive the precise form of the expression for the potential energy, viz. $9 \times 10^9 \frac{q_1 q_2}{d}$. We suggest that it be pointed out to pupils that this expression represents the amount of work that would be done by the electric field when a charged particle of charge q_2 moves a distance d in a uniform field of strength $9 \times 10^9 \frac{q_1}{d^2}$, and that this amount of work would certainly be at least of the same order as that which would be done by the field of a fixed

/point.....

point charge q_1 on a charge q_2 that goes from a distance d from q_1 to a very great distance away.

The equation to be written down is then

$$9 \times 10^9 \frac{Ze^2}{d} = \frac{1}{2} mv^2$$

where m and v are respectively the mass and initial velocity of the alpha-particles, e is the electronic charge and Z the atomic number of the nucleus. By substituting the appropriate values for e, m, v and Z in this equation, using $Z = 79$ for gold and $v = 1.6 \times 10^7$ m/s, as in the original experiments of Geiger and Marsden, one finds that d is roughly equal to 4×10^{-14} m. This, then, is the upper limit to the nuclear radius set by the meesurements of Geiger and Marsden. Since the radius of a gold atom is of the order of 10^{-10} m, it is readily seen that the nuclear dimensions are four orders of magnitude smaller than those of the atom. Here is the justification for asserting that the atom consists largely of empty space.

It is worth mentioning to pupils that in later experiments in which nuclei were bombarded by more energetic alpha-particles and other subatomic "bullets", it was found that when the "bullets" had enough energy, there were significant departures of the scattering pattern from that predicted by Rutherford's formula. This showed that some of the incident particles had evidently come close enough to the nuclei to be acted upon by forces of other than electrostatic origin. These experiments have enabled more precise determinations of nuclear size to be made, so that it is now possible to say that the radii of nuclei range from about 10^{-15} m for the lightest nuclei to about 10^{-14} m for the heaviest. It is a considerable achievement of modern physics to be able to specify the sizes of atomic nuclei, and the fact is worth underlining.

/10.3.7.....

10.3.7 A critical survey of some possible approaches to atomic structure

The planner of a syllabus having the scope of that discussed in this thesis is confronted with certain important decisions when considering what should be taught under the heading of "atomic structure". Should he recommend the teaching of the Bohr-Rutherford model, on the grounds that at least this gives the pupils something that they can readily visualize? If he does, then it is with the awareness that he is advocating the teaching of an atomic model whose successes were virtually confined to hydrogen-like atoms, and which has been known for some forty years to be incorrect. (As long ago as 1929, we find Westawey (1929) warning teachers that "The planetary picture [of the atom] is just a pretty analogy borrowed from astronomy, and pupils should be advised not to assume that it is really anything more".) Or should he recommend the use of the "orbital" approach, explaining the existence of energy levels in hydrogen-like atoms by using a picture of electronic standing-waves fitting round the nucleus? As we shall see below, this approach is not unobjectionable either. Or should he decide that the safest way out is to concentrate on the teaching of those aspects of the atom which are accessible to direct or nearly direct measurement: its energy levels, its ionization energy, the frequencies of the lines in its spectrum? There is a great deal to be said for this last approach (it is strongly advocated by Lewis (1964)), but if it is dealt with to the exclusion of any attempt to say what the atom is like, the pupil is bound to feel cheated.

Of course, the above three approaches are not necessarily mutually exclusive; and the obvious way out of the difficulty is to try to find an approach which combines the best features

/of.....

of all of them. Let us first consider what might be worth saving from the Bohr theory of the atom. Bohr (1913) assumed that when an atom was in a stationary state, it would not radiate; that the atom would only radiate when passing from one such state to another; that when the atom was in a stationary state its "dynamical equilibrium" was "governed by the ordinary laws of mechanics"; that the radiation emitted would have a frequency ν determined by the relation $\Delta E = h\nu$, where ΔE was the energy emitted and h was Planck's constant; and the allowed stationary states would be determined by the condition (in the case of circular orbits) that "the angular momentum of the electron round the nucleus is equal to an entire multiple of $\frac{h}{2\pi}$." Of this list of assumptions, only the third would now be rejected outright. We know now that we can obtain satisfactory descriptions of the behaviour of microscopic systems only if we assume that the ordinary laws of mechanics do not apply at all to such systems. The last of Bohr's assumptions has to be somewhat modified to bring it into line with present ideas. We would now make no mention of "orbits", and formulate the "quantization" condition for the atomic angular momentum by saying that the maximum value of the component of the orbital angular momentum in a particular direction must be an integral multiple of $\frac{h}{2\pi}$. Thus Bohr's theory of the atom is far from being unacceptable in its entirety; its most serious defects are its adherence to the idea of "orbits" governed by the laws of classical mechanics, and the ad hoc character of its basic assumptions. (Those assumptions of the Bohr theory that are now known to be correct can be deduced in quantum mechanics from basic postulates of a much more general nature.)

It would seem that only a fanatical adherence to the historical development can justify the full-scale teaching of the Bohr theory in a modern physics course at whatever level.

/In.....

In a syllabus such as the one under discussion, it would in any case be impossible to teach the quantization of angular momentum, as the latter concept does not appear in the syllabus. We shall consider below whether any of our teaching of atomic structure ought to make explicit use of Bohr's ideas.

Turning now to what we can call a "simplified wave-mechanical approach" to atomic structure, we meet the serious difficulty that the wave-mechanical treatment of even the simplest of atoms cannot be satisfactorily simplified. One can give a beginner who is familiar with the de Broglie relation derivations of the expressions for the energy levels of a particle in a square well or in a box, which do not necessitate the solution of the Schrödinger equation; essentially, this is possible because the Schrödinger equation reduces to a particularly simple form in these cases. The situation is exactly analogous to finding the allowed wavelengths of standing waves on a string stretched between two points; one can do this very simply by "inspection", instead of solving the appropriate differential equation. But for the hydrogen-like atom we have an entirely different state of affairs; the solution of the Schrödinger equation for this case is lengthy and uses sophisticated mathematical procedures, and there are no honest short cuts to the correct expression for the energy levels.

This last remark may occasion some surprise, as the practice of fitting electron standing waves around the nucleus and thence calculating the energy levels has been much used in elementary texts and courses, and it indubitably gives the right answers. Moreover, its ancestry is honourable - it originated with de Broglie. We list below those respects in which it can be held to be defective:-

(i) It gives no indication of the fact that the true "wave" is one which represents the two-particle system (electron plus nucleus) and not just the electron by itself.

/(ii).....

(ii) In order to calculate the energy associated with each electronic standing wave "fitting on to itself" around the nucleus, it is necessary to find an expression for the electron velocity by equating the centripetal force on the electron to the electrostatic attractive force between nucleus and electron; to substitute the resulting expression for the velocity into the classical expression for the kinetic energy and to add this to the electrostatic potential energy of the atom. But to carry out the first steps of this procedure one has to treat the electron like a classical particle subject to "the ordinary laws of mechanics", which is precisely what we ought not to do in a proper wave-mechanical approach to the atom.

(iii) It is often pointed out in connection with this approach to the atom that it represents a <u>derivation</u> of the Bohr quantization condition for angular momentum instead of a mere postulation of this condition. It is much less often pointed out that although by use of the incorrect procedure mentioned in (ii) this approach gives the right (non-relativistic) expression for the energy levels, it does not give the right value for the maximum orbital angular momentum associated with any given energy level. It is not just that the true maximum value of the angular momentum is not a rational multiple of $\frac{h}{2\pi}$; even the maximum value of the component of angular momentum in a given direction is incorrectly given. Whereas according to the above derivation one would expect this value to be $\frac{nh}{2\pi}$ for the n^{th} energy level, it is in fact $(n-1)\frac{h}{2\pi}$; in particular, the ground state of the hydrogen atom has zero orbital angular momentum whereas if the above derivation were correct we would expect it to be $\frac{h}{2\pi}$. To put it in its simplest terms, the de Broglie derivation of the Bohr quantization condition predicts the wrong number of nodes in the wave function for a given energy level.

(iv) This approach gives no hint of the existence of other kinds of standing wave in the space surrounding the nucleus.

/It....

It is presumably for these reasons that Elton (1967) has referred to the idea of standing waves in a closed orbit as "quantum mechanical nonsense". In fairness we must record that some at least of those writers who describe it are aware of its defects, though they may not necessarily refer to these in detail. The Nuffield Guide (V), for example, does say that "Even for hydrogen, the picture of waves on circular 'orbits' proves incomplete and partially incorrect."

The above critical examination of the "standing-wave-patterns-in-closed-orbits" approach to the atom has shown that it is, after all, not much of an improvement on the Bohr theory. Its only advantage is that it enables us to derive the correct expression for the energy levels of hydrogen-like atoms in a simple way, by making some use of the "waviness" of matter and without our having to appeal explicitly to the Bohr quantization conditions. One might argue that by carrying out this derivation, one at least has produced a theoretical prediction whose agreement with experiment can be studied; but since the theory in question is bogus, as we have shown, this exercise would seem to be of doubtful value. From time to time, attempts are made to produce simplified accounts of the hydrogen atom which do not falsify the theory and which do give beginning students some idea of the properties of the atom as predicted by quantum mechanics. A thorough-going attempt along these lines has recently been published by Cohen (1967). This makes use of the concept of "equivalent square well" potential energy functions to discuss the dependence of the wave function on the r and θ coordinates ("r and θ directions", as he calls them), and by this means is able to arrive at an approximate expression for the energy levels. But it still includes a number of mathematical and physical subtleties which rule it out of consideration for use of

/schools.

schools. Indeed our only reason for mentioning it here is that it illustrates the essential impossibility of producing a worthwhile, simple, quantitative treatment of the hydrogen atom that can be understood by pupils.

There remains to be considered the approach to the atom via the experimental facts concerning atomic energy levels. The Nuffield Guide(V) has come out strongly against this, claiming that "To our pupils the argument from frequency differences through the quantum idea to energy levels would be new and far too difficult." With this view we concur. We feel that pupils should certainly know of the connection between the frequencies of spectral lines and the energy levels of atoms, but that to expect of them to infer the existence of the latter from observations of the former, even with careful guidance, is probably unreasonable. Mott (1964) has advocated that the existence of quantized energy states should be presented to pupils as a straightforward deduction from the experimental fact that the (molar) specific heat at constant volume of a monatomic gas is $\frac{3}{2}$ R. The argument is that this value of the specific heat shows that "when atoms collide in a gas, they do not apparently transfer any of their kinetic energy to the energy of internal movement of whatever the internal constituents of the atom are." Therefore one can postulate that a minimum amount of energy ΔE has to be transferred to an atom to give energy to its constituent parts, and that ΔE must be much greater than the mean kinetic energy of a gas atom at room temperature. We suspect that this reasoning might also be found to be too subtle by the average pupil, at any rate by the average O-level or South African matriculation pupil. Being wise after the event, it is easy for one to agree that the thermal behaviour of monatomic gases is suggestive evidence for energy quantization, but one may doubt whether the newcomer to modern physics will see this in

/quite.....

quite the same light. (Mott himself remarks that it would be an "interesting historical study" to enquire why it was not realized prior to Bohr's paper of 1913 that quantization of the ground state of an atom could be deduced from the observed specific heats of gases.)

10.3.8 A suggested teaching approach to atomic structure

In an attempt to combine the most valuable features of the above-discussed approaches to atomic structure in an approach which would still be assimilable by pupils at South African schools, we suggest the following outline:-

One should begin with some discussion of the picture of the nuclear atom suggested by Rutherford's alpha-particle scattering experiments, since in this way the difficulties connected with a purely "classical" particle model of the atom can be brought home to the pupil. These difficulties reside in the inherent instability of such a model. A simple "particle electron" travelling round the nucleus is an accelerated particle, and an accelerated charged particle radiates electromagnetic waves and so loses energy; as the electron loses energy it must move closer to the nucleus and eventually coalesce with it. According to this model, any given atom would very quickly collapse into the volume occupied by the nucleus, but the observed behaviour of matter is not consistent with this prediction at all. Nevertheless atoms do emit light: atoms of a particular element emit characteristic "light patterns", or spectra, by which these elements may be identified, and these spectra consist of lines, corresponding to electromagnetic waves of well-defined frequency. So atoms do radiate away some energy, but they do not do this in accordance with what would be expected of a "classical atom".

It can now be suggested to pupils that since they have already learned to be suspicious about the application of "ordinary" physical laws to microscopic phenomena, they should

/be.....

be similarly suspicious here, and that a way out of the difficulties mentioned above is to assert that an atom can only have certain allowed energies, and that when it is in one of these allowed energy states, it cannot radiate. These, of course, are two of the key ideas of the Bohr theory and are among the features of that theory which are in accord with the newer quantum mechanics. To introduce the idea of energy levels in this way certainly amounts to "teaching by assertion", but we do not see how it can be introduced otherwise in a reasonably short space of time. It would be patently unrealistic to assume that pupils could see in advance of anybody telling them that the existence of stationary energy states is a necessary consequence of "wave mechanics." It seems to us a fair compromise between "discovery" and "dogmatic" teaching to suggest to pupils that there <u>are</u> energy levels and then to show them that this idea is consistent with observation and theory.

The demonstration that energy levels in the atom are consistent with what is observed has two aspects: comparison with data from atomic spectra, and study of collision experiments of the Franck-Hertz type. The first of these, if it is to be done properly, requires the introduction of the idea of a spectral series, the demonstration that the lines in a given series correspond to transitions to a particular lower energy level, and that the frequency of a line in a given series is equal to the difference in the frequencies of two of the lines in one of the other series. This discussion can be based on the spectrum of a particular element, for which the obvious choice would be hydrogen. Photographs must be used to display these series, since only a few lines of the hydrogen spectrum appear in the visible region. Data giving the actual frequencies or "wave numbers" of the spectral lines are readily available (e.g. in Lewis (1964).)

/Naturally.....

Naturally it is now necessary to introduce Bohr's postulate that atoms radiate when undergoing transitions between energy states, and the frequency of the radiation is given by the energy difference divided by Planck's constant. The existence of a series limit is significant and should be explained as being the result of the tendency of the energy levels of the atom to become more and more closely spaced the higher their energy, so that they tend towards a limiting value of the energy. It can then be suggested that a reasonable physical interpretation of the existence of an upper limit to the (internal) energy of the hydrogen atom is that this upper limit corresponds to the removal of the electron, i.e. to the ionization of the atom. Finally the relationship between the ionization energy of the atom and the frequency limit of the spectral series representing transitions to the ground state should be made plain.

The second aspect of the experimental confirmation of the existence of atomic energy levels can, in principle, be dealt with by direct demonstration of the Franck-Hertz experiment. Several firms manufacture a version of this apparatus suitable for demonstration of the energy levels of mercury atoms. (In certain English sources it is implied that one of these firms (Phywe) also make a Franck-Hertz helium tube (cf Lewis 1964, Duncan 1967), but this firm has resolutely denied this when its South African agents have questioned it on this point.) The Nuffield Guide (V) is not in favour of the performance of this experiment in its O-level course, on the grounds that it looks "too complicated and too special". It certainly does seem unwise to recommend the purchase of the appropriate equipment by schools, since it is both expensive and of strictly limited usefulness. The P.S.S.C. film on the Franck-Hertz experiment presents a

/palatable.....

palatable account of it and should be used if available. Failing this, pupils should have the experiment briefly described to them. Warren (1965) has warned against the description of "imaginary experiments" in this connection, pointing out that it is risky to describe a Franck-Hertz experiment as though it were done with gaseous hydrogen, since this ordinarily consists of diatomic molecules having very different energy levels from those of atomic hydrogen.

It remains to show that the existence of atomic energy levels can be accounted for theoretically. A discussion of this point can be introduced, as in the P.S.S.C. course, with the demonstration that stationary energy states for an electron in a "well" can be predicted from the wave properties of matter. If an electron is confined between two rigid walls, it is to be expected that the "probability wave" associated with it must be a standing wave, which must satisfy the condition that only an integral number of half-wavelengths can fit into the space between the walls. This readily yields the result that the electron can only have certain allowed values of the magnitude of its momentum when it is confined in the "well", viz. $p = \frac{nh}{2\ell}$, where n is an integer whose lowest value is unity, ℓ is the distance between the walls, h is Planck's constant and p is the magnitude of the electron's momentum. If we use the classical formula for the relation between momentum and kinetic energy (quantum mechanics actually confirms its correctness in this case), viz. $K.E. = \frac{p^2}{2m}$, where m is the mass of the electron, we can use the above result to find the allowed energies of the electron. They are given by

$$E_n = \frac{n^2h^2}{8m\ell^2}$$

Two points in particular should be brought to the pupil's attention in connection with the above derivation and result.

/(i).....

(i) The electron is confined or "bound", and it is the fact of its confinement which has had the direct effect of limiting its wavelength and hence its momentum and energy to certain discrete values only.
(ii) The lowest energy of the electron is not zero; even when it has its least energy, it apparently does not lapse into inertness inside the well.

This treatment has shown that in a very simple case, the occurrence of discrete energy levels can be deduced from the wave aspect of electron behaviour. One can now turn to the hydrogen atom and point out that the electron in it is also to some extent "bound", since it is subject to the Coulomb interaction between the nucleus and itself. Therefore it might be expected that here, too, we would have standing waves corresponding to certain discrete energy levels. Only this time, it is not very easy to visualize these standing waves, since they are now in three-dimensional space. Also, the electron is not confined within a rigid enclosure, and we must expect that the nucleus must itself have some "motion" under the influence of the Coulomb interaction between it and the electron. At this point we think it would be wise for the teacher to tell his pupils that it will be impossible for him and them to continue further along this road, as the problem is one which can only be solved properly by the use of the techniques of "wave mechanics", which are too advanced to be dealt with at school. (cf. our remarks in Section 10.3.7 on this subject.). Certainly he can and should show them the pictures of the "probability patterns" for the electron in a hydrogen atom; and he should assure them that there is very satisfactory agreement between the quantum-mechanical calculations of atomic energies and the values of these that are inferred from spectrometric measurements. That the pupil will have to take this on trust need not be deplored too deeply; it is simply a reflection of

/the.....

the fact that one cannot fully teach all aspects of physics at school. (If chemistry teachers insist, something might be said also about different species of "quantum number", about "electron spin" and about the Pauli Exclusion Principle and the building-up of the Periodic Table. But we would not regard these as being urgently necessary in a physics syllabus, especially as anything that would be "taught" under these headings could be given little or nothing in the way of supporting argument.)

10.3.9 The composition of the nucleus

Concerning the atomic nucleus, the pupil now knows the following facts:

(i) The nucleus has a positive charge whose magnitude is an integral multiple of the electronic charge. The number of electronic units of charge on the nucleus is called the "atomic number".

(ii) Nearly all of the mass of the atom is associated with the nucleus.

(iii) Measurement of the masses of positive ions shows that the relative masses of atoms (and therefore of nuclei) are very nearly whole numbers.

In addition, the pupil will have to know something about the discovery and properties of the neutron. Historically, the discovery of the neutron did not come about as a result of a search intended to solve the puzzle of the composition of the nucleus, and one should not attempt to falsify history by presenting it as such, tempting though this might be. Nor is it necessary to tell the detailed story of how the neutron eventually came to be identified after the initial discovery by Bothe and Becker of the mysterious, highly penetrating radiation ejected by alpha-particles from beryllium. It will suffice to summarize the main points, viz:-

/(i).....

(i) Highly penetrating and very weakly ionizing "particles" can be ejected from certain nuclei, e.g. those of beryllium and boron, when the latter are bombarded with alpha-particles.
(ii) By means of the usual tests, it was established that these "particles" were uncharged.
(iii) In 1932, Chadwick was able to determine the masses of these particles by allowing the latter to collide with hydrogem and nitrogen nuclei and analysing the collisions in the light of the principles of conservation of momentum and energy. The relevant calculations are very simple and pupils should see them done as an illustration of how an important result can be deduced from an entirely straightforward and simple analysis (cf P.S.S.C. 1960, 1965).
(iv) The discussion of point (iii) would not be complete without some account of the method of measurement of the recoil velocities of the hydrogen and nitrogen nuclei, which constitute the actual data obtained in the experiment. Chadwick measured the recoil velocities of the hydrogen nuclei by finding the thickness of absorber required to prevent them from entering the ionization chamber used to detect them. The recoil velocities of the ionized nitrogen atoms were deduced from the lengths of the tracks they made in a cloud chamber. Pupils may wonder why these velocities were not deduced from measurements of the curvature of the cloud chamber tracks in a magnetic field. It would be an instructive exercise for them to calculate the strength of the magnetic field required to give these particle tracks a measurable curvature, given the velocities of the hydrogen and nitrogen atoms as being respectively 3.3×10^7 and 4.7×10^6 m/s (the values obtained in the early experiments).

(Up to this point, we have used the term "hydrogen nucleus" to refer to the nucleus of the (light) hydrogen atom. It is

/up.....

up to the teacher to decide whether he would like to use the term "proton" instead in the course of his treatment of the foregoing material. Certainly it is appropriate that he should use it thereafter.)

Pupils curious about the motivation of the experiments that led to the production of the mysterious radiation later identified as neutrons, can be told that these experiments were attempts to probe the nucleus itself by using alpha-particles.

Prior to the discovery of the neutron, it was supposed that the nucleus must be made up of protons and electrons, the residual charge on the nucleus being accounted for by the assumption that the number of protons exceeded the number of intra-nuclear electrons. This model was unable to provide a satisfactory explanation for the observed values of intrinsic nuclear angular momentum (nuclear spin); moreover an electron inside the nucleus should have had a far greater energy than that which was typical of the electrons emitted in beta- decay processes. It is impossible at school level to give a satisfactory account of the difficulties encountered by the proton-electron model of the nucleus. Fortunately there is no logical need to approach the problem of nuclear composition from the historical point of view. The pupil knows the facts concerning the masses of atoms and the charges of their nuclei, and he knows of the existence of the neutron and the fact that its mass is approximately equal to that of the proton. From these facts it is clear that the simplest picture of nuclear composition is that which assumes the nucleus to be made up of protons and neutrons in appropriate numbers. It is true that to proceed directly to the proton-neutron model of the nucleus is to short-circuit history, but it does not short-circuit the logic. As we have remarked, the neutron was not discovered as a direct result of desperate attempts to solve the problem of nuclear composition; the

/discovery.....

discovery of the "radiation" was essentially a fortuitous occurrence, unrelated to any speculations concerning what was inside the nucleus. Therefore it is perfectly fair to accept the existence of the neutron without any prior consideration of the proton-electron model of the nucleus. (In an intermediate course in nuclear physics, it is instructive for students to consider the reasons why electrons cannot be found in the nucleus, but such considerations are not of importance in an elementary course.)

The teacher could present the topic of nuclear composition to pupils in some such terms as these:-

"We know that atomic nuclei have masses that are nearly integral multiples of the mass of the hydrogen nucleus, or proton. But the helium nucleus, for example, although it has four times the mass of the proton, has only twice its charge, so that we cannot simply imagine the helium nucleus to be made of four protons. We encounter similar difficulties with all nuclei heavier than the proton. The way out is provided by the neutron, which is electrically neutral and has very nearly the same mass as the proton. Perhaps the nucleus is simply made up of the right numbers of protons and neutrons; for example, the helium nucleus or alpha-particle must consist of two protons and two neutrons."

It must now be "explained" how it is possible for protons, presumably packed very close to one another inside a nucleus, to stay close together despite the strong electrostatic repulsion between them. Naturally, this matter can only be dealt with very superficially, but one can at least say that very strong short-range forces must act between nucleons (protons or neutrons), that these forces are attractive over most of their range - this accounts for the ability of nucleons to stick together - but must apparently become repulsive at very short inter-nucleon distances. These nuclear forces form a category of their own; they are very much stronger

/(by.....

(by a factor of 100 or more) than electrostatic forces, which are in turn enormously stronger than gravitational forces. It is necessary to stress that the nuclear force acts between proton and proton, proton and neutron, and neutron and neutron. (We have quoted in Section 2.7 a passage from a textbook which indicated the author's belief that the internucleon force acted between proton and neutron only.) It is also worth mentioning that the "strong interaction" is not yet fully understood; there are many unanswered questions about nuclear forces which continue to occupy the minds of a great many theoretical and experimental physicists. This is one of the frontiers of knowledge of the subject.

Acceptance of the proton-neutron model of the nucleus makes it an easy matter for the teacher to give a more fundamental interpretation to the meaning of the term "isotope", which the pupil will already have met, and "isobar", to which he should now be introduced. The "atomic number" can now be identified with the number of protons in the nucleus; the "mass number" of an atom can be defined as the nearest whole number to its atomic mass when the latter is expressed in atomic mass units, and it can be pointed out that it is equal to the total number of nucleons in the nucleus. The fact that when talking about "mass numbers" and "isobars" one is really concerned with numbers of nucleons rather than with masses should be stressed by teachers; this idea will become particularly important when nuclear reactions are dealt with.

10.3.10 Nuclear reactions and "nuclear energy"

Many pupils will have heard of, and had their imagination fired by "nuclear reactions". It is possible to give an adequate introduction to this topic at school level, and we think that it is sensible to include such an introduction in a school syllabus. (At the moment, this and certain

/related.....

related topics have "square bracket" status in the Common Basic Syllabus for Physical Science.)

Since many of the reactions at present under study in nuclear physics research laboratories are initiated by charged particles which have been accelerated in a "particle accelerator", it is sensible to preface a discussion of nuclear reactions with some account of the principles of particle acceleration. There are two main points to be covered: the need for particle accelerators and the principles on which they work. Concerning the first point, reference can be made to Rutherford's alpha-particle scattering experiments and to the experiments in which certain elements were found to emit neutrons after bombardment with alpha particles. The bombardment of nuclei with alpha-particles led to an increased understanding of sub-atomic structure, so that it is conceivable that the technique of shooting the right kind of bullets at nuclear targets might be fruitful in yielding yet more information about the nucleus. If bombarding particles could actually penetrate the nucleus, their scattering pattern might be able to disclose something of the details of nuclear structure; also they might initiate nuclear reactions which could be studied in detail. Alpha-particles emitted by naturally radioactive elements have certainly shown their usefulness as sub-atomic bullets, but they have their limitations. If one simply takes the alpha-particles as they come, their energy is determined by their source, whereas it might be advantageous to have the energy of the bullets under direct control. Indeed it might be advantageous to give these "bullets" as much energy as possible, since they are positively charged and experience electrostatic forces of repulsion from the target nuclei.

The pupil might ask about the usefulness in this connection of neutrons, electrons, protons and other sub-atomic

/particles.....

particles besides alpha-particles. Neutrons are obvious candidates for use as sub-atomic bullets, since they are electrically neutral and would not be repelled by the positive charge on the nucleus; they have been used with some success in this field. High-energy electrons are useful for studying the distribution of electric charge in the nucleus. Electrons of sufficiently high energy have wavelengths comparable with the dimensions of the nucleus, and it is to be expected that the way in which the nuclear charge scatters these "electron waves" will yield information about the way in which this charge is spread through the nucleus. The successes achieved with this technique led to the construction of the two-mile long "linear" accelerator at Stanford University in the U.S.A., with which it is now possible to accelerate electrons to previously unattainable kinetic energies.

These are the terms in which the need for particle accelerators can be presented to pupils. It is necessary to describe only the basic principles of operation of these machines. First, it should be clear to pupils that the only feasible way to accelerate microscopic particles and thereby give them added energy is to subject them to electric fields. This means that one can only accelerate charged particles, such as electrons, protons, deuterons or alpha-particles; neutrons cannot be accelerated after emission from their source. Clearly any accelerator must have a source of particles to be accelerated; in any accelerator other than those specifically intended for electrons the particles employed will be ions - e.g. protons, heavy hydrogen ions (deuterons) or helium ions (alpha-particles). Therefore an essential requirement for such an accelerator is an ion source.

In principle, the simplest way of accelerating a charged particle is to subject it to an electric field of very high intensity applied over the greatest possible distance.

/Then.....

Then the energy acquired by the particle from the field is given simply by the charge on the particle multiplied by the potential difference through which it "falls". This principle is applied in the Cockcroft-Walton and Van de Graaff accelerators, of which there are several in South Africa, but there is a limitation on the energetic capabilities of these machines set by the fact that it is impracticable to generate potential differences of greater than a few mega-volts in the laboratory. In the most powerful machines at present in use, the same electric field is applied over and over again to the particle, giving it a little more energy each time. In "linear" accelerators, such as the two-mile one at Stanford already referred to, the same comparatively weak electric field is applied continuously to the particle over an appreciable distance until it has gained enough energy. The former method ((a)) can be compared to repeatedly hitting a ball forward as it moves along; the latter method ((b)) can be compared to continuing to push a trolley while it moves faster and faster. Method (a) is used in those types of particle accelerator in which the particles are made to go round in spiral or circular orbits. Magnetic fields are used to guide the particles into their orbits, and the accelerating electric field is applied to the particles either once or twice per orbit. The simplest of these machines, the cyclotron (one example in South Africa), and the "alternating gradient proton synchrotrons" at the C.E.R.N. laboratories at Geneva and the Brookhaven National Laboratory in the United States come into this general category.
Method (b) might appear to be no different in principle from that employed in Van de Graaff or Cockcroft-Walton accelerators. However, there is a considerable difference between the means of the application of the electric field to the particles in the two cases. In linear accelerators used to

/accelerate.....

accelerate electrons, the electrons ride along with electromagnetic waves which are arranged to have a component of electric field opposite to their direction of travel. The electrons in such a machine quickly cease to undergo acceleration in the kinematic sense, i.e. they are soon travelling with velocities exceedingly close to that of light. But their momentum and energy continue to increase as they travel down the track of the accelerator.

In its most general sense, the term "nuclear reaction" refers to any process in which the nucleus undergoes some form of change - a change in energy or a change in composition. In this context it is intended to apply to processes in which a target nucleus is bombarded by some or other "particle". In lieu of actual demonstrations of such reactions in the school laboratory, pupils should be shown cloud-chamber, bubble-chamber or nuclear emulsion photographs of the tracks of incident and product particles of such reactions. The main purposes in teaching pupils about nuclear reactions should be to show the operation of certain conservation laws and to explain the facts of energy release or absorption in these reactions. For these purposes it is only necessary to consider a few selected reactions. As an introductory example, that in which Rutherford succeeded in transmuting nitrogen into oxygen by alpha-particle bombardment will serve as well as any. (Cloud-chamber photographs showing this reaction are reproduced in Rogers (1960).) It can be pointed out that the sums of the atomic and mass numbers must be the same on both sides of the reaction equation, this being the consequence of two conservation laws, the conservation of mass number (sometimes called nucleon or baryon number) and the law of conservation of charge. The first of these is new to pupils, and it should be stressed to pupils that it is a new and distinct conservation law and not, say, a

/consequence.....

consequence of "conservation of mass". According to it, the number of nucleons at the end of the reaction should be the same as the number at the beginning; i.e. we can neither create additional nucleons nor destroy all or part of any nucleon in a nuclear reaction.

Other examples of interesting nuclear reactions are neutron-producing reactions (e.g. ${}^{4}_{2}He + {}^{9}_{4}Be \rightarrow {}^{12}_{6}C + {}^{1}_{0}n$) and the Cockcroft-Walton splitting of the lithium nucleus with accelerated protons (${}^{1}_{1}H + {}^{7}_{3}Li \rightarrow {}^{4}_{2}He + {}^{4}_{2}He$). The latter is of interest not only because it was the first example of so-called "atom-splitting" by artificially accelerated particles, but because it provides a direct experimental verification of Einstein's mass-energy relation. This brings up the important questions of how this relation should be introduced in such a course as this and how one should describe the energy transformations that occur in nuclear reactions. On the first of these questions, Guggenheim (1968) has written: "There is no reason why [Einstein's relation between internal energy and mass] should not be taught at school without any mention of the word 'relativity'." His article implies that this relation should be taught as an experimental fact; thus he writes: "Although this relation was brought to light through special relativity, it is independent of special relativity and could have been established by experiments such as that of Cockcroft and Walton." One begs leave to wonder whether he is right in claiming that the mass-energy relation is "independent of special relativity". It is true as Guggenheim points out that it is "independent of velocity", and that we could deduce it from experiments such as that of Cockcroft and Walton by postulating that the principle of conservation of energy must hold in all circumstances. But within the context of special relativity, we are led to expect that with a system of mass m there must be associated an amount

/of.....

of rest or internal energy mc^2, so that it is surely more accurate to regard the Cockcroft-Walton and other experiments of this type as experimental verifications of this relation, rather than as establishing it independently of special relativity. We can, at any rate, follow Guggenheim in his implied suggestion that we introduce the Einstein mass-energy relation by consideration of the energy balance of the Cockcroft-Walton reaction, pointing out that if the Einstein relation is assumed, we get energy conservation; but we feel it is no more than honest to point out to pupils that this relation can, in fact, be inferred from Einstein's theory. (In any case, many of them will have seen "$E = mc^2$" before and will have come to associate it with Einstein's name.)

Demonstration of the relation between internal energy and mass from the Cockcroft-Walton (or any other) nuclear reaction is indeed simple enough. It is found that the sum of the kinetic energies of the two alpha particles resulting from the reaction exceeds the kinetic energy of the incident protons by some 17 MeV. On the other hand, the sum of the rest masses of incident proton and lithium target nucleus exceeds that of the two alpha particles by 0.0186 u. Conversion of the latter to kilograms and multiplication by the square of the speed of light leads to a figure which, within experimental error, is equal to 17 MeV (expressed in joules). It can now be asserted that the gain in kinetic energy of the system is accompanied by a corresponding loss in internal energy of the system, equal in magnitude to the decrease in mass multiplied by the speed of light squared.

In our view, the teaching of this aspect of nuclear reactions should certainly not stop here, for if it did, pupils could easily be left with the impression that the release of energy in certain nuclear reactions is "explained" by the

/conversion.....

conversion of matter into energy, or that there is a special kind of energy called "nuclear energy" whose source is mass. (Cf Mossop (1968): "The ultimate source of nuclear energy is mass".) If this were to happen, it would be preferable not to discuss the energetic aspects of nuclear reactions at all. Fortunately it is quite easy to present the correct view. One must begin with the correct interpretation of mass-energy "equivalence", which can be stated as follows:- "If a system gains or loses internal energy of amount ΔE, its mass increases or decreases by an amount $\frac{\Delta E}{c^2}$." (There are those who would wish to remove the qualification that the energy gain or loss must be internal, so that if a system were to gain kinetic energy, its mass could also be said to increase by the appropriate amount. We pointed out in Section 3.7.2 that whether or not one wished to use a "relativistic mass" defined in this more general way was to some extent a matter of taste. In the present instance we are concerned only with changes in internal energy.) By "internal energy" is meant any energy that can rightly be regarded as being stored within the system, viz. rest energy associated with any particles that are present, kinetic energy of internal motion and mutual potential energy of the particles composing the system. If, therefore, a two-particle system has mutual potential energy and negligible kinetic energy, its overall mass must be equal to the sum of the rest masses of the particles and the mass of their mutual potential energy. As an example one may consider the deuteron, which is the nucleus of "heavy" hydrogen and contains one proton and one neutron. If the supposition is correct that the neutron and proton are subject to strong attractive forces when close together, one would expect the mutual potential energy of the two nucleons to be very much less when they form a deuteron than

/when.....

when they are very widely separated. Consequently the internal potential energy should have a large negative value with respect to a zero of potential energy that corresponds to the nucleons being far apart. Now the mass of the deuteron is equal to the rest mass of the proton plus the rest mass of the neutron plus the mass of the internal potential and kinetic energy. Suppose the magnitude of the (negative) potential energy exceeds that of the kinetic energy. Then the net internal energy is negative, and the mass of the deuteron must therefore be less than the sum of the rest masses of proton and neutron. Accurate data of the masses of the particles concerned show this to be the case.

In trying to explain this, the teacher could simplify somewhat by treating the internal energy as entirely potential. In view of the obscurity of what is happening inside the deuteron this would perhaps be pardonable. Another point to note is that the mass of the neutron cannot be determined independently of the mass of the deuteron or of Einstein's mass-energy relation with an accuracy comparable with that to which the masses of the proton and deuteron are known (cf Taylor and Wheeler 1966). We would therefore be on dangerous ground if we tried to use standard data on these masses to prove Einstein's relation; but since we are here concerned only with illustrating the application of this relation to the simplest multi-particle nucleus that there is, it will do no harm to use standard data for this purpose.

Generalizing from the case of the deuteron, one can now explain the discrepancy between the mass of a nucleus and the sum of the rest masses of its constituent nucleons by saying that when a number of nucleons come together to form a nucleus, the resulting loss of internal energy of the system is so great that the associated decrease of mass of the system is

/readily.....

readily detectable. It is also possible to explain in general terms why it is possible for energy to be released or absorbed in certain nuclear reactions. In any such reactions it is reasonable to expect that the rearrangement of nucleons that takes place must be accompanied by an overall change in mutual potential energy of the nucleons and that the overall total/internal energy of the system will be different. As a result, the mass will be different. If the final energy of the system is less than its initial energy, energy has been released in the reaction. If the masses of the incident particle, target nucleus, product nucleus and product particle are known, it is convenient to use them to predict the amount of energy released in the reaction, but it is a strange misuse of words to say that mass is the source of this energy. One might with equal justice say that mass was the source of the energy lost by a body when it cools down. It is no less misleading to say of common energy-producing nuclear reactions that in them, "matter is converted into energy." There are circumstances in which the use of this unfortunate phrase might almost seem justifiable, viz. reactions in which particles cease to exist and their rest energy appears as radiation. Here we have a disappearance of "matter" and the "production" of energy as a result; in such cases, mass may be said to be a source of energy, although one would prefer that this statement would not be made without an immediate explanation of what it really means. But in conventional nuclear reactions in which we simply have a rearrangement of nucleons, with no particles disappearing and no new particles appearing, it makes no sense to talk of matter being converted into energy. All that has happened is that the energy of interaction of the participating nucleons has altered. To refer here to matter being converted into

/energy.....

energy or vice versa can only be defended on the grounds of the most naive association between mass and "quantity of matter." (Frisch (1964) has made the same point in reference to the reaction in which two deuterons give rise to a proton and a triton: "It may be worth stressing that this is not a transformation of matter into energy; the components of the two nuclei are still the same and have merely become rearranged to form two other nuclei - just as atoms become rearranged in a chemical reaction - and energy is freed thereby, which causes the final nuclei to be a little lighter than the initial ones.")

Typical fission and fusion reactions should be quoted on account of their practical interest. Pupils can also be told something of what happens in nuclear reactions initiated by bombarding particles of extremely high energy, the main point of interest here being the creation of new particles from some of the energy made available in the reaction. In terms of conservation of energy, such processes can be described as the conversion of some of the available energy into rest energy of the new particles. In this connection it is possible to mention the existence of some of the bewildering array of "elementary particles" which have been identified.

With the consideration of these matters, we think that the limit has now been reached of what can and should be formally taught to pupils under the heading of "nuclear physics". Enough has now been said to enable them to derive beneift from further reading on the topic. It would be better for them to pursue the matter in this way than to listen to what must inevitably become more and more like a mere recital of facts as the teacher tries to go deeper into the subject. The Common Basic Syllabus lists artificially induced

/radioactivity.....

radioactivity and cosmic rays as suitable topics for optional extension. We think that both of these should be set simply as reading assignments.

CHAPTER 11: RELATIVITY IN THE SCHOOL SYLLABUS

11.1 Introduction

Special relativity has not yet become a part of school syllabuses in physics and physical science. It is now accepted that certain predictions of relativistic mechanics, viz. the variation of "relativistic mass" with velocity and the mass-energy relation, should be mentioned in school physics, but no firm attempt seems yet to have been made to include special relativity as such in school syllabuses. The primary purpose of this chapter is to discuss whether such an attempt would be worthwhile.

A point that should be made at the outset is that the scientific respectability of special relativity has long ceased to be in doubt. Not only have its predictions been amply confirmed by experiment, but they are constantly being employed with complete success in the design of new equipment and experiments. Occasionally the theory is attacked by those who claim to have discovered that it contains internal contradictions (cf Burniston Brown 1967 and Dingle 1967c) but their objections invariably turn out to be based on elementary misapplications of the fundamentals of the theory, and as such have value only as tutorial examples of the "find the fallacy" kind. (Of Dingle's opposition to the theory it can fairly be said that he has at last demonstrated to his own satisfaction that the brand of relativity he has been expounding for so long (cf Dingle 1940, 1956) is "untenable", to use his own favourite word in this connection. Indeed, Dingle's only error has been to confuse his own brand of relativity with Einstein's "Special Theory".) One would not therefore be "misleading the young" by formally incorporating relativity in an elementary physics syllabus.

Because of the fundamental importance of special relativity, there has in recent years been an increasing tendency

/to.....

to teach this subject at lower levels in universities than was hitherto thought safe. The many introductory textbooks that have recently appeared are an indication of the need that is currently being felt for clear presentation of special relativity at an elementary level. Experience in the Department of Physics at Rhodes University suggests that special relativity can be successfully taught in the second or even in the first year in a South African university provided that enough time and tutorial attention is devoted to it.

Can it be successfully taught at a lower level and ought one to try? Concerning South African school syllabuses, Professor J.A. Gledhill (1964) has remarked that while "the fundamentals of physics appear to be reasonably well covered... There is an exception to this - the theory of relativity. While this is fundamental to physics, it is not taught in schools." He goes on to say that "While I am sure that it will eventually have to appear in the school syllabuses, I very much doubt whether it will do so in the next <u>ten</u> years, let alone five." The assumption underlying these statements is clearly that it is <u>desirable</u> that special relativity should eventually find a place in the school curriculum. The idea is not new. Westaway (1929) suggested forty years ago that "Inasmuch as relativity is transforming our fundamental notions of physics, and therefore of science generally, it is desirable that at least an outline of the subject should be presented to Sixth Form boys." This argument is given in its most extended form by Rekveld (1965), who forcefully puts the case for teaching special relativity in schools. After asserting that "the choice of a certain topic should be motivated by pointing to the specific and necessary contribution it adds to the understanding of the picture that modern man has made of the phenomena in his natural surroundings", he lists "three specific and necessary contributions which the teaching of relativity could provide." First, relativity is

/"a.....

"a typical example of a theory showing how man has been able to surmount his everyday experience, namely his biologically founded conceptions of space, time and motion." Second, although much time has to be spent on giving adequate coverage of the laws of Newtonian dynamics, these laws have restricted validity, and one should make it plain to pupils that they "do not present the last word in mechanics". Connected with the second point is the third: that courses in modern physics are bound to include mention of the "inertia of energy", and that it is fitting that the pupil should not only "know about" this, but also understand that it is a "logical consequence of the fundamental postulates of relativity theory."

These arguments are convincing. Special relativity is no mere ornament added to physics by the efforts of Einstein and others; it is part of the very foundations of the subject. Not only are "relativistic" effects commonplace in the laboratories of high energy physicists; they are part and parcel of our everyday lives, to the extent that the latter depend on the laws of classical electromagnetism. If, then, special relativity interpenetrates with our experience in this way and is fundamental to the outlook of the contemporary physicist, we ought to do something at school towards preparing the way for its appreciation at a higher level. At the very least we ought to teach physics in such a way as not to inhibit understanding and acceptance of special relativity at a later stage. More positively, we might note Rekveld's point that special relativity is a "typical example of a theory showing how man has been able to surmount his everyday experience" and try even at school to make this apparent to some extent. We can also agree with Rekveld that it is desirable, if possible, to demonstrate the origin of such results as the variation of relativistic mass with velocity and the mass-energy relation. We add a further point: the present writer's

/experience.....

experience of teaching relativity to university students suggests that there are few aspects of physics which stimulate the imagination of students as much as special relativity does. Once they have overcome the psychological hurdles which relativity appears to present, students seem to enjoy grappling with the intellectual challenge presented by some of the paradoxes of the subject and in arguing with one another about them. If a comparable degree of interest could be stimulated among pupils, the effort will have been worthwhile indeed.

Boulind (1957) has remarked that "[relativity] does not seem to enter naturally into any school physics syllabus." By this he presumably means that there were no obvious links between relativity and the type of school physics syllabus which was common at the time he was writing. He does say that "It may be mentioned briefly with reference to gravitation: the effect of the sun on light rays from stars and the anomalous behaviour of the planet, Mercury." But this would be of no real help to pupils, to whom the connection between gravitation and relativity could not be convincingly explained in a few sentences. On the other hand, there are links between special relativity and particle and nuclear physics, and the "Principle of Relativity" arises in limited form in Newtonian mechanics, so that here, too, the theory has a definite link with knowledge already acquired by the pupil. In this case we could use the link as a starting-point for the teaching of the theory (cf Section 11.4.1). Properly approached, special relativity can be made to seem as "natural" a part of the syllabus as, say, electricity.

11.2 The difficulties presented by elementary relativity

The arguments presented above suggest that it is desirable to teach special relativity at school; the feasibility of this must now be considered. The present writer has not come across any accounts of attempts that have been made in

/this.....

this direction partly, he suspects, because there have been few such attempts as yet. Perhaps the main reason for this timidity is a prevalent feeling that the topic is difficult - difficult to teach and difficult to understand. Rekveld (1965) accepts the alleged "difficulty". "It cannot be denied", he says, "that relativity theory forms a difficult chapter of physics"; but he adds: "We feel that our reaction to this fact should not be that we had better avoid teaching this topic! On the contrary: we should be encouraged by this circumstance to exert ourselves as much as we can to find convenient ways and means to bring these arduous conceptions before our students and make them understandable." We shall try below to pinpoint these "arduous conceptions" or difficulties and to show that although these are genuine difficulties, it is possible for them to be eased or overcome in elementary presentations of the subject.

No treatment of special relativity, however elementary, can avoid confrontation with the fundamental postulates of the theory, and acceptance of these is undoubtedly one of the "difficulties" experienced by the beginner. Unless he is properly prepared for it, the Principle of Relativity itself could be a stumbling-block to a pupil, while the Principle of the Constancy of the Velocity of Light conflicts with what the pupil has learned about the velocities of material objects and of sound signals. The actual derivation from these postulates of such predictions of relativistic kinematics as the relativity of simultaneity and time dilatation need pose no serious mathematical or physical problems for the pupil. On the other hand, the solution of problems (other than "plug-in" ones) based on these results is found awkward by the beginner, mainly because of the initial difficulty in adjusting to the need to allow for relativity of simultaneity when going

/from.....

from one frame of reference to another. In the present writer's experience, even the better students have trouble to begin with in trying to cope with this type of problem.

11.3 The use of inertial frames

We suggested in Section 4.6 that explicit teaching of the concept of an inertial frame was now necessary in teaching elementary classical mechanics, and implied there that this concept might present some difficulty to pupils if explicitly taught. However, "inertial frames" are indispensable to any satisfactory exposition of special relativity, so that this matter must be dealt with at the onset of any relativity course. We are inclined to feel that once the pupil has studied classical mechanics and electromagnetism, he should be ready for the concept of an inertial frame. Formal introduction of it might be appreciated by him as bringing into the foreground something that he now recognizes as always having been in the background of what he has been taught.

Certain sophisticated expositions of special relativity (e.g. those of Taylor and Wheeler (1966) and Woolfson (1968)) use "Einsteinian" inertial frames from the beginning (i.e. frames defined locally by freely-falling particles). We believe that this approach would only confuse the average South African matriculation pupil, since it requires him at the start of a new and none too easy topic to alter his standpoint on gravitational forces. It is probably wisest in a school course to stick to the conventional method of presentation of special relativity and use "Newtonian" inertial frames, which have indefinitely large spatial extent and relative to which gravitational forces are treated as real forces. This kind of inertial frame is what has been implicitly used by the pupil up to this point, so that the presentation of special relativity in terms of such frames would

/in.....

in this respect follow smoothly on what has gone before.

11.4 The postulates of special relativity

11.4.1 The "Principle of Relativity"

It would seem that the teacher of elementary relativity must be prepared for the possibility that his pupils or students may find the fundamental postulates of special relativity hard to accept. A possible remedy for this difficulty has already been hinted at, viz. adequate preparation of the pupil for the acceptance of these postulates. One way of preparing him would be to draw his attention to the "Principle of Relativity" that is implicit in Newtonian mechanics, so that he will be ready in due course to consider the suggestion that this Principle might be valid for all physics. This is essentially the approach so strongly urged by Bondi (1965a,b, 1966, 1967). As he says, "... surely, it would be good teaching to emphasize the continuity with earlier thought" (Bondi 1966). By "drawing the attention of the pupil to the Principle of Relativity" we mean that in the course of learning about mechanics, the pupil should be led to think about whether he could carry out a dynamical experiment in, say, a Jumbo Jet, which will enable him to tell whether or not the vehicle is "moving". Our experience of discussing this point with university students suggests that the pupil might not easily be convinced of the correct view. (As Bondi (1967) remarks, "the really difficult step is the step taken by Newton and Galileo, and not the step taken by Einstein.") But if he can be convinced, it should be less troublesome to convince him that "the principle of the unity of physics requires that systems that cannot be distinguished by internal dynamical experiments should be indistinguishable by any internal experiments" (Bondi 1965a). Naturally one would not put the

/point.....

point to the pupil in such sophisticated language, but the idea itself is not a difficult extension of "Newtonian relativity". One could lead up to it by using Bondi's argument that there is no such thing as a purely dynamical experiment or phenomenon; "even in the prototype of all dynamical phenomena, the collision between two billiard balls, you rely on the material properties of billiard balls, which as we know are due to complicated quantal electromagnetic forces" (Bondi 1967).

In advocating that pupils should be introduced to the "Principle of Relativity" by way of Newtonian dynamics, we are not suggesting that there should be explicit and detailed teaching of "Galilean" or "Newtonian" relativity, including a derivation of the Galilean transformation. We do not think that any benefits are likely to accrue to the pupil from acquaintance with the latter, which might actually be harmful at this stage. Rekveld (1965) has reported that "there are some teachers who are inclined to assume that a confrontation with classical relativity could reinforce the vague conceptions of our students about absolute time and mass. Consequently the acceptance of the revolutionary ideas of Einstein might perhaps rather be hindered than prepared by this discussion." But if the subject is introduced in the manner we have suggested above, the implications for space and time of the indistinguishability of inertial systems by means of "purely dynamical experiments" will not be discussed at all, so that there is no chance of pupils having their "vague concpetions" reimforced. (To undermine rather than reinforce these "vague conceptions", the teacher could while teaching the fundamentals of kinematics and dynamics warn against making unwarranted assumptions about space and time, especially the latter.)

/11.....

11.4.2 The constancy of the velocity of light

Bondi (1965a,b, 1966, 1967) also draws attention to the implications of the "principle of the unity of physics" for the velocity of light: the latter must be the same in all directions relative to any inertial frame of reference, independent of the velocity of the source. Actually this is not an inevitable conclusion from the premise that "[inertial] systems cannot be distinguished by any internal experiments". If the laws of electromagnetism are correctly given by Maxwell's equations, then the Principle of the Constancy of the Velocity of Light does indeed follow from the Principle of Relativity (cf Rosser 1967). (As Rosser (1967) points out, Einstein could, in fact, have taken as his second postulate that the laws of optics and electromagnetism were Maxwell's equations.) Other formulations of electromagnetism, such as those of Ritz (cf Fox 1965), in which it is presumed that the velocity of light depends on the velocity of the source, are also possible; indeed the Ritz theory is actually in conformity with Galilean relativity. It is partly the impressive experimental support for special relativity that sustains one in the view that classical electromagnetism is correctly described by Maxwell's equations. One can hardly discuss rival theories of electromagnetism in an elementary treatment of special relativity; what one can do is to remind the pupil that in the treatment of the generation and propagation of electromagnetic waves, no mention was made of any influence of the velocity of the source on the propagation of the waves, and to tell him that the full theory implies that there is no such influence. The similarity between this case and that of sound can be pointed out. Thus the pupil can be led to accept as reasonable the principle that "in a vacuum light is never overtaken by light", which is Bondi's

/(1966.....

(1966, 1967) non-metrical formulation of the postulate that the velocity of light is independent of the velocity of its source.

The teacher might wish to back up this assertion with some mention of relevant experimental evidence. The evidence often cited in this connection (e.g. by Bondi 1966, 1967) consists of the results of de Sitter's measurements on the spectra of double stars. However, this may not be as clear-cut as has long been assumed: Fox (1962, 1965) has argued that the effective sources of the light received on earth are not the double stars themselves but the interstellar matter through which the light has passed. In recent years a good deal of work has been done on checking the "source independence" postulate by measuring the velocity of gamma-rays "emitted" by fast-moving particles. Dingle (1967c) claims that these experiments "all involve a circular argument" in that they supposedly invoke special relativity at crucial points in their interpretation. It is true that in the interpretation of these experiments (cf Farley et al 1968) it is customary to use the relativistic formula for the Doppler shift to deduce the velocity of the source of the gamma-rays; but these sources are produced in high-energy collisions in which the velocity of the centre of mass is a substantial fraction of the speed of light so that, as Farley et al remark, "it can hardly be doubted that the gamma-rays come from a moving source." The measurement of the velocity of the gamma-rays was carried out by a time-of-flight method which does not seem to depend on any assumptions derived from special relativity, and it was found that within the limits of experimental error (1.3 parts in 10^4) the measured velocity of these gamma-rays is in agreement with the accepted value of the velocity of light. We may take it, then, that these

/experimental.....

experimental results are consistent with the "source independence" postulate. As the principle of these experiments is straightforward, the teacher could mention them briefly, without giving any experimental details.

The other aspect of the "Constancy of the Velocity of Light", viz. that the velocity of a light signal is the same in different inertial frames, is less easy to make plausible to the pupil, since it is not in accordance with the pupil's previous experience of the behaviour of waves. A possible solution is to point to the experimental evidence. It is generally held that the relevant evidence is provided by the Michelson-Morley and the Kennedy-Thorndike experiments (cf Robertson 1949, Grünbaum 1955, Taylor and Wheeler 1966). (Specifically, experiments of the Michelson-Morley type establish that the round-trip speed of light is isotropic in different inertial frames, and the Kennedy-Thorndike experiment establishes that the round-trip speed of light has the same numerical value in different reference frames.) However, there are two possible objections to detailed discussion of these experiments in an elementary course on relativity. First, such a discussion would take up a considerable amount of time. Second, as Bondi in particular has pointed out (cf Bondi 1959, 1965a,b, 1966, 1967), these experiments come close to being "Zanzibar" experiments (cf. Section 2.5.) although this could not have been apparent when they were first carried out. Bondi's argument rests mainly on the fact that "the size of our measuring rods is determined by atomic interactions which themselves are fully characterized in principle by the atomic frequencies" (Bondi 1959). Hence the lengths of measuring rods are "determined by a superposition of radar methods" (Bondi 1965a), i.e. by the time of travel of light signals. But the essence of experiments of

/the.....

the Michelson-Morley type is to compare the travel times of light signals over paths of supposedly fixed length; if the lengths of these paths are fundamentally determined by the travel times of light signals, then it is not at all surprising that these experiments should have established that the speed of light in different inertial frames is isotropic and has the same numerical value.

Apart from this objection, Bondi also makes the point that in the performance of a Michelson-Morley experiment it would be desirable to ensure that the distance from the source to the mirrors is kept constant, and that it seems to be impossible to conceive of a check on this aspect of the experiment which does not depend on the properties of light. In the light of this criticism, it is interesting to consider a recently-performed modern version of the Michelson-Morley experiment (cf Jaseja et al 1964). In this experiment the two legs of the Michelson interferometer were replaced by He-Ne masers operating in the infra-red region. The frequency of each maser depends on the spacing of its mirrors and the effective velocity of light in the space inside it; the presence of an "ether drift" should therefore cause a change in the beat frequency between the masers as the apparatus is turned through 90°. The experimenters apparently did not try to check that the lengths of the masers were unchanged by the rotation. They do discuss the effect on the frequency spread of the maser output of the thermal vibrations of the spacers between the maser mirrors, and also consider the effect on the spacers of magnetostriction as a result of the earth's magnetic field. Evidently they assume that only effects of this type or a possible anisotropy of space could affect the effective lengths of the masers. This is reasonable, although it would have been better to have had an independent check; in this case, direct measurement of the frequency of

/the.....

the masers could not have been used for this purpose, as this frequency was about 3×10^{14} Hz. (From their results, these experimenters concluded that "there is no anisotropy or effect of ether drift larger than one-thousandth of the small fractional term $\frac{v^2}{c^2}$ associated with the earth's orbital velocity.")

We incline to the view that Bondi's main criticism does not entirely invalidate modern repetitions of the Michelson-Morley and Kennedy-Thorndike experiments. (Bondi has made it clear that he does not wish to minimize the historical importance of the original experiments of this type.) At the very least, null results of the Michelson-Morley and Kennedy-Thorndike experiments can be regarded as confirmation of the self-consistency of our view of nature. The beginner who is groping his way into special relativity might be glad of the assurance provided by this confirmation; but in a school course in relativity, there is unlikely to be enough time to allow for an adequate treatment of these experiments. We would accordingly recommend that these experiments be given only the briefest possible mention in the classroom; interested pupils can read about them if they wish.

It may be thought that Bondi's above-quoted "operational" argument concerning measuring rods provides the teacher with a convenient justification of the constancy of the velocity of light: the lengths of our measuring rods are effectively determined by a "superposition of radar methods", so that any measurements we make of the velocity of light with the aid of these rods are bound always to yield the same answer. There are at least two reasons why it would be inadvisable to use this argument in an elementary course. First, the argument is subtle - it might not be easy to persuade a beginner that the length of a measuring rod really is determined by a "superposition of radar methods". Second, one may claim that it misses the real point at issue. Neither special relativity nor Maxwellian electromagnetism are based

/upon.....

upon any assumptions concerning the nature and structure of measuring rods or clocks, and it is therefore misleading to justify the constancy of the velocity of light with an argument which does depend on such assumptions. Surely the correct view to take is that space and time are so interrelated that the postulates of special relativity must always be valid in the appropriate circumstances.

We are still left with the problem of how to make the "constancy of the velocity of light" plausible to the pupil. A possible approach is to contrast the propagation of light with that of sound. The characteristic velocity of a sound wave in a given medium is evaluated relative to the medium: what is the velocity of light "relative to"? There is no material medium for light propagating in vacuo. One possibility is to assert that the velocity of light is expressed relative to its source (thus denying the validity of Maxwellian electromagnetic theory); but experiment shows that the velocity of light is independent of that of its source. Another possibility is to postulate an "ether" as the non-material medium for light propagation, relative to which the velocity of light should be specified. But this implies that one can determine one's velocity relative to a "special" reference frame (the ether) by an optical experiment, whereas this cannot be done by dynamical means. This contradicts the principle that "systems that cannot be distinguished by internal dynamical experiments should be indistinguishable by any internal experiments." If this principle is to be retained, it must mean the rejection of the idea of an ether, at least in the form of a medium whose presence can be detected by direct measurement. This in turn suggests that the same value should be obtained for the velocity of light in different inertial "frames".

/The.....

The previous paragraph gives the standard argument for the constancy of the numerical value of the velocity of light in all inertial frames. There is little chance of simplifying it further. If one is to try to convince pupils of this aspect of the "uniqueness" of light, then one has to try to suggest to them why light cannot be expected to behave like sound in all respects; in particular, the implications of the absence of a material medium for light waves should be clearly brought out.

The argument given above could be expanded by giving pupils a qualitative version of an argument based on the application of the Principle of Relativity to the laws of electromagnetism. As they stand, these laws imply that electromagnetic disturbances are propagated in vacuo with velocity "c" relative to the given inertial frame. If the laws of electromagnetism are the same in all inertial frames, then it follows that the velocity of light must be the same in all inertial frames. Saying that the velocity of light is not the same in all inertial frames is tantamount to saying that the laws of electromagnetism are not the same in all inertial frames, which is contrary to the "principle of unity of physics" (Principle of Relativity). Acceptance of this argument would depend on pupils accepting that the correct laws of electromagnetism are those they have been taught, and that it does follow from these laws that light travels with a velocity "c" relative to a given frame of reference. Although demonstration of the latter point cannot be included in the syllabus, it might comfort some pupils to know that such a deduction can be made, which might make them feel happier about the connections between the "principle of the unity of physics" and the constancy of the velocity of light.

Bondi (1965a) has given a summary of this kind of argument which might be helpful to a teacher: "Just as one can

/talk.....

talk as easily on a plane as on the ground, so one can read as easily, although this involves the propagation of light." So neatly does this express the essence of the matter that a teacher might feel tempted almost to say just this about the subject and leave it at that. However, this is a summing-up, not an argument: the sceptical pupil would want to know whether possible changes in the velocity of light brought about by the motion of the plane could affect his reading!

Bondi (1965a,b, 1966, 1967) actually prefers an alternative approach to the one we have outlined above. He bases this on the measurement of distance by radar methods, arguing that by adopting the velocity of light as a "natural standard of velocity" and setting this equal to unity, we reach the happy condition of measuring distance by time, taking as the unit of distance the light-year, light-second or light nanosecond, and "all the talk about the velocity of light dissolves into nothing." The yardstick is really a very complex system, far more complex than a radar set combined with an atomic clock; with the velocity of light set equal to unity by definition, "measurement of the velocity of light" becomes instead the determination of the length of one's yardsticks in "light-seconds" or similar units. Bondi claims that "the constancy of the velocity of light" is indeed obvious once it is agreed that distance measurements should be made by radar methods, partly, it would seem, because the velocity of light is now "a purely conventional (and arbitrary) number", and partly because of the requirements imposed by the "principle of the unity of physics". Bondi does not explicitly discuss how application of the latter principle establishes the point at issue. One can do this by discussing the question that is raised by the above account, viz. how can we be sure that the radar method of distance

/measurement.....

measurement will give consistent answers for the length of a given object if the measurement is carried out first in one inertial frame, then in another which is moving with respect to the first? For an answer one can appeal to the "principle of the unity of physics" - if the answers for the length were different, this would be a way of telling one's "absolute" velocity by a purely internal experiment.

Bondi's argument looks simpler than the "standard argument". But the teacher who decides to present it will probably find that he has to back it up by discussing the "uniqueness" of light in much the same terms as we have done in presenting the standard argument, so that "all the talk about the velocity of light" will not quite "dissolve into nothing".

The teacher may wonder what effect, if any, the recent re-definition of the standard metre in terms of the wavelength of a standard light source has on the teaching of the "constancy of the velocity of light". The answer here is that it really has no effect on the basic point at issue, viz. the "uniqueness" of light. If any pupil were curious about this, the teacher could help him by discussing the implications of defining length standards in terms of the wavelength of a source of sound.

We conclude that the principle of the constancy of the velocity of light can be rendered plausible by careful presentation, provided that the pupil can accept that internal dynamical experiments cannot disclose what the observer's velocity is relative to any other frame, and that it would be very odd if any other form of internal experiment could provide this information. The "constancy of the velocity of light" may still be difficult for pupils to accept, but at least they should be willing to consider its consequences.

It may worry some teachers that they should be expected to introduce the postulates of special relativity with very little experimental support, even of the "reported experiment"

/variety.

variety. But it should be remembered that Einstein, the prime creator of special relativity, did not lay down its postulates as inevitable generalisations from experiment. He stated (cf Einstein 1905, 1923) that the unsuccessful attempts to discover any motion of the earth relative to the "light medium" suggested that the same laws of electrodynamics and optics would be valid for all frames of reference for which the equations of mechanics held good. He then "raised this conjecture ... to the status of a postulate" and introduced another postulate, viz. that the velocity of light was independent of the motion of the source. Thus the first postulate was suggested, not compelled, by the experimental evidence available at the time, while convincing direct evidence for the second postulate became available only after Einstein had died (see above). In formulating special relativity, Einstein evidently acted on the conviction that his two postulates must be valid, a conviction that was based not on clear-cut experimental evidence but on profound contemplation of the established theories of the day and of what had been established experimentally. We are not obliged to follow history in our teaching of physics, but from history we may learn that we give a false picture of the way science works if our teaching implies that theories are created only on the basis of well-established experimental evidence. "Conjecture", as we have seen, may play an important part (cf also Maxwell's postulation of "magnetoelectric induction") and we would do well to show pupils how plausible conjecture can form the basis of a theory which afterwards turns out to be amply confirmed by experiment.

11.4.3 Relativity and history

The presentation of the postulates of relativity suggested in the preceding sections has not been based on an historical approach, essentially because we believe that an

/historical.....

historical approach does not in this case bring with it any special advantages. Cf. Rekveld(1966): "It is true... that a discussion of the historical process motivates the need for a new and revolutionary theory, but it does not in the least prepare or promote an understanding of relativity proper"; and Rosser (1967): "The author has found that, with elementary students, it is better to proceed as directly as possible from Newtonian mechanics to the theory of special relativity. With such students it is hardly worthwhile cultivating the confusing ether points of view in detail, merely to replace them with the conventional point of view." In the preceding sections we have seen how one could lead in to special relativity from previously acquired knowledge, so that the introduction of the subject to pupils need not seem arbitrary or forced in the absence of an historical introduction. This is just as well, as the time available for the teaching of relativity in any future South African school syllabus is not likely to allow anything resembling an extended treatment of the historical background of the subject.

11.5 Problem-solving in elementary relativity

We remarked in Section 11.2 that beginners in relativity are likely to find problems in it hard to solve. In our view, the only way to deal with this situation is not to set such problems to any but the brightest pupils. The extent to which someone can cope with problem-solving in special relativity is certainly an excellent means by which to gauge his understanding of the subject, but the experiences of university students suggest that we cannot reasonably expect that the understanding of special relativity by the average pupil can go very deep. The best we can hope for is that a way to an ultimately "deep" understanding will have been successfully paved in the classroom. This view may appear unduly pessimistic. Cf. Bondi (1967), for example: "I have always said

/that.....

that my ultimate aim is to get special relativity into the primary school syllabus. What one wants to do in order to achieve this has also seemed to me quite simple. If somebody could devise a cheap and safe toy which is an accelerator fast enough to show relativistic effects so that children of five, six or seven could play with it, then special relativity would be regarded as obvious and a suitable subject for the primary school curriculum." No doubt Bondi is consciously exaggerating here. Many of the toys children play with provide happy illustrations of the principles of Newtonian dynamics, yet no-one has tried to capitalize on this by introducing Newtonian dynamics into the primary school -- nor is it likely that anyone ever will. Presumably Bondi is trying to dramatize his view that special relativity is simpler than one might think, and perhaps he is right. But there are limits to what can be assimilated and understood at a given age and, though nothing that could be called "scientific evidence" is yet available on this point, it does seem as though it is unrealistic to expect more than a rather superficial understanding of special relativity from school pupils. Certain facts should serve as a salutary warning in this connection. We noted in Section 11.1 that certain physicists have claimed to have shown that special relativity is self-contradictory. If physicists of the standing of Dingle and Burniston Brown can persuade themselves of this as a result of confusion about what special relativity has to say about clock synchronization, we should allow for the possibility that pupils' grasp of the fundamentals of special relativity may not be too secure.

Until quite recently physicists who accepted special relativity seem generally to have assumed that the length contraction was a phenomenon which, in principle, could be "seen";

/that....

that is, if one took a photograph of a fast-moving object, the latter would appear on the photograph to have been contracted in the direction of motion. This idea was uncritically accepted and described and illustrated in many books, popular or otherwise, until Terrell (1959) and others pointed out that it failed to take into account the fact that light signals reaching the camera at the same instant from the ends of the object must have started out at different times, and that if this were taken into account, it would be realized that the object would appear undistorted in shape but rotated. Once this was pointed out, it seemed obvious enough and in one's wisdom after the event, one wondered why it had taken so long for this point to be appreciated. That it did take so long shows that special relativity does after all demand very careful thought if its implications are to be properly grasped. Evidently we should not set an unreasonably high standard for our pupils in this respect.

11.6 The teaching of relativistic kinematics

11.6.1 The relative velocity of inertial frames

Whichever way one wishes to teach relativistic kinematics, one must make use throughout of the result that if an inertial frame S has a velocity $\vec{v}$ relative to another inertial frame S', S' will have a velocity $-\vec{v}$ relative to S. No doubt the average pupil would be happy to accept this - it agrees with his experience and what he has learnt in classical kinematics - and it would probably be wise of the teacher to treat this point initially as though it were self-evident. (In this he would be following the practice of many textbooks on the subject.) But the bright pupil who succeeds in due course in grasping some of the predictions of relativistic kinematics may become uneasy about it. It has been impressed upon him

/that.....

that time and space measures are not the same in different inertial frames; on what grounds can he be certain these measures nevertheless combine in such a way that a pair of inertial "observers" will always agree on the magnitude of their relative velocity?

To answer such a pupil, one must appeal to the principle of relativity (cf Smith 1965). Suppose the inertial observers did get different values for their relative speed. Then the principle of relativity requires that these values must satisfy the following condition: if v_A and v_B are the values of the relative speed measured by A and B respectively, then $v_A = F(v_B)$ and $v_B = F(v_A)$, where "F" denotes the same function in each case. For if they did not meet this condition, e.g. if A and B agreed that v_A was bigger than v_B, then there would be a means of telling the two frames apart by a physical experiment, which is contrary to what the principle of relativity asserts. (This situation can be compared with the reciprocity of "time dilatation" and "length contraction"; each observer finds that the other's clocks "go slow" and that his measuring rods are contracted in the direction of relative motion.) The only function which yields a physically meaningful solution of the above pair of equations is just the variable itself, i.e. we have $v_A = v_B$.

We feel that there is no special need for this argument to be given at the beginning of a treatment of relativistic kinematics, but that it could be introduced after the pupil has had the time dilatation and length contraction formulae derived for him.

11.6.2 Should one start with the Lorentz transformation?

Conventionally, the teaching of relativistic kinematics begins with a derivation of the Lorentz transformation. The case for beginning in this way has been summarized thus by Sears and Brehme (1968): "... the Lorentz transformation

/is.....

is relativity. Unless one understands the meaning of the transformation,he cannot understand relativity". Bondi (1967), on the other hand, refers to "another old myth, namely that you must understand the Lorentz transformations in order to understand relativity."

The fundamental significance of the Lorentz transformation for special relativity is undeniable, but it is going too far to imply that one cannot acquire some understanding of special relativity without understanding, or knowing about the Lorentz transformation. Bondi (1966) has strongly argued the case for not beginning with the Lorentz transformation. "My own experience", he writes, "even with mathematics students, indicates to me that although they are very much used to working with coordinates, a coordinate-free approach tends to go home more readily." He points to E.A. Milne's distinction between a "world map" and a "world picture" of events: "the world picture is simply what the observer sees at one instant, so that it refers to distant events at an earlier time and to near events at a later time", whereas the world map, which is essentially a coordinate representation of events, "is made.... by a survey, as it were, putting on one sheet all the events to which the observer assigns the same time". Finally, he points out that "there is a tendency throughout the whole subject for the world picture to be simpler than the world map in relativistic theory." For this reason, he adopts a world-picture at the beginning, i.e. a non-coordinate approach. We shall examine Bondi's approach in due course. In the meantime, we note his observation that the coordinate-free approach succeeds better initially than the coordinate approach, and believe that this is likely to apply even more to South African matriculants, whose mathematics is not particularly sophisticated and who would probably

/be.....

be bewildered and discouraged by a presentation of special relativity which starts by concentrating on the transformation between coordinate systems.

11.6.3 Derivation of some of the standard results directly from the fundamental postulates

The main short-term aim of teaching relativistic kinematics is presumably to derive its most striking predictions, viz. the relativity of simultaneity, the dilatation of time and the contraction of length. All of these can be derived directly from the fundamental postulates of the theory, i.e. without reference to the Lorentz transformation. For example, the time dilatation formula can be derived by consideration of the time elapsed between the emission of a light signal from the origin of an inertial frame and its return after reflection at a distant mirror (Katz 1964, Smith 1965). The derivation depends on the result that a length measured at right angles to the relative velocity between a pair of inertial frames is the same in both frames (this is a consequence of the principle of relativity - see below), and on the constancy of the velocity of light. Once these premises have been accepted, the desired result follows easily from the geometry of the situation. It is therefore tempting to use this derivation in an elementary presentation of the subject. However, the first of the above-mentioned premises would not (and indeed should not) be regarded as obvious by a pupil, so some thought must be given to the matter of how to present it convincingly. (In fact, this result has to be used somewhere in virtually any presentation of special relativity one chooses to follow, so the teacher of elementary relativity can scarcely avoid facing this problem.)

In the most extended of his popular expositions of relativity, Bondi (1965a) discusses this problem as follows:-

/...let.....

"... let us suppose that Brian has a ruler marked in feet and held at right angles to the direction in which he sees Alfred, and that Alfred has a similar ruler held parallel to Brian's, and therefore also at right angles to the direction in which he sees Brian. [Brian and Alfred move at constant velocity relative to each other.] Then a foot in Brian's ruler looks a foot to Alfred, or vice versa, for if Alfred moves a foot along his ruler, then in the direction at right angles to his ruler there will be the foot mark on Brian's ruler. Thus there is no difficulty in translating distances along these two rulers into each other - they are simply the same". This sounds very plausible, but - with the greatest respect to Bondi - it is difficult to escape the conclusion that what is supposed to have been proved has really been assumed from the start: now do we <u>know</u> that "if Alfred moves a foot along his ruler, then in the direction at right angles to the ruler, there will be the foot mark on Brian's ruler?" (In fairness to Bondi, we must point out that he has implicitly acknowledged the inadequacy of this approach, because in his 1966 article on the teaching of special relativity (Bondi 1966) he remarks that "the fact that the coordinates at right angles to the direction of motion are not changed is... rather more difficult to present convincingly [than the one-dimensional Lorentz transformation]!")

A more carefully considered approach to this problem deals with two rods with identical rest-lengths which are transversely mounted in their respective inertial frames, S and S' , and are allowed to move past each other (Kittel et al 1965, Smith 1965). For argument's sake it is supposed that one rod will appear shorter in the other frame, and that it is arranged that as the rods pass each other, the ends of the shorter rod will make marks on the longer rod. This makes it possible to say that the "moving" rod is definitely shorter than the "stationary" rod. But the principle of relativity requires that the rod that appears shorter in the one frame should

/appear.....

appear longer in the other frame. This cannot be reconciled with the physical record that the one rod is actually shorter than the other. Therefore the lengths of the rods must appear to be the same when viewed from either frame.

This argument is correct as far as it goes, but it fails to make explicit one important point. If a similar experiment had been conducted with two measuring rods aligned parallel to the direction of relative motion of the two frames, a physical record could have been obtained apparently showing that one rod is shorter than the other; what is more, this "physical record" is reconcilable with the fact that in each frame, the rod at rest in the other frame appears shorter. This apparent contradiction is easily resolved when the role played by simultaneity is taken into account (cf Smith 1965, Taylor and Wheeler 1966 and many other sources for full discussions of this point). How, if at all, does simultaneity enter into the case under consideration? First we note that clocks mounted along a straight line in S', say, perpendicular to the direction of relative motion of S and S', will appear synchronized in S if they have been synchronized in S'. (This follows readily from Einstein's criterion of synchronization (Einstein 1905) and the constancy of the velocity of light.) Hence observers in S and S' will agree that the ends of a rod transversely mounted in one frame are simultaneously marked on a similar rod in the other frame. This justifies one in applying the argument cited above, viz. that a physical record showing one rod shorter than another would be a violation of the principle of relativity. Accordingly, it is misleading to say that "With the metre stick along the y-axis, we did not have to worry about questions of simultaneity in the comparison of a moving and a stationary metre stick" (Kittel et al 1965). Simultaneity is surely implicit in any comparison of lengths between a "moving" and a

/"stationary".....

"stationary" rod, whether the rods are transversely or longitudinally placed with respect to the direction of their relative motion. It is just that no problems arise in connection with relativity of simultaneity when "the metre stick is along the y-axis".

To make clear the role played by simultaneity in the case of length comparisons perpendicular to the direction of relative motion, some treatments (e.g. Panofsky and Phillips 1962, Arons 1965, Rosser 1967) use thought experiments in which the marking of the positions of the ends of the rod is triggered off by light signals from its mid-point. Then the Einstein criterion for synchronization or simultaneity can be directly applied in the course of the analysis.

It would appear that the teacher who wants to give a reasonably rigorous derivation of the time-dilatation formula based on the reflection of light signals, has first to take up the matter of the comparison of lengths perpendicular to the direction of relative motion of the reference frames. From the foregoing discussion it can be inferred that this might be a tricky issue to handle in the classroom. The teacher might be tempted to gloss over the subtleties in the argument, or even to try his luck at deriving the time dilatation formula without explicitly mentioning the matter of length comparison. This would undoubtedly be dangerous. In view of the startling nature of some of the predictions of special relativity, brighter and more sceptical pupils in the class are likely to watch very closely every step in the arguments presented to them, and to challenge anything that seems suspect. The teacher committed to a "simpler" approach may find himself pulled up short by challenges on the points discussed above, and if he has not budgeted for this kind of thing, his schedule of presentation may be seriously and

/perhaps.....

perhaps disastrously interrupted. (One need hardly add that the teacher who cannot cope with such challenges is unlikely to be able to do the cause of special relativity much good.) Thus he might find it better after all to go into some detail on the subject of length comparison, perhaps along the lines of the approach of Smith (1965) and Kittel et al (1965). Like many other aspects of the teaching of special relativity, this is certainly a point worth investigation in the reality of the classroom.

Once these difficulties have been overcome, it is necessary for the teacher to make the interpretation of "time dilatation" as clear as he can. This is a crucial step in the teaching of special relativity, as misconceptions acquired at this stage about what relativity is alleged to say about time and clocks may prove impossible to root out later on. Such misconceptions may be fostered by what the pupil reads in popular books on the subject. Consider the following statements on time dilatation taken from one such book (Coleman 1959): "If A then looks at his clock and compares its reading with B's, he will be surprised that B's clock appears to be running slower. This is precisely what the Special Theory predicts, since mathematical results show that the times given by the clocks are related by.... $t' = t\sqrt{1 - \frac{v^2}{c^2}}$, where t' is the time A reads for B's clock, and t the time A reads on his own clock." Nothing is said by the author of this book about the process by which A compares his clock reading with B's or even about what exactly is meant by "the time A reads for B's clock." But without an understanding of these points, the reader cannot have any real idea of what the author is talking about.

Perhaps the first point to be emphasized in any discussion of time and relativity is that the predictions of the time dilatation formula can apply to the actual readings of actual clocks. Authority for this statement comes from

/Einstein's.....

Einstein's first paper (Einstein 1905, 1923), in which the discussion throughout is formulated in terms of measuring rods and clocks and the readings on the latter. Einstein makes the point perfectly clear in his discussion of the definition of simultaneity. Here he refers to the determination of times at points A and B by means of clocks located at those points, and then describes the process of synchronizing these clocks by means of light signals. In other words, he is concerned with the synchronization of the actual readings of clocks at different places. The point may seem obvious enough, but it has been denied by one former adherent of the theory. Dingle (1956) has contended that the time assigned to an event by an observer not coincident with it "is essentially unobservable"; one can only calculate it by applying Einstein's criterion of synchronization, and the result of the calculation will not in general agree with a reading of any clock actually present at the event. There is no warrant for this claim in Einstein's paper, and it is small wonder that Dingle should eventually have shown that arguing from such premises as these must inevitably lead to contradictions (cf Section 11.1). The truth of the matter cannot be too strongly emphasized, and opportunity for bringing it out further arises in connection with the next "teaching point" to make: that time dilatation is concerned with the comparison between the 'proper" and "non-proper" time intervals between events; that is, between the time interval between these events as recorded in a frame of reference in which they occur at the same place, and as recorded in a frame of reference having a non-zero velocity relative to the first. To measure a "proper time", only one clock is needed; to measure a "non-proper time", one needs two synchronized clocks. Once the pupil has grasped this distinction, it should be apparent to

/him.....

him how the familiar statement "a moving clock runs slow" must be interpreted, and the confusions that arise in connection with it can be avoided (cf Smith 1965). The present author's experience suggests that this pedagogical use of the concept of "proper time" would help to clarify matters for the pupil and that, although the latter might have difficulty in applying these ideas in certain circumstances, he might at least grasp and remember the fact that "the proper time is always less than the non-proper time."

The remaining results from relativistic kinematics of interest to the pupil are the length contraction and the addition formula for collinear velocities. The former can readily be derived from the time-dilatation formula; Eisberg (1961) gives a particularly simple derivation which would be eminently suitable for an elementary course. The velocity-addition formula can be derived at this level by an "intelligent guesswork" method which is based on the results that the formula must reduce to a simple addition at low velocities and that it must be symmetrical with respect to the interchange of velocities in the formula (cf P.S.S.C. 1968).

The above scheme of presentation of relativistic kinematics does seem as though it might be workable at school, subject to the difficulties mentioned above. Apart from the latter, its main drawback is that it operates very much in the realm of highly-idealized thought experiments, with its use of light signals and sets of synchronized clocks.

11.6.4 Bondi's K-calculus

An alternative approach to relativistic kinematics which has recently come into prominence is Bondi's "K-calculus." Bondi has published simplified expositions of it in recent books (Bondi 1965a,b, 1967); more sophisticated accounts are given by Bohm (1965) and Rosser (1967). The K-calculus develops special relativity in terms of a "one-dimensional set

/of.....

of inertial observers" who, in an elementary treatment, can be supposed to be travelling in space ships. Communication between them is by means of radar signals. Suppose that we consider two of these space ships which happen to be receding from each other with constant relative velocity v. Let one of them, A, be sending out radar signals at regular time intervals T as recorded on his own clock. Because the space ships are moving apart, we would expect that when these signals are received by the other space-ship, B, they will be separated by a somewhat longer time interval as recorded on B's clock. Let the time interval between successive signals received by B be KT, as measured by B. According to the postulates of special relativity, the factor K can only depend on the relative velocity of the two space ships. This implies that if B sends radar signals to A, and these signals are separated by time intervals T' according to B, they must be received by A at time intervals KT' as recorded by A. It is this K-factor which gives its name to Bondi's method of presenting relativistic kinematics.

The formula for time dilatation can be derived very simply with the aid of the K-calculus. Suppose space ships A and B agree that when B receives a signal from A, he will send one back, and suppose that the first such exchange is made when A and B coincide. After a time T on his clock, A transmits another signal, which B receives after a time KT on his clock, and his return signal to A is received by the latter at a time K^2T. B measures the time between his receipt of the two signals from A as KT; this is the proper time between these "events" (i.e. the arrival of the signals at B). A was present at the first event but not at the second, so he has to compute the time at which, according to his clock, his second signal arrived at B. Since the signals travelled throughout at the speed of light, A is bound to conclude that

/his.....

his second signal reached B midway between the time he sent out the signal (T) and the time he received B's answering signal (K^2T), i.e. at time $\frac{1}{2}(K^2 + 1)T$. This is the crucial step in the argument. As Bondi (1967) says, "please realize that this is something complicated;" A has to assign a time to an event at which he is not present, and he does this by using Einstein's (1905) criterion for the simultaneity of spatially separated events.

From these results, it emerges that the ratio of A's non-proper time to B's proper time for the interval between the chosen pair of events is $\frac{K^2 + 1}{2K}$. To express this in terms of v, it is necessary to find K in terms of v. This relationship is easily shown to be $K = \left(\frac{c+v}{c-v}\right)^{\frac{1}{2}}$, where c is the speed of light (cf Rosser 1967), so that the above ratio of non-proper to proper time becomes $\left(1 - \frac{v^2}{c^2}\right)^{-\frac{1}{2}}$.

One would hesitate to describe this approach as "easy". The teacher would have to lead his pupils carefully through it, stressing particularly those points at which the postulates of relativity have to be invoked; but the approach is essentially straightforward. This may seem a longer way round to time dilatation than the approach mentioned in Section 11.6.3, but this one-dimensional approach does have the advantage of not depending on the result that lengths perpendicular to the relative velocity are unaltered.

The length contraction and velocity addition formulae can be derived with comparative ease by means of the K-calculus. Besides the above-outlined derivation of time dilatation, it is possible with K-calculus to show the "route-dependence" (Bondi's phrase) of time in other simple but striking ways.

The official K-calculus derivation of the length contraction formula makes use of the idea that radar signals are sent out from space ship A so as to arrive simultaneously (according to A) at the ends of the space ship B, where they are reflected back to A (cf Bondi 1965a, Rosser 1967).

/The.....

The length contraction result is also implicit in the K-calculus derivation of the time dilatation, since A and B disagree on the distance that separates them when B receives A's second signal; according to A, this distance is $\frac{1}{2}(K^2+1)vT$, while according to B it is KvT, and the ratio of the latter distance to the former is $(1 - \frac{v^2}{c^2})\frac{1}{2}$. The teacher might well prefer to derive this result in this shorter way. In that case it would be as well for him to appreciate that simultaneity plays a concealed but still fundamental role in this situation. The position of B in A's frame when it receives A's second signal can, in principle, be directly recorded on a sufficiently long measuring rod stationary relative to A, or by a space ship A' at rest relative to A and by happy coincidence exactly opposite B when the latter receives A's signal. On the other hand, the position of A in B's frame when B receives this signal can only be inferred, say from knowledge of the relative speed of A and B. Let L' be B's calculated value of A's distance from him when B receives A's signal, and suppose there is a space ship B' stationary relative to B and located at distance L' from B on (or, strictly speaking, near) the line joining A to B. Let B' have a clock that has been synchronized with B's clock. Then B''s clock must read KT when A is opposite B'; i.e. clocks in B and B' simultaneously read KT when B' is opposite A and B is opposite A'. On the other hand, the appearance of A opposite B' and A' opposite B will not be adjudged simultaneous in A's frame. As can easily be shown by using the results of the derivation of time dilatation, A's clock must read $\frac{2K^2T}{K^2+1}$ when A is opposite B'. (Non-proper time between events "A opposite B" and "A opposite B'" is KT; therefore proper time is $\frac{2K}{K^2+1} \times (KT) = \frac{2K^2T}{K^2+1}$. Likewise, if A' is equipped with a clock synchronized with A's, it must read

/$\frac{1}{2}$ .....

$\frac{1}{2}(K^2+1)T$ when A' is opposite B. Thus in each frame, the clocks in the other frame appear not to be synchronized; each "observer" can partly account for the discrepancy between his and the other "observer's" value for their separation in terms of "incorrect" synchronization of the other's clocks.

The analysis just given of the further implications of the problem of the receding space ships illustrates the subtleties which lie in wait for the student of relativity. The teacher will certainly have to master this kind of thing if he is to feel confident about teaching relativity to beginners. There must always be a distinct possibility that some member of a class will ask awkward questions which are fundamentally concerned with relativity of simultaneity, and it is thus vitally important for the teacher to know what he is talking about in this connection. It is the merit of the K-calculus approach to relativity that these subtleties are not initially apparent; in particular, the pupil does not have to familiarize himself from the start with the use of the conceptual device of sets of clocks which are synchronized in a particular frame of reference. A further merit of this approach is that it can, as has been done above, be presented in terms of space ships sending radar signals to each other, a picture which is not so far from reality as to strike pupils as ludicrous. True, the relative speeds with which these space ships would have to move to produce detectable relativistic effects are not those characteristic of present-day space vehicles, but these speeds are readily attained by charged particles in particle accelerators.

Reporting on the use of K-calculus in teaching special relativity to university students, Rosser (1967) states: "The author has found that students find this method of measuring the positions and times of distant events easier to follow than using spatially distributed synchronized clocks.

/It.....

It is to be hoped that more and more teachers will try the approach of [the K-calculus] ". Elsewhere (Rosser 1968b), he describes the K-calculus as "ideal" for the discussion of the kinematic aspects of special relativity. Here, then, is direct evidence in favour of the effectiveness of this approach to relativistic kinematics. If this topic is to be taught at school, it does seem as though the K-calculus approach ought to be tried in preference to more traditional methods.

11.6.5 Experimental evidence for relativistic kinematics

A school course in relativity should include some mention of the experimental evidence that confirms some of the startling predictions of relativistic kinematics. At the time of writing, no evidence is available concerning time dilatation experienced by conventional clocks or by laboratory atomic clocks, but this effect has been observed with what Sherwin (1961) calls "disappearing clocks", viz. decaying muons and pions. This evidence has been dubbed "inconclusive" by Prokhovnik (1967), but he considered no evidence of a later date than 1941. Of the experiments on muon decay, those by Farley et al (1968) and Frisch and Smith (1963) present data that appear to be much more reliable than those of Rossi and Hall (1941). An experiment on the decay of pions was carried out by Durbin et al (1952), who used pions created in the laboratory; the results confirm the expected dilatation of the pion mean life, but the interpretation of the results is complicated by the fact that a large correction has to be made for decay muons continuing in the direction of the parent pions.

The principle of the meson-decay experiments is that the number of fast-moving mesons which survive after having travelled a certain time (in the laboratory frame) is compared with the number that set out, and the effective "lifetime"

/of.....

of the particles in the laboratory frame is inferred from this information. In the experiments of Farley et al (1968), these measurements were carried out on muons injected into circular orbits in the muon storage ring of CERN, and it was found that the circulating muons had lifetimes of 12 times as long as those of stationary muons; this was within 2 per cent of the predicted factor of dilatation. Frisch and Smith (1963) compared the number of muons arriving per unit time at the top of Mount Washington with those arriving at sea level. Since the distance travelled by a muon before decay is equal to its time of survival multiplied by its speed, it is easily seen that these measurements likewise permit a deduction of a dilatation factor (8.8 ±0.8 in this case, again in agreement with the predicted value).

In explaining the background of these experiments to pupils, one can point out that unstable particles of a given kind do not all "live" equally long, but that one can forecast to a high degree of accuracy what fraction of the original number will have decayed after a certain time. The time for a collection of particles to have diminished to 37% of its original number is called its "mean life." If one knows the "mean life" from measurements already carried out, one can predict the number of particles that will survive after a certain time. In the experiment of Frisch and Smith, some 560 muons per hour arrived at their detector on Mt. Washington; of these, only 27 could be expected to arrive at sea level if one assumed that the "muon clocks" recorded the same time as the synchronized clocks on Mt. Washington and at sea level. Actually, some 408 muons arrived per hour at the detector at sea level, showing that the mean life of fast-moving muons is appreciably longer than that of stationary muons. The latter mean life is about 2.2 μs, which means

/that.....

that in the absence of time dilatation, a collection of muons travelling with nearly the speed of light would be reduced to 37% of their original number after travelling a distance of 660m, and to only about $(0.37)^3$ of the original number after travelling a distance equal to the height of Mount Washington (1910m). That the numbers surviving are far larger is a clear indication that according to the laboratory clocks, the moving muons have a mean life far longer than 2.2μs.

Naturally one would like to clinch the argument by showing that the measured value of the mean life is equal to the predicted value. This could not conveniently be done in a school course, because of the manner in which the mean life is inferred from the data and because of the details of the measurement of muon velocity. In the Frisch-Smith experiment, the velocity is inferred from measurements of the stopping distance of muons in iron; this gives the value of the energy, from which the velocity is calculated by application of the relativistic formula. In the experiment of Farley et al, the <u>momentum</u> of the muons is known, and the velocity is inferred from the appropriate relativistic formula. If relativistic kinematics is treated before relativistic dynamics, these formulae would not be at the disposal of the teacher. But it is not essential for the full quantitative story to be told to pupils, for whom it is chiefly important that they should gain the impression that "fast-moving" unstable particles live longer than slow-moving ones.

Which experiment or experiments should be described to pupils? The fact that we have devoted an appreciable amount of space to the Frisch-Smith experiment indicates our own bias towards it. This derives mainly from the homely simplicity of the basic arrangement. Also, the experiment

/forms.....

forms the basis of a film entitled "Time Dilation - An Experiment With μ-Mesons", and consequently photographs of the apparatus and examples of the measurements taken are available in textbooks (e.g. French 1968). The experiment of Farley et al is more direct than that of Frisch and Smith, since the numbers of muons surviving at different known times are measured, but the experimental arrangement is more sophisticated than that using natural muons. Moreover the muons travel in circular orbits, and although it is not difficult to prove that the time dilatation factor ought to be the same for particles travelling in such orbits as in straight-line paths, this proof would complicate the discussion and must be regarded as outside the scope of a school course.

The famous experiment of Ives and Stillwell (1938) on the longitudinal and transverse Doppler effect of light from fast-moving hydrogen atoms is commonly regarded as the first experimental test of time dilatation. The data obtained are beautifully consistent with special relativity, but the details of the analysis are too complex in the case of the longitudinal Doppler shift to be conveniently summarized in a brief account (cf Rosser 1967, French 1968). The "transverse Doppler shift" does not present such problems, since the frequency shift to which it gives rise is simply that resulting from time dilatation. However, the present writer has not seen any convenient summary of the relevant data in any book on relativity that might be accessible to teachers. There is clearly a case for making these data more widely known.

In any discussion of experimental evidence relating to time dilatation, it should be made clear that the time dilatation factor has to be allowed for in the design of certain experiments involving very fast-moving particles, and that this has never led to any anomalies. High-energy and particle

/physicists.....

physicists apply special relativity as a matter of course in designing their experiments and interpreting the results, and the fact that they do so successfully is itself a highly convincing proof of the validity of special relativity, including relativistic kinematics.

If the teacher wants to discuss evidence for the length contraction, he can refer again to the experiments on muon decay; from the point of view of the travelling muons, their lifetimes are normal, but the distance they travel is contracted by the appropriate factor.

11.7 Relativistic dynamics

11.7.1 Mass and momentum: thought experiments

One of the arguments advanced for teaching relativity at school was that it enabled one to show pupils where the mass-velocity and mass-energy equivalence formulae came from. We now consider whether one can find a simple way of deriving these results or making them plausible. The standard elementary methods of "deriving" the relativistic expression for the momentum of a particle are generally one or other version of that originally proposed by Lewis and Tolman (1909). This is the analysis of a perfectly elastic, symmetrical collision or "interaction" between two identical particles, subject to the requirement that the principle of conservation of momentum remains valid at relativistic speeds. (Full textbook accounts of this analysis are given by Sherwin (1961) and Arons (1965).) In simplified form, we may summarize it as follows:-

Consider a two-dimensional interaction between two identical particles A and B which, in one inertial frame (S', say), appears to be completely symmetrical; i.e. the x-components of the velocities of A and B are unchanged by the interaction, while the y-components are simply reversed.

/(As.....

(As a special case, the interaction may take the form of imparting equal and opposite y-components of velocity to A and B, while leaving their x-components of velocity unchanged.) It is possible to find another inertial frame (S, say) in which particle A has no x-component of velocity. In this frame, the effect of time dilatation is such that the y-velocities of A and B cannot be equal and opposite; in fact, the y-velocity of B must be less than that of A. If it is assumed that the principle of conservation of momentum can validly be applied to this interaction, and that the classical definition of momentum still applies, viz. momentum equals mass times velocity, then the inequality of the y-velocities of A and B as measured in S implies that the mass of B as measured in S must be bigger than the mass of A as measured in S. If the magnitudes of the y-velocities of A and B are small compared with that of the relative velocity $\vec{v}$ of S and S', it is easily deduced from the above argument that the mass of B measured in S must be equal to its "rest mass" multiplied by the time dilatation factor $(1 - \frac{v^2}{c^2})^{-\frac{1}{2}}$.

Certain points call for comment in connection with the role played by time dilatation in this argument:

(i) In frame S', the y-velocities of A and B are equal; they cannot be compared in S unless one compares the times taken by A and B to travel the same distance in S, or the distances travelled by A and B in the same time in S. Of these possibilities, the first is the obvious one to use, because of the fundamental result that distance measurements in different frames will agree provided that they are carried out in a direction perpendicular to the relative velocity of the frames. The dependence of the argument on this point will have to be explained if the derivation of the mass-velocity formula is to be understood (cf what was said on this matter in Section 11.6.3).

/(ii).....

(ii) If the pupil understands "time dilatation" to be concerned with the comparison between proper and non-proper time intervals between events, then it must be explained to him why it is that the time taken by A to travel a given y-distance in S can effectively be treated as a "proper time" in S, and why the obviously non-proper time taken by B to travel an equal y-distance in S must be related to the above proper time for A by the usual time-dilatation factor. These points are usually dealt with rather cursorily in elementary textbooks. Thus Rogers (1960) simply says"...E, watching E' at work, sees that E' uses a clock that runs slowly", and Bondi (1965a) similarly says only that "the time that in Brian's reckoning is only 60 minutes is 100 minutes in Alfred's reckoning". The time for A's journey will effectively be a proper time in S provided that the speed of A relative to S is small compared with that of light. This is always assumed in presentations of this argument; e.g. in Rogers's (1960) numerical example, the value of this speed is taken to be $3\ m\ s^{-1}$. To compute B's time as measured in S, one notes first of all that because of the assumed symmetry of the interaction, a clock travelling with B will <u>register</u> the same time interval as a clock travelling with A when A and B travel equal y-distances in S. (In other words, the proper times for their respective journeys are the same for A and B.) The time interval recorded by A's clock agrees with that measured in S, since the speed of A relative to S is small (see above). If B has an appreciable x-velocity relative to S, one must use the time dilatation formula to find the ratio between the time interval recorded on B's clock and the corresponding non-proper time interval measured in S. This tells us that in S, the time elapsed during B's journey is $(1 - \frac{v^2}{c^2})^{-\frac{1}{2}}$ times as great as the time-interval recorded on B's clock. Since the latter time is equal to that recorded

/by.....

by A's clock, which in turn is equal to that measured in S for A's journey, one finally obtains that in S, the ratio of the times taken by B and A respectively to travel equal y-distances is $(1 - \frac{v^2}{c^2})^{-\frac{1}{2}}$.

In Rogers's (1960) simplification of the above approach, one considers two observers riding on platforms moving parallel to each other and in opposite directions. Each observer has a platinum standard kilogram which rests on a frictionless table. While the platforms pass each other, the observers attach a long light spring to the two masses, along a direction perpendicular to their relative velocity, and allow the spring to give each mass a momentary inwards tug. Each then measures the transverse velocity of his standard kilogram; on comparing notes afterwards, these observers find that each has obtained the same value for the transverse velocity in his own frame of reference. Each now applies the principle of conservation of momentum, deducing from it that the other's standard kilogram must have acquired an equal and opposite momentum to his own. From here onwards, the argument follows the same lines as in that given above.

This treatment is shorter and easier to follow than the "standard" one, while making all the essential points of the latter. The deliberately-contrived symmetry of the situation may still puzzle beginners, who may find this kind of artificiality (perfectly permissible in a "thought experiment") and the unreal nature of the whole arrangement somewhat off-putting. Bondi (1965a), in his version of the "interaction" approach, deals with an even simpler and more realistic situation than Rogers does, viz. a bullet fired at a sheet of armour-plating. The latter is at rest in one of the reference-frames and is so placed that in the other frame it will be seen edge-on. In the rest-frame of the sheet, the bullet

/is.....

is fired at right angles to it. In each frame, it is agreed that the maximum thickness of armour plating penetrated by the bullet is a measure of its (transverse) momentum. The same value is obtained in each frame for the depth of penetration of the bullet, implying that the transverse momentum in each frame is the same. From this point on, the argument follows similar lines to the others outlined above.

Of these simplifications, it is probably Bondi's which beginners would find easier to follow and accept, for reasons implied above. Bondi (1966) himself remarks that he has not seen a "simple dynamical experiment" which serves its purpose in this connection as well as his example of the penetrating power of bullets, but he feels "that it should be possible to refine and improve the procedure, and get a simpler and more convincing dynamical experiment to show the increase of mass with velocity." No one seems yet to have found one.

In following Bondi's approach, the teacher is still faced with the need of having to explain why lengths transverse to the direction of relative motion are the same in both frames (assuming that he has not already had to do this, which would be the case if he has introduced his pupils to kinematics via the K-calculus); and if he is conscientious, he will want to take particular care in explaining how time dilatation enters into the argument. We have indicated how these problems can be solved in principle, but their existence means that even Bondi's approach to "relativistic mass" is not altogether straightforward.

11.7.2 Mass and momentum: empirical approach

It has been shown that even the simplified forms of the "interaction" approach to relativistic mass have difficulties lying just beneath the surface. It is not surprising to find Bondi (1966) reporting that "the introduction of mass seems to give rise to some difficulties of presentation."

/We.....

We can be guided towards a suitable alternative approach by noting that the methods of presentation of special relativity outlined so far have been essentially deductive. Such experimental evidence as is adduced in this connection is drawn from books or, at best, from films, and is cited to support the predictions and not to point the way towards them. Is it not possible to present certain aspects of special relativity, such as dynamics, by way of discussion of appropriate experiments? In this way one might be able to by-pass the awkward questions that arise in connection with a deductive presentation.

Rosser (1968b) is firmly in favour of the "empirical" approach: "In the true tradition of physics teaching," he writes, "the variation of mass with velocity can be presented as established by experiment." He actually advocates <u>starting</u> one's teaching of special relativity with dynamics. We shall consider the latter question below (Section 11.8); for the moment, we consider the possibility of presenting the variation of mass with velocity as a result established by experiment. How should such an experiment be introduced? Suppose one decides to describe Bucherer's (1909) experiment, as Rosser (1967) has specifically suggested; how does one lead up to it? Simply to describe the experiment, report its results and show that they are consistent with the relevant formula would surely not be in "the true tradition of school physics teaching." When Bucherer performed his experiment on the ratio of charge to mass for beta-rays, he did so against a background not only of Einstein's newly-developed special theory of relativity but also of Lorentz's earlier theory of the electron, which predicted the same mass-velocity relationship for electrons as was later to be predicted by special relativity for particles in general; also, earlier experiments by Kaufmann had suggested that the mass of an

/electron.....

electron depended on its velocity. Thus Bucherer's experiment was not done in a vacuum (in a metaphorical sense, at any rate); he had some idea of what to expect and two theories to help him interpret the results. Since this background is not expected to be known by the pupils to whom the experiment is to be described, one must try to think of some other way of placing the experiment in a context. One might tell pupils that since it has been found that deviations from classical kinematics become manifest at speeds comparable to that of light, it might be reasonable to expect corresponding deviations from classical dynamics at such speeds. But even then it will be difficult to justify why one should specifically set out to investigate the variation of electron mass with velocity when, unless one has already sought to apply special relativity to dynamics, one has no reason for picking on mass as something likely to vary significantly with speed at very high speeds.

It would be better if a more general experiment on the behaviour of electrons (say) at high speeds could be discussed. Such an experiment is available, viz. Bertozzi's (1964) on the limiting speed attained by electrons accelerated to high energies. But an approach to relativistic dynamics through this experiment is better discussed under the heading of energy (see below, Section 11.7.3). Meanwhile, we must stress that to justify the formula $m = \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}$ simply by describing Bucherer's experiment is scarcely any improvement on not doing any relativity at all.

11.7.3 Energy

If a deductive approach has been followed in teaching the mass-velocity relationship, the latter result can be taken as the point of departure for the deduction of the mass-energy relation. At university level that deduction would be carried

/out.....

out by finding the work done on a particle by an accelerating force, using the relativistic form of Newton's Second Law to relate force and mass. This shows that the kinetic energy gained by the particle is equal to $mc^2 - m_0c^2$, where m and m_0 are respectively the relativistic and rest masses of the system, and this result can be interpreted by saying that the extra mass gained by the particle is just the gain in kinetic energy divided by c^2, and that m_0c^2 is the "rest energy" of the particle. Much of the mathematics of this deduction is beyond the reach of South African pupils, but it is possible to short-circuit the detailed argument by using the familiar trick of taking the expression for relativistic mass and showing that for low speeds, it reduces to $(\frac{1}{2}m_0v^2+m_0c^2)$ divided by c^2. This should properly be done by using the binomial theorem to expand $(1 - \frac{v^2}{c^2})^{-\frac{1}{2}}$, but this theorem is not officially in the mathematics syllabus of South African schools. This is not an insuperable obstacle, as pupils can be shown with the aid of arithmetical examples that $(1 - x)^{-\frac{1}{2}}$ equals $1 + \frac{x}{2}$ when x is very much smaller than unity.

The most interesting aspect of mass-energy equivalence (some - e.g. Taylor and Wheeler (1966) - would say, the really fundamental aspect) is the prediction that energy is associated with the rest mass of a particle or system, and in particular that if the rest mass of the system were to change, there must be a corresponding gain or loss of internal energy. This could be treated as a result to be verified by experiment - ample experimental evidence for it is available (cf Section 10.3.10) - but it is also possible to give a more direct hint of it from theory, by showing that the total rest mass of a two-particle system is larger after a perfectly inelastic collision between the particles than it is before the collision, the increase in rest mass being equal to the loss in

/kinetic.....

kinetic energy of the particles divided by c^2. Arons (1965) gives a treatment of an inelastic collision between identical particles which is symmetrical in one frame of reference. This works out straightforwardly enough in the low-velocity approximation, but the analysis involves viewing the collision from different frames of reference and taking the limit as the relative velocity of the frames tends to zero, which must be regarded as too complex for an average schoolboy to follow. Pupils could be given a qualitative account of inelastic collisions from a relativistic standpoint, with the point being put to them that if no energy escapes from the system as heat, the added internal energy resulting from the loss of kinetic energy is accompanied by an equivalent increase in rest mass.

Instead of following the above essentially deductive approach to mass-energy equivalence, it is possible to follow a quasi-experimental, inductive approach, based on the above-mentioned experiment of Bertozzi (1964). An approach of this type is employed by the P.S.S.C. in their "College Physics" text (P.S.S.C. 1968), and is well worth considering for use in school. Because of the importance of Bertozzi's experiment in this connection, we discuss it in some detail first. The experiment, which demonstrates the existence of a limiting speed for electrons, was planned as the basis of one of a series of films on special relativity; the film itself is called "The Ultimate Speed, An Exploration With High-Energy Electrons." As this experiment seems to have been the first directly to show the existence of a limiting speed for particles, one may remark that in this case purely educational needs have given rise to the performance of an experiment of fundamental significance. The principle is very simple. Electrons are accelerated in a Van de Graaff machine and then emerge into a field-free space in which their time of flight over a known distance is measured. The kinetic energy of

/the.....

the emerging electrons is also measured, so that the relationship between the kinetic energy of the electrons and their speed is determined directly in this experiment. As the time of flight of the electrons is measured by measuring the distance between "blips" on an oscilloscope trace, and the kinetic energy is measured by a calorimetric method the principle of which is easy to understand, pupils should have no difficulty in following a simplified account of Bertozzi's experiment. We say "simplified" because there are certain details of the experiment with which one need not trouble the beginner. First, there are some technical subtleties, such as the method of calibration of the oscilloscope sweep and the details of the method used for finding the number of electrons responsible for giving up a known total amount of kinetic energy. Second, to obtain electron kinetic energies in excess of 1.5 MeV Bertozzi had to use an electron linear accelerator to give additional kinetic energy to the electrons from the Van de Graaff machine; these additionally-accelerated electrons gained kinetic energy over part or all of the path over which they were timed. As Bertozzi's measurements showed that these more energetic electrons had much the same average speed (viz. the speed of light) over this path as electrons of kinetic energy 1.5 MeV, it is justifiable in an elementary account to omit all mention of the "Linac" and its function.

We have implied above (Section 11.7.2) that it might be pedagogically advantageous to use Bertozzi's experiment in teaching relativistic dynamics. One advantage is that we can quite easily justify the performance of an experiment of this type as a logical further step in the exploration of special relativity. (Here we follow what is essentially the P.S.S.C. line of reasoning.) In relativistic kinematics, the factor $\left(1 - \frac{v^2}{c^2}\right)^{-\frac{1}{2}}$ emerges. This quantity becomes imaginary as v exceeds c, implying that there cannot be a "relativistic kinematics" for speeds greater than c. A possible

/implication.....

implication of this is that particle speeds cannot exceed c in any inertial frame. Newtonian dynamics imposes no upper limit on the speed of a particle - in principle, it can be accelerated to any speed one pleases by the application of a force for the appropriate length of time. Thus the implications of relativistic kinematics are in conflict with the predictions of Newtonian dynamics, and there is the possibility that at high speeds Newtonian dynamics might cease to be valid. One might investigate this by accelerating particles to very high speeds to see whether or not their behaviour is in accord with Newtonian expectation. Bertozzi's experiment does just this; a known electric force is applied to electrons over a fixed distance, and the kinetic energy and speed acquired by the electrons are measured. The results can be compared with the Newtonian relation between kinetic energy and speed.

The experiment plainly shows that electrons do attain a limiting speed as their kinetic energy is increased. It also reveals a relationship between kinetic energy and speed which is at variance with Newtonian dynamics at very high speeds. Now a formula must be found to fit Bertozzi's results, since this might give a clue as to how Newtonian mechanics should be modified in relativistic situations. The vital factor in the correct formula is $\left(1 - \frac{v^2}{c^2}\right)^{-\frac{1}{2}} - 1$, and the pupil must somehow be led to this in a manner which is an acceptable compromise between simply telling him it is so and giving him a formal derivation. The guiding principles to be laid down are that the new formula for kinetic energy must reduce to the classical one at low speeds, and that the value of the kinetic energy must become very large as the speed of the particle approaches "the ultimate speed." As a first possibility, one can suggest the factor $\left(1-\frac{v^2}{c^2}\right)^{-\frac{1}{2}}$

/as....

as a quantity which does become very large as v tends towards c. It can be shown that for low speeds, multiplication of this factor by mc^2, where m is the mass of the particle, yields an expression which is equal to $mc^2 + \frac{1}{2}mv^2$. (The introduction of "mc^2" here need not seem too artificial; if $(1 - \frac{v^2}{c^2})^{-\frac{1}{2}}$ is first expanded in its low velocity approximation, "mc^2" is seen to be the term by which to multiply it in order to obtain "$\frac{1}{2}mv^2$" in the resulting expression.) This expression looks promising, but contains an unwanted "mc^2"; to remove this, it is necessary to assume that mc^2 must be multiplied by $(1 - \frac{v^2}{c^2})^{-\frac{1}{2}} - 1$. To show that the resulting formula fits Bertozzi's results very closely is a matter of not very troublesome arithmetic; this is an important step which should on no account be omitted.

The pupil would now know, as a matter of experimental fact, that the kinetic energy of a very fast-moving particle is given by the expression

$$E_k = \frac{mc^2}{\sqrt{1-\frac{v^2}{c^2}}} - mc^2$$

The interpretation of the two terms in this expression could be given to pupils along the lines sketched out earlier in this section, or left for later expansion in connection with nuclear physics.

11.7.4 Relativistic momentum and Bertozzi's experiment

In the P.S.S.C. (1968) treatment, the above analysis is extended to arrive at the relativistic expression for the momentum of the particle. As a fundamental assumption, the classical equality between force and rate of change of momentum is retained. The time for which the accelerating electric force F acts on the particle is very nearly $\frac{d}{c}$, where d is the distance over which the force acts on the particle. The momentum acquired is therefore $\frac{Fd}{c}$, which in turn is equal to $\frac{E_k}{c}$. But for such faster-moving particles E_k

/is.....

is very nearly equal to E, the total energy (a little arithmetic readily shows that the rest energy is small compared with E_k if the electron has an energy of several MeV), so that the momentum can be set equal to $\frac{E}{c}$ very nearly, i.e. to $\frac{mc}{\sqrt{1-\frac{v^2}{c^2}}}$. On the other hand, for low velocities the expression must reduce to mv. The expression $\frac{mv}{\sqrt{1-\frac{v^2}{c^2}}}$ is seen to satisfy both conditions.

An important advantage of approaching relativistic dynamics in this way is that it is the expression for momentum as a whole that is under discussion. One does not start off by assuming that momentum must be thought of as "mass times velocity"; instead, an expression for momentum is found which is consistent with the classical relation between force and rate of change of momentum and also with the experimental requirements imposed by the results of Bertozzi's experiment. Once the correct expression has been arrived at, it can be interpreted as "relativistic mass times velocity", as "mass times proper velocity", or simply as "momentum". (In the P.S.S.C. treatment, "relativistic mass" is relegated to a brief footnote.)

11.7.5 Relativistic momentum: experimental confirmation

Direct evidence for the correctness of the relativistic expression for momentum is provided by the results of experiments of the type initiated by Kaufmann and Bucherer. In these experiments, fast-moving charged particles (usually electrons) are projected at right angles to a uniform magnetic field, and the radius of curvature of the orbit of the particle is measured. The equation relating the momentum of the particle to the magnetic field strength (induction) B, the charge on the particle q and the orbital radius R is very simple, viz. $p = Bq\,R$. This result follows easily from non-

/relativistic.....

relativistic mechanics, and pupils should already be familiar with it. What is of interest here is that it is also valid for particles moving with relativistic speeds. Elementary- and intermediate- level textbooks on general or modern physics tend to be somewhat casual in their treatment of the latter fact. Some (e.g. Bitter 1956, Semat 1962, Arons 1965, Halliday and Resnick 1966) make use of it in discussing Bucherer's experiment, but do not attempt to justify this usage. At least one (Blanchard et al 1969) actually says, "It is by no means obvious, but this same formula continues to hold even for speeds near the speed of light if we mean by p the relativistic momentum", but does not explain further. One has to turn to textbooks on relativity to see this point adequately discussed.

This is surprising, because although the result may be "by no means obvious", it is far from difficult to derive. The main fact on which the result rests is that the speed and the magnitude of momentum of the particle remain the same, since the force is always at right angles to the velocity of the particle. This is obvious in classical dynamics, but cannot be taken for granted in relativistic dynamics, in which the acceleration of the particle is not necessarily in the direction of the resultant force on it. From the fact that the force must be perpendicular to the velocity, it follows that the particle does not gain kinetic energy; from the relativistic expression for kinetic energy it follows that the speed of the particle must stay the same, which in turn means that the magnitude of its momentum must stay the same. Once these points are established, one can proceed in a variety of ways to the derivation of $p = Bq\,R$. By considering the changes in direction of the velocity and momentum

/vectors.....

vectors as the radius vector of the particle describes a small angle, it is easy to show that the centripetal force on the particle is given by $\frac{pv}{R}$. Or one may make use of the fact that the "relativistic mass" of the particle is constant in this situation, so that the rate of change of momentum of the particle is simply equal to its "relativistic mass" multiplied by its centripetal acceleration. In either case the desired result follows directly.

Since in experiments of the Bucherer type an independent measurement of the speed of the particle is made, the results of these experiments make it possible to deduce the relationship between p and v or, if one prefers to put it that way, to find $\frac{p}{v}$ (the "relativistic mass") and compare it with $\frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}$. This is subject to the assumption that the charge on the particle is independent of its velocity. One might feel that this seems to be implied by certain aspects of Bertozzi's experiment, in which the measured values of the electron kinetic energy agree with those calculated by multiplying the accelerating potential difference by the charge on the electron. Either the charge stays the same, or the force varies with speed in such a way as to compensate for a corresponding variation of electric charge with speed. Bertozzi actually takes the charge on the electron to be constant and regards this part of the experiment as establishing that the electric force on the electron is independent of the speed of the latter. We must therefore seek independent evidence for what Rosser (1967) calls the "principle of constant charge". This is provided by experiments such as those of J.G. King (1960), who showed that the charges on electrons and protons in hydrogen molecules have the same magnitude to within 1 part in 10^{20}. As the velocities of electrons in hydrogen molecules or atoms are expected to be vastly different from those

/of.....

of protons, we may take it that the electrical neutrality of the hydrogen molecule demonstrates the validity of the "principle of constant charge." King also performed a similar experiment on helium, with a similar result. Another method of showing the charge neutrality of atoms and molecules and so confirming constancy of charge is that of Zorn, Chamberlain and Hughes, who have measured the deflection of neutral molecular beams passing through an electric field perpendicular to the direction of motion, and found it to be zero (cf Hughes 1964).

We see that although the application of the formula $p = Bq R$ to relativistic particles orbiting in a magnetic field does need justification that goes a little beyond the obvious, this justification is not difficult to provide. But the very fact that one has to dig a little deeper into the theory at this point may deter the syllabus planner from including discussion of the above-mentioned experiments in his exposition of relativity. (It is perhaps significant that the P.S.S.C. treatment does not mention them.) If it is thought worthwhile to discuss these experiments, then it would be well to mention other, more recent experiments than Bucherer's, such as those of Zahn and Spees (1938) and M.M. Rogers et al (1940).

11.7.6 Photon mechanics

It is customary to extend the treatment of relativistic dynamics to include a brief discussion of the mechanics of photons. The P.S.S.C. method of doing this is quick and easy: comparison of the expressions for momentum and total energy of a particle shows that $p = \frac{Ev}{c^2}$, from which it follows that if v equals c, as in the case of photons, $p = \frac{E}{c}$. Substitution of $v = c$ into the originally-obtained expression for momentum leads to an embarrassing infinity on account of

/the.....

the vanishing of the denominator, to be avoided only by setting the rest mass of the photon equal to zero. That the photon has zero rest mass and has momentum $\frac{E}{c}$ are the prime facts that special relativity teaches us about the photon; in connection with South African syllabuses, they would be best derived and discussed when the "photon" aspects of light are being treated. In such a context, the Compton effect, the annihilation of positron-electron pairs to produce pairs of gamma-ray photons and the converse process of "pair production" can be cited as examples of the successful application of the ideas of relativistic dynamics to photons.

This view of photon mechanics is that which is to be found in virtually any intermediate-level textbook on special relativity, to say nothing of more general text books which treat special relativity in passing. It has recently been sharply criticized by Bunge (1967), according to whom "special relativity mechanics" refers only to material systems: entities such as photons and neutrinos "are not supposed to 'obey' dynamics but quite different theories." Consequently the concept of mass (rest or relativistic) can apply only to material particles or bodies; we are not justified in equating the energy of a photon divided by c^2 to the relativistic mass of a photon, because "the mass concept is as out of place in the theories of the radiation field as the concept of grandfather." Now it is true that the behaviour of photons is properly described by quantum electrodynamics, according to which the energy and momentum of the electromagnetic field are respectively quantized in units $h\nu$ and $\frac{h\nu}{c}$. But the latter prediction coincides with that which was obtained above by extrapolating relativistic dynamics appropriately, and is moreover confirmed by such experiments as the Compton scattering of photons. Indeed, the results of these experiments can be satisfactorily interpreted within the framework of

/relativistic.....

relativistic dynamics on the assumptions that the photon is a "particle" of zero rest mass, energy $h\nu$ and momentum $\frac{h\nu}{c}$. May we not be dealing here with an acceptable overlap between theories?

One could answer this question in the affirmative if one could somehow show experimentally that the concept of mass was directly relevant to the photon, e.g. that a photon behaved as if it had a relativistic mass $\frac{h\nu}{c^2}$. The possibility of its behaving in this way seems to be explicitly denied by Bunge. "If the field in a hollow cavity increases in energy by ΔE", he writes, "the field in it gains no mass although the walls do lose mass by the amount $\Delta m = \frac{\Delta E}{c^2}$". In other words, he seems to be denying the possibility that the energy of the electromagnetic field in the cavity can contribute to the rest mass of the system cavity-plus-field. This denial has disturbing implications; for example, it implies that if a pulse of radiation were to be emitted by one wall of the cavity and absorbed by the other, the gravitational attraction of another body for the box enclosing the cavity would be less while the pulse was in transit than before its emission or after its absorption. This particular question cannot be settled directly by experiment, but it has been claimed (e.g. by the P.S.S.C. (1968)) that the recent experiments of Pound and Rebka (1960) and Pound and Snider (1965) on the gravitational red-shift of photons do provide supporting evidence for the idea that the photon has a relativistic mass $\frac{h\nu}{c^2}$.

In discussing this point, it is important to distinguish carefully between what has been observed in these experiments and how one interprets the observations. What has been observed is that if a stationary atom at the bottom of a tower emits a photon, this photon will not be absorbed by a similar stationary atom at the top of the tower. If the source of

/radiation.....

radiation at the bottom of the tower is moved upwards at just the right speed, however, the photons arriving at the top will in fact be absorbed.

It can immediately be inferred from these observations that the frequency of the radiation emitted by a stationary atom at the bottom of the tower is too <u>low</u> when it arrives at the top to cause an absorptive transition between the corresponding energy states of a similar atom at the top of the tower. This fact, as well as the amount of the discrepancy in frequency, can be explained within the framework of relativity theory in various ways. One such explanation, given by the P.S.S.C. (1968) and others, is based on the premise that a photon has a relativistic mass $\frac{h\nu}{c^2}$. As the photon rises a height d in the earth's gravitational field, it gains potential energy of amount equal to the above mass multiplied by gd. (The variation in energy of the photon is assumed to be very small compared with h in order to ensure that the "mass" remains effectively constant while the photon rises.) This increase in potential energy can only be gained at the expense of the kinetic energy of the photon. Therefore at a height d above its source, the photon's kinetic energy must have been reduced by $(\frac{h\nu}{c^2})$ gd, and the frequency of the radiation must correspondingly be less by $(\frac{\nu}{c^2})$gd than its value just after emission. This is exactly the amount of the redshift observed by Pound and his co-workers.

The P.S.S.C. (1968) sum up by saying that the results of these experiments "confirmed to within 1 per cent that photons indeed behave as though they have a mass of $m = \frac{E}{c^2}$". But as Bondi (1965b) in particular has insisted, this is not the only way in which one can interpret these results. Bondi is disturbed by the idea that the radiation actually loses frequency on the way up: "We cannot, after all, envisage the electromagnetic waves making up the radiation actually to

/lose.....

lose cycles on the way." He prefers to take the view that "In a static gravitational field one cannot see from the top a clock at the bottom losing more and more time unless it actually does so." For him, "the only tenable interpretation" of the gravitational red shift is that if one has two clocks of identical construction occupying different positions in a gravitational field, the clock in the position of lower potential energy will actually go slower than the other. This interpretation is also consistent with the experimental facts.

Frisch (1962) has also discussed this point, but takes a less rigid view than Bondi. In an operational approach to the problem, he suggests that whether or not one chooses to regard the radiation as losing frequency depends on how one decides to establish one's standards of time in places of different gravitational potential. If one does this by adopting a single master clock for one's fundamental standard and using "light" signals to synchronize all other "stationary" clocks with it, then the radiation emitted by an atom at the bottom of a tower is unaltered in frequency when it arrives at the top, by definition. But if independent clocks of identical construction are used to define frequency at the places where they are kept, then the clocks at the top and the bottom of the tower go at the same rate, by definition, and the frequency of the radiation changes in transit between the two.

Which of these views one prefers to adopt is clearly a matter of taste. (The present writer is strongly inclined towards Bondi in this matter.) This being the case, one can argue that the P.S.S.C. go too far in claiming that the results of the gravitational red-shift experiments "confirm" that photons behave as though they have a relativistic mass.

The bending of light in a gravitational field is often interpreted in simple accounts (e.g. Bergmann 1968) in terms of photons having a gravitational mass. However, naive

/application.....

application of this idea leads to the wrong result; as Weber (1964) points out, the photon actually behaves in this case as though its gravitational mass is twice its inertial mass. Within the framework of general relativity, this prediction of photon behaviour is obtained not by making any assumptions concerning the supposed mass of the photon, but by making appropriate use of the fact that for a light signal, the "interval" ds always vanishes.

It appears that experimental work on photon behaviour cannot as yet provide any firm answer to the question of whether or not the photon has some kind of a mass. It is therefore advisable to proceed cautiously in this matter, especially in a school course. Perhaps the line to take is that <u>if</u> one had "particles" capable of travelling with the speed of light, special relativity predicts that such particles must have momentum $\frac{E}{c}$ and rest mass zero. (Even this is not an unexceptionable way of putting it, at any rate as far as photons or neutrinos are concerned; there is no inertial frame in which these particles can appear at rest, so that the phrase "rest mass" is not really applicable to them. It is safest to refer to these entities as having "zero mass", but for this to be properly understood the concept of mass must previously have been treated as a velocity-independent property.) The experimental evidence concerning the interaction of photons with matter can be cited in support of the possession by the photon of momentum $\frac{E}{c}$, and it can be stressed that this is consistent with these predictions.

11.7.7 <u>Which approach to relativistic dynamics should one use?</u>

Bondi's approach to this topic starts off with an attractively simple situation (bullet penetrating armourplate), the analysis of which depends on certain facts which may not be easy to explain to a pupil. The "quasi-experimental" approach described above starts from Bertozzi's experiment,

/which.....

which is simple in principle, and analyses its results in a way which is a good deal better than just telling pupils the answer, although it is based on wisdom after the event. The factor $(1-\frac{v^2}{c^2})^{-\frac{1}{2}}$ comes into the analysis as a hint offered by relativistic kinematics; no awkward points concerning length and time measurement have to be considered. No doubt it is a disadvantage of this approach that it does not properly show how the "essentially kinematic nature of relativity" (Brehme 1968) determines the nature of the relativistic expressions for momentum and energy. But the pupil would know from studying the results of actual experiments that for fast-moving particles, classical dynamics must be modified, and from those results he can (with assistance) deduce the nature of the modification. That certain subtleties are bypassed in this approach is surely acceptable in a "first-order" approach to relativity.

We are inclined to favour the "quasi-experimental" approach, on the grounds that it appears to raise fewer problems of exposition than "interaction" approaches, of which Bondi's is an example. Ultimately this question should be settled by classroom trial.

11.8 Special relativity via mechanics?

We have noted above (Section 11.7.2) Rosser's (1968b) suggestion that special relativity should be taught via mechanics. It seems to us that this is open to objection. Although special relativity derives very strong support from the results of dynamical experiments, one cannot easily find a convincing way of introducing these results at the beginning of a discussion of special relativity. (This applies as much to Bertozzi's experiments as to those of the Bucherer type.)

We can, however, agree with Rosser that there may be pedagogical advantages in initially replacing the "principle of the constancy of the velocity of light" with what he calls

/"the.....

"the principle of the constancy of limiting speeds of particles" (Rosser 1967, 1968b). As enunciated by him, this principle states that "the numerical value of the limiting speed of particles is the same for all inertial reference frames and is equal to 3.0×10^8 ms^{-1}, the speed of light in empty space". As Rosser says, it is a reasonable extension of the principle of relativity to postulate that the result of Bertozzi's experiment should hold in all inertial frames. Pupils might find it easier to understand and accept this principle than the principle of the constancy of the velocity of light. If, as Rosser implies, pupils are likely to wonder about a medium for the transmission of light, they might tend to confuse its behaviour with sound, which would certainly hinder their acceptance of Einstein's second postulate. In Section 11.4.2 we have suggested ways of overcoming this difficulty, but it might be convenient to sidestep it by using the "constancy of limiting speed" postulate.

Rosser suggests that in presenting the K-calculus one might have the space ships firing bursts of accelerated electrons at each other from ray-guns instead of communicating with each other by radar signals. Were one to introduce relativistic kinematics in this way it would mean preceding this introduction with a description of Bertozzi's experiment. It would not be necessary to infer from Bertozzi's results what the relativistic expression for kinetic energy must be; only the fact that there is a limiting speed is relevant to what would follow. Bertozzi's experiment could be introduced for this purpose by discussing with pupils whether one ought to take for granted that particles would be accelerated to infinitely high speeds, and whether it would not be a good idea to examine some relevant experimental evidence.

In our view, approaching the K-calculus by way of space fiction electron guns and Bertozzi's experiment is a

/pedagogically.....

pedagogically acceptable way of introducing "special relativity via mechanics". If the introduction of special relativity into the school syllabus is to be attempted, this is certainly one of the approaches to the subject that should be tested.

11.9 Conclusions

(i) The above survey suggests that it might be possible, in principle, to teach elementary special relativity to matriculation pupils. There seems to be at least an even chance of such pupils being able to achieve a "first-order" understanding of the subject.

(ii) Special relativity has great potential for stimulating the imagination of pupils, and this is one of the strongest arguments for trying to teach it at school. If it is to be taught there, something must make way for it. The present author has had experience of trying to teach a class of third-year university students a somewhat more advanced version of the "syllabus" considered above. He was allotted six forty-five minute periods for the purpose, and found that the material could be covered - but only just. It would be wise to allot at least double this amount of teaching time to a school course in relativity. This should not occupy too many weeks of the teaching year, as pupils cannot do any laboratory work which will illuminate the subject for them. Still, this is time that would otherwise have been given to something else, so that some decision will have to be made concerning what is to be left out.

(iii) It would be unwise to present special relativity to the pupil without preparing him in some way for acceptance of some of the predictions of the theory. In particular, classical mechanics should not be taught to him as though it were the last word on the subject. As he studies it, classical

/mechanics.....

mechanics deals only with slow-moving particles; as far as laboratory experiments are concerned, with very slow-moving particles indeed. He should not therefore be led to assume that classical mechanics must apply without modification to fast-moving particles. Above all, he should not assume that Newtonian assumptions about time which work so well in ordinary experience must apply to all experience (cf Section 3.6). The pupil could also be prepared for relativity by a tentative introduction of the Principle of Relativity in the context of Newtonian dynamics.

(iv) Contrary to the view expressed by Rosser (1968b), it seems best for the pupil to start his formal study of special relativity with relativistic kinematics. It has been suggested that at school level it is not desirable to start relativistic kinematics with a derivation of the Lorentz transformation. Two possible approaches were considered: of these, it was thought that Bondi's "K-calculus" for collinearly moving observers was better suited to elementary teaching.

(v) It is important to include some form of derivation or justification of the key predictions of relativistic dynamics: the so-called "variation of mass with velocity" and "mass-energy equivalence". Two elementary approaches were considered: simplified versions of the "collision" or "interaction" approach, devised by Rogers (1960) and Bondi (1965a) respectively, and a "quasi-experimental" approach, used by the P.S.S.C. (1968). In view of certain snags inherent in the former approach, it was thought that the P.S.S.C. approach might prove easier to use.

(vi) One should not expect that a very profound understanding of special relativity would be attained by pupils who have been taught the kind of course considered here. One should especially not expect them to grapple too successfully with problems in special relativity which deal with certain paradoxical situations. Problems given to pupils should be kept

/simple.....

simple and comparatively free of subtlety.

(vii) Although we have concluded that certain methods of approach to special relativity are more likely to succeed than others, we have done so without experience of teaching the subject at school. Therefore we feel that while the above discussion may be helpful in pinpointing some of the difficulties that the prospective teacher of special relativity will face, its conclusions concerning teaching approaches must be submitted to experimental test. We accept Bekveld's (1965) summing-up: "There is no tradition on which we can draw for the teaching of relativity at the level of secondary education. It seems advisable to start pilot courses in all countries, trying perhaps several approaches to the teaching of this topic. After these teaching experiments we might find one or a few sure methods which seem to be appropriate and effective for our purposes."

(viii) Both special relativity and quantum theory offer challenges to "common sense." This is an argument in favour of trying to teach these topics at school, since a picture of science which suggests that science accords only with everyday common sense is certainly a false picture. No doubt we need more evidence than is at present available on the extent to which the matriculation pupil can reconcile himself to "challenges to common sense." There is much to be said for gaining some of this evidence from actual classroom experience.

(ix) Relativity may well offer a bigger challenge to pupils than quantum physics. In teaching the latter, one can start with certain experimental facts and derive certain conclusions from them which conflict with conventional views of light and matter, and then show how all is reconciled within the framework of the quantum theory. The pupil may find this puzzling, but at least he is not required to change his attitude to <u>familiar</u> things; water waves remain water waves, and billiard

/balls.....

balls remain billiard balls. On the other hand, relativistic kinematics leads quickly to a challenge to the accepted view of time and space, and to the fact that relativistic effects are "reciprocal". Thus in learning special relativity the pupil may have to work even harder to adjust himself to new views than he has to in learning quantum physics, with a correspondingly harder effort from the teacher.

(x) Some years ago, Boulind (1957) observed that "few teachers know about relativity." One suspects that in Boulind's own country (Great Britain) the picture may have changed a little in view of the tendency to teach more relativity at university, but in Britain and in many other countries, including South Africa, there must be many practising teachers who have never had the opportunity properly to come to grips with relativity. Such teachers ought not to attempt the teaching of relativity without working carefully through some kind of preparatory course, for which such books as those of Bondi (1965a), Smith (1965), Taylor and Wheeler (1966), Resnick (1968) and certain others should be required reading. A course of this kind should include plenty of problem-solving and opportunities for discussion, in order to give the teacher the confidence he needs to teach relativity adequately. Until there are enough teachers who have had this kind of preparation, relativity cannot be successfully taught in our schools. The problem is not insoluble: if the step can be taken of including certain aspects of quantum physics in the school curriculum, we can certainly consider doing the same with relativity, although, as we have suggested above, relativity probably presents a bigger challenge to teacher and pupil than elementary quantum physics. One way of clearing the ground for the acceptance of special relativity as a topic in school physics would be to encourage discussion of it among teachers by means of suitably-aimed articles in journals, and

/to.....

to suggest to teachers that they urge their brighter pupils to read about it. If the latter suggestion were followed, teachers might find themselves having to take special relativity seriously in order to be able to discuss it properly with interested pupils, and would then be in a better position to teach special relativity when the need arose.

(In connection with points (iv) and (v) above, it should be mentioned that Rekveld (1969) has recently published "a brief survey.... of four new approaches to the teaching of special relativity", in which he briefly describes four "lines of approach" to the subject; viz. the K-calculus, "the invariance of the space-time interval", the "dynamical approach" (i.e. Rosser's (1968b) suggestion to introduce relativity by means of experimental results that support relativistic mechanics), and "thought experiments" (i.e. the derivation of the "fundamental effects of relativity" by the exclusive use of thought experiments). Of these, we have not troubled to consider the second in this chapter, believing that its reliance on space-time diagrams of one kind or another (e.g. "Minkowski", "Brehme" or "Loedel") would place it beyond the comprehension of most South African school pupils. In our opinion, Rekveld's discussion of the other three approaches is valuable chiefly because it draws attention to their existence. He does not mention any of the difficulties inherent in these approaches, nor does he attempt a comparative assessment.)

CHAPTER 12: SUMMARY AND CONCLUSIONS

In the main body (Chapters 3 to 11) of this thesis, we have attempted a detailed critical analysis of teaching approaches to the "standard" physical topics included in the new South African Common Basic Syllabuses in Physics and Physical Science, and have also considered possible teaching approaches to certain "new" topics which have been or might conceivably be included in these syllabuses. We have argued that in the case of certain topics, teaching approaches currently in use or recommended by textbooks and teachers' guides are misleading or, at worst, unlikely to promote understanding of these topics, and we have suggested that these approaches should be suitably modified or replaced by others which we have outlined. Even for topics for which it has not been necessary to recommend any major modification of teaching practice, we have thought it appropriate to discuss the background in considerable detail, and to draw attention to certain pitfalls that lie in wait for the unwary teacher.

In dealing with the main subject-matter of this thesis, we have devoted considerable attention to philosophical problems attached to particular physical concepts. Our justification for this is our belief that sooner or later the able and conscientious physics teacher with a concern for the structure of his subject will find himself driven to think about these problems, and that his confidence and effectiveness as a teacher will depend in part upon his success in achieving a consistent and defensible philosophical standpoint. The teacher who feels that he has attained clarity on such matters as the nature of definitions, physical laws and theories, the "reality" or otherwise of physical concepts and the interpretation of certain aspects of modern physics has some chance of sowing the seeds of a similar clarity of mind in his pupils.

/To.....

To some teachers, the stress laid in these pages on careful logic and the need to avoid circular arguments may seem excessive. But the development of a critical attitude is an essential aspect of the training of the pupil in scientific thinking; and one way to encourage this development is to pay careful attention to logic of presentation and to take pupils fully into one's confidence when outlining arguments the logic of which is known to be suspect or incomplete. Brighter pupils could be invited to apply their minds to the criticism of those "Zanzibar" experiments and arguments which can still be found in textbooks. Even teachers who recoil from the thought of discussing such matters with their pupils must surely grant the desirability of knowing about them themselves.

The South African physics teacher must use as little mathematics as possible in his teaching, as the mathematical knowledge he may expect from his pupils falls short of the differential and integral calculus, the binomial theorem or even the formulae for the sines and cosines of the sums and differences of angles. In our view, this should be regarded as a challenge, not a handicap. In outlining and criticizing teaching approaches in the foregoing chapters, we have made but little use of mathematics, even when venturing beyond the confines of the syllabuses in discussing the specified topics. We hope that this treatment has provided a satisfactory illustration of the use of primarily physical thinking in the elementary presentation of key topics. The trainee teacher needs to be exposed to such largely non-mathematical discussions of physical topics, not only for the obvious reason that he cannot rely heavily on mathematics in his teaching, but also because his recent training as an undergraduate student of physics may have left him with the impression that <u>any</u> discussion of the subject without the aid of

/sophisticated.....

sophisticated mathematics is bound to be superficial. To be sure, it is impossible formally to state the fundamental postulates of modern physical theories such as quantum mechanics without making use of advanced mathematical concepts, and a proper grasp of these concepts and of the appropriate mathematical theory is essential if one is fully to understand what these theories are trying to say. But even here it is possible, as we have tried to show in Chapter 10, to tell something of the story to the mathematical novice without misleading him as to its content.

In our opinion, there is an urgent need in South Africa for systematization of the "method" training of teachers of physics, particularly in regard to instruction in the content and philosophy of the subject. There should be an agreed minimum syllabus for these aspects of "physics method" and, needless to say, competent instructors should be appointed to present it and be allotted adequate time in which to deal with it. Ideally, textbooks dealing with this material should be made available to trainee teachers. Nothing of this kind is currently available in English or Afrikaans; indeed, if it was decided to provide such a textbook in a hurry, the latter would probably have to consist of a translation from one of the many Russian books that have appeared on the subject in the last few decades! (Cf the review by Galanin (1959) of the content of some of these books.) The problem cannot conveniently be solved by giving the teaching diploma student a list of suitable references to work his way through; in the absence of a suitable textbook, such a list would have to be enormously long in order adequately to cover the content of the course. We have expressed the view earlier (Section 2.9.5) that the material of the main

body of this thesis could serve as the basis for a suitable textbook. Whether or not this material is ever used in this way, we would suggest that the standpoint it adopts merits serious consideration. We have taken the view that increased rigour is desirable in the teaching of elementary physics, and we have tried to show that a willingness to be careful about matters of accuracy is by no means incompatible with clarity of presentation. There will always be room for further research into this and other aspects of elementary physics teaching, and we hope that this thesis might help to stimulate such research in South Africa.

"... and so, my friends, I cease." *

* (Beaumont and Fletcher, "The Knight of the Burning Pestle"; cf. also the end of Benjamin Britten's "Spring Symphony.")

REFERENCES

(The following abbreviations have been used for the titles of certain journals that appear in the list:

Amer. J. Phys.	=	American Journal of Physics
Contemp. Phys.	=	Contemporary Physics
Phil. Mag.	=	Philosophical Magazine
Phys. Rev.	=	Physical Review
Phys. Teacher	=	The Physics Teacher
Rev. Mod. Phys.	=	Reviews of Modern Physics
School Sci. Rev.	=	The School Science Review

The titles of all other journals and of all books cited appear in full. In this connection we must point out that the journal formerly known as the "Bulletin of the Institute of Physics and Physical Society" changed its title to "Physics Bulletin" in January, 1968. For convenience we have used the shorter title in all references to this journal. In accordance with prevailing practice the journal known since 1940 as the "American Journal of Physics" is referred to throughout under its current (abbreviated) title; its earlier title was "The American Physics Teacher."

With certain exceptions, titles of articles do not appear in the list. The exceptions are articles which appear in books, and two anonymous articles in "The Physics Teacher" on "Charging an electroscope". The latter will be found in the list under the letter "C".)

Adam, N.K. : "The Physics and Chemistry of Surfaces", Oxford University Press; Third Edition, 1941.

Ahlgren, A. : Phys. Teacher, 1, 183 (1963).

Alexander, L.M. : Amer. J. Phys., 9, 380 (1941).

Allen, H.S. : Nature, 152, 354 (1943).

Alonso, M., and Finn, E.J. : "Fundamental University Physics", Volume I, Addison-Wesley Publishing /Company.....

1967.
: Company, Inc., Reading, Massachusetts,/
American Association of Physics Teachers: Amer.J. Phys., 25, 417 (1957).

Anderson, D.L. : "The Discovery of the Electron", D. van Nostrand Company, Inc., Princeton, New Jersey, 1964.

Andrade, E.N. da C. : "Rutherford and the Nature of the Atom", Heinemann Educational Books Limited, London, 1965.

Andrews, C.L. : "Optics of the Electromagnetic Spectrum", Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1960.

Armstrong, H.L. : Amer. J. Phys., 33, 1074 (1965a).
: " " " 33, 832 (1965b).
: " " " 34, 364 (1966a).
: " " " 34, 1031 (1966b).
: School Sci.Rev., 48, 272 (1967).

Arons, A.B. : "Development of Concepts of Physics", Addison-Wesley Publishing Company, Inc., Reading, Mass., 1965.

Arons, A.B., and Bork, A.M. : Amer. J. Phys., 32, 313 (1964).

Ashford, T.A. : "From Atoms to Stars," Holt, Rinehart and Winston, New York; First Edition, 1960; Second Edition, 1967.

Association for Science Education: "An Expansion and Teachers' Guide to Physics for Grammar Schools", John Murray (Publishers) Ltd., London, 1963.
: "The 1966 Report on the Teaching of Electricity", John Murray (Publishers) Ltd., London, 1966.

Atkins, K.R. : "Physics", John Wiley & Sons, Inc., New York, 1965.

/Austern,.....

Austern, N. : Amer. J. Phys., 29, 617 (1961).

Bailey, C.H. : "The Electromagnetic Spectrum and Sound", Pergamon Press Ltd., London, 1967.

Balkenkohl, K.H. : Annalen der Physik, 14, No. 1-2, 17 (1954).

Band, W.C. : "Introduction to Quantum Statistics," D. van Nostrand Company, Inc., Princeton, N.J., 1955.

Barbour, I.G. : "Issues in Science and Religion", Prentice-Hall, Inc., Englewood Cliffs, N.J., 1966.

Barnell, H. : Contemp. Phys., 3, 415 (1962).

Barton, V.P. : Amer. J. Phys., 17, 318 (1949).

Belham, N.D.N. : Physics Education, 1, 87 (1966a).

: " " 1, 284, (1966b).

Bennet , G.A.G. : "Electricity and Modern Physics", Edward Arnold (Publishers) Ltd., London; First Edition, 1965; M.K.S. Edition, 1968.

Bergmann, G. : "Sense and Nonsense in Operationism". In: "The Validation of Scientific Theories", edited by P.G. Frank, Collier Books, New York, 1961.

Bergmann, P.G. : "The Riddle of Gravitation", Charles Scribner's Sons, New York, 1968.

Bernal, J.D. : Scientific American, 203, 124 (August, 1960).

Bertozzi, W. : Amer. J. Phys., 32, 551 (1964).

Bewley, L.V. : "Flux Linkages and Electromagnetic Induction", Dover Publications, Inc., New York, 1964.

Beyer, R.E. : Amer. J. Phys., 18, 25 (1950).

Birch, C. : Physics Education, 3, 274 (1968).

: " " 4, 185 (1969).

Birss, R.R. : Contemp. Phys., 2, 286 (1961).

Biser, R. : Phys. Teacher, 4, 304 (1966).

Bishop, G.D. : "Physics Teaching in England From Early

/Times.....

Bishop, G.D. : Times up to 1850", P.R.M. Publishers, London, 1961.

Bitter, F. : "Currents, Fields and Particles", John Wiley & Sons, Inc., New York, 1956.

Blanchard, C.H., Burnett, C.R., Stoner, R.G., and Weber, R.L. : "Introduction to Modern Physics", Prentice-Hall, Inc., Englewood Cliffs, N.J; Second Edition, 1969.

Bleaney, B.I., and Bleaney, B. : "Electricity and Magnetism", Oxford University Press, 1957.

Blitz, J. : Physics Education, 3, 164 (1968).

Bohm, D. : "Quantum Theory", Prentice-Hall, Inc., Englewood Cliffs, N.J., 1951.

: "The Special Theory of Relativity", W.A. Benjamin, Inc., New York, 1965.

Bohr, N. : Phil. Mag., 26, 1 (1913).

: "Atomic Physics and Human Knowledge", John Wiley & Sons, Inc., New York, 1958.

Bolton, W. : Contemp. Phys., 7, 210 (1966).

Bondi, H. : Rev. Mod. Phys., 29, 423 (1957).

: Reports on Progress in Physics, 22, 97 (1959).

: "Relativity and Common Sense", Heinemann Educational Books Limited, London, 1965a.

: "Relativity and Time."
In: "Time: Selected Lectures", edited by S.T. Butler and H. Messel, Pergamon Press Ltd., London, 1965b.

: Physics Education, 1, 223 (1966).

: "Assumption and Myth in Physical Theory", Cambridge University Press, 1967.

/Bopp.....

Bopp, F. : "The Principles of the Statistical Equations of Motion in Quantum Theory." In: "Observation and Interpretation in the Philosophy of Physics", edited by S.Körner, Dover Publications, Inc., New York, 1962.

Bork, A.M. : Amer. J. Phys., 31, 854 (1963).
: " " " 32, 959 (1964).
: " " " 35, 844 (1967).

Born, M. : "Atomic Physics", Blackie and Son Limited, London; Fourth Edition, 1946.
: "Physics in My Generation", Pergamon Press Limited, London, 1956.

Bosco, Brother. : Spectrum, 6, 377 (1968).

Boulind, H.F. : "Introduction to Electricity", John Murray (Publishers) Ltd., London, 1948.
: "The Teaching of Physics in Tropical Secondary Schools", Oxford University Press, 1957.
: Physics Bulletin, 12, 115 (1961).
: Nuffield Physics Question Books I-V, Longmans, Green and Co., Ltd. and Penguin Books, Ltd., London, 1966.

Braithwaite, R.B. : "Scientific Explanation", Cambridge University Press, 1953.

Brehme, R.W. : Amer. J. Phys., 36, 896 (1968).

Bridgman, P.W. : "The Logic of Modern Physics", The Macmillan Company, New York, 1927.
: "The Present State of Operationalism". In: "The Validation of Scientific Theories," edited by P.G. Frank, Collier Books, New York, 1961.

/Brommert.....

Brommert, J.W. : Spectrum, 3, 92 (1965).

: "Physics for South African Schools", Nasou Beperk, Johannesburg: Volume I, 1968; Volume II, 1969.

Brown, G. Burniston : Science Progress, 176, 619 (1956).

: " " 181, 15 (1958).

: Amer. J. Phys., 28, 475 (1960a).

: " " " 28, 820 (1960b).

: Contemp. Phys., 5, 15 (1963).

: Physics Bulletin, 16, 319 (1965).

: " " 18, 71 (1967).

Brown, F.L. : Amer. J. Phys., 31, 400 (1963).

Brown, R.C. : Physics Education, 2, 290 (1967).

Brown, S.C., and Clarke, N. (Editors) : "International Education in Physics", Massachusetts Institute of Technology and John Wiley & Sons, Inc., New York, 1960.

Brown, S.C., Clarke, N., and Tiomno, J. (Editors) : "Why Teach Physics", M.I.T. Press, Cambridge, Mass., 1964.

Buchdahl, H.A. : "The Concepts of Classical Thermodynamics", Cambridge University Press, 1966.

Bucherer, A.H. : Annalen der Physik, 28, 3, 513 (1909).

Bunge, M. : Amer. J. Phys., 24, 272 (1956).

: " " " 34, 511 (1966).

: "Foundations of Physics", Springer-Verlag, Berlin, 1967.

: Amer. J. Phys., 36, 167 (1968).

Burn, C. : "Crystals: Their Role in Nature and Science", Academic Press, Inc., New York, 1964.

Burke, J.G. : Amer. J. Phys., 34, 390 (1966).

Burling, E.M. : Amer. J. Phys., 31, 217 (1963).

/Caratheodory,.....

Caratheodory, C. : Mathematische Annalen, 67, 355 (1909).

Caro, D.E : Contemp. Phys., 5, 38 (1963).

Carr, H.Y., and Sells, R.L. : Amer. J. Phys., 28, 727 (1960).

Chalmers, J.A. : "Atmospheric Electricity", Pergamon Press Ltd., London, 1957.

Chapin, E.K. : Amer. J. Phys., 27, 617 (1959).

: Phys. Teacher, 4, 271 (1966).

"Charging an Electroscope by Induction", Phys. Teacher, 3, 29 (1965).

"Charging of an Electroscope", Phys. Teacher, 3, 185 (1965).

Charles, R.J. : Scientific American, 217, 129 (September, 1967).

Chaundy, D.C.F. : "Electrical Oscillations". In: "Modern Physical Science Report No.2", Association for Science Education, 1964.

Chraplyvy, Z.V. : Amer. J. Phys., 20, 562 (1952).

Christiansen, F.E. : Phys. Teacher, 1, 172 (1963).

Coffman, M.L. : Amer. J. Phys., 18, 378 (1950).

Cohen, B.L. : Amer. J. Phys., 35, 849 (1967).

Cohen, I.B. : "The Birth of a New Physics", Heinemann Educational Books Limited, London, 1961.

Colbeck, J.E. : School Sci. Rev., 49, 270 (1967).

: Physics Education, 3, 218 (1968).

Coleman, J.A. : "Relativity for the Layman", Penguin Books Limited, London, 1959.

Condon, E.U. : Amer. J. Phys., 1, 18 (1933).

Constable, A.R. : Physics Education, 1, 284 (1966).

Constant, F.W. : "Fundamental Laws of Physics", Addison-Wesley Publishing Company, Inc., Reading, Mass., 1963.

Cook, A.H. : Physics Education, 2, 261 (1967).

/Corson,.....

Corson, D., and
Lorrain, P. : "Introduction to Electromagnetic Fields and Waves", W.H. Freeman and Company, San Francisco, 1962.

Cottrell, A.H. : "The Mechanical Properties of Matter", John Wiley & Sons, Inc., New York, 1964.
: Scientific American, 217, 90 (September, 1967).

Cotts, R.M., and
Detenbeck, R.W. : "Matter in Motion", University of Washington, Seattle, 1966.

Coulomb's Law Committee of the American Association of
Physics Teachers : Amer. J. Phys., 18, 1 (1950a).
: Amer. J. Phys., 18, 69 (1950b).

Cowling, T.G. : "Molecules in Motion", Hutchinson & Co. (Publishers) Ltd., London, 1950.

Crawford, F.S., Jnr. : "Waves", Volume 3 of the Berkeley Physics Course, McGraw-Hill Book Company, New York, 1968.

Cullwick, E.G. : "Electromagnetism and Relativity", Longmans, Green and Company, London; Second Edition, 1959.

Dadourian, H.M. : Amer. J. Phys., 7, 241 (1939).

Daw, H.A. : Amer. J. Phys., 31, 867 (1963).

Dempster, R.R. : Amer. J. Phys., 19, 64 (1951).

Department of Education, Cape of Good Hope:
: Senior Secondary Course, 1961.
: Senior Secondary Course: Physics: Refresher Course, 1967.

Derieux, J.B. : Phys. Rev., 11, 276 (1918).

Dicke, R.H. : Scientific American, 205, 84 (December, 1961).

Dijksterhuis, E.J. : "The Mechanization of the World Picture", Oxford University Press, 1961.

Dingle, H. : "The Special Theory of Relativity",

/Methuen.....

Dingle, H. : Methuen & Co. Ltd., London, 1940.
: Proceedings of the Physical Society, 69A, 925 (1956).
: Science Progress, 50, 568 (1962).
: Quarterly Journal of the Royal Astronomical Society, 8, 252 (1967a).
: British Journal for the Philosophy of Science, 18, [illegible] (1967b).
: Nature, 216, 119 (1967c).

Dingle, H., and Samuel, H. : "A Threefold Cord", George Allen & Unwin Ltd., London, 1961.

Dirac, P.A.M. : Proceedings of the Royal Society, A133, 60 (1931).

Dorling, G.W. : Physics Education, 1, 236 (1966).

Drake, S. : Amer. J. Phys., 32, 610 (1964).

Duff, A.W. : Amer. J. Phys., 6, 219 (1938).

Duffin, W.J. : Physics Education, 1, 283 (1966).

Duncan, T. : "Practical Modern Physics", John Murray (Publishers) Ltd, London, 1967.

Duncan, J., and Starling, S.G. : "A Text Book of Physics", Macmillan & Co., Ltd., London, 1918.

Durbin, R.P., Loar, H.H., and Havens, W.W., Jnr. : Phys. Rev., 88, 179 (1952).

Eddington, A. : "The Mathematical Theory of Relativity", Cambridge University Press, 1923.

Edwards, J.D. : IEEE Student Journal, 3, 29 (November, 1965).

Ehrenhaft, F. : Journal of the Franklin Institute, 233, 235 (1942).
: Nature, 154, 426 (1944).

Ehrenhaft, F., and Banet, L. : Science, 96, 228 (1942).

Ehrlichson, H. : Amer. J. Phys., 35, 359 (1967).

/Einstein.....

Einstein, A. : Annalen der Physik, 17, 891 (1905). Translated in: "The Principle of Relativity", Methuen & Co. Ltd., London, 1923.

: "Relativity, the Special and General Theory", Methuen & Co. Ltd., London; First Edition, 1920.

: "The Meaning of Relativity", Methuen & Co. Ltd., London, 1922.

Eisberg, R.M. : "Fundamentals of Modern Physics", John Wiley & Sons, Inc., New York, 1961.

Eisenbud, L. : Amer. J. Phys., 26, 144 (1958).

: "The Conceptual Foundations of Quantum Mechanics", University of Washington, Seattle, 1966.

Eldridge, J.A. : Amer. J. Phys., 16, 463 (1948).

Ellis, B.D. : "Basic Concepts of Measurement", Cambridge University Press, 1966.

Elton, L.R.B. : Physics Education, 2, 292 (1967).

: Physics Bulletin, 19, 183 (1968).

English, A.C. : Amer. J. Phys., 35, 326 (1967).

Ericson, T.J. : Physics Education, 3, 238 (1968a).

: " " 3, 294 (1968b).

Evans, J.C. : "The Units and Standards of Measurement in the Physical Sciences." In: "Penguin Science Survey A", Penguin Books, Ltd., London, 1964.

Farley, F.J.M., Bailey, J., and Picasso, E. : Nature, 217, 17 (1968).

Feather, N. : "The Physics of Mass, Length and Time", Edinburgh University Press, 1959.

: "The Physics of Vibrations and Waves", Edinburgh University Press, 1961.

: Contemp. Phys., 7, 29 (1965).

: " " 7, 294 (1966).

Feather, N. : "Electricity and Matter", Edinburgh University Press, 1968.

Ference, M., Lemon, H.B., and Stephenson, R.J.: "Analytical Experimental Physics", University of Chicago Press; Third Edition, 1956.

Feynman, R.P., Leighton, R.B., and Sands, M.: "The Feynman Lectures on Physics", Addison-Wesley Publishing Company, Inc., Reading, Mass.: Volume I, 1963; Volume II, 1964; Volume III, 1965.

Ford, K.W. : Scientific American, 209, 122 (December, 1963).

Foucault, L. : Comptes Rendus, 30, 551 (1850).

Fowler, R.H. : Phil. Mag., 43, 785 (1922).

: "Statistical Mechanics", Cambridge University Press, 1929.

Fox, J.G. : Amer. J. Phys., 30, 297 (1962).

: " " " 33, 1 (1965).

Freier, G.D. : "University Physics", Meredith Publishing Company, New York, 1965.

French, A.P. : "Special Relativity", Thomas Nelson & Sons Ltd., London, 1968.

French, A.P., and Hudson, A.M. : "Physics - A New Introductory Course; Parts I and II", M.I.T. Science Teaching Center, 1965.

French, A.P., and Tessman, J.R. : Amer. J. Phys., 31, 201 (1963).

Frisch, D.H., and Smith, J.H. : Amer. J. Phys., 31, 342 (1963).

/Frisch,.....

Frisch, O.H. : Contemp. Phys., 3, 194 (1962).
: " " 6, 81 (1964).
: " " 7, 45 (1965).

Galanin, D.D. : "The Present Situation in the Method of High-School Physics and Its Bearing on the Polytechnical Instruction System."
In: "Methods of Teaching Physics in Soviet Secondary Schools", Proceedings of the Institute of Teaching Methods, Moscow, 1959, published by Israel Program for Scientific Translation, 1966.

Ganley, W.P. : Amer. J. Phys., 31, 510 (1963).

Gardner, Martin : "The Ambidextrous Universe", Penguin Books Limited, London, 1967.

Gardner, T.J. : School Sci. Rev., 3, 53 (1921).

Geiger, H., and
Marsden, E. : Proceedings of the Royal Society, A82, 495 (1909).
: Phil. Mag., 25, 604 (1913).

Gerjuoy, E. : Amer. J. Phys., 17, 477 (1949).

Glasser, L. : "Order into Chaos", Inaugural Lecture as Professor of Physical Chemistry, Rhodes University, 1969.

Gledhill, J.A. : Private communication to the author(1961).
: "Trends in the Matriculation Syllabuses for Physics and Physical Science over the Next Five Years", paper read at a Conference of the South African Institute of Physics, July, 1964.
: Private communication to the author(1965).
: " " " " " (1968).
: " " " " " (1969).

/Goodlet,.....

Goodlet, B.L. : "Basic Electrotechnics", Edward Arnold (Publishers) Ltd., London, 1955.

Gordon, W., and Neser, G.O. : "A Modern Textbook of Physical Science for Senior Certificate", Maskew Miller, Cape Town, no date.

Gordon, W., Neser, G.O., and Pienaar, H.N. : "Supplement to Gordon and Neser's Senior Physical Science", Maskew Miller, Cape Town, no date.

Gottlieb, H.H. : Phys. Teacher, 3, 380 (1965).

Grünbaum, A. : Amer. J. Phys., 23, 450 (1955).

Guggenheim, E.A. : Physics Education, 3, 10 (1968).

Halliday, D., and Resnick, R. : "Physics", John Wiley & Sons, Inc., New York, 1966.

Halligey, P. : Contemp. Phys., 7, 161 (1966).

Hanson, N.R. : "Patterns of Discovery", Cambridge University Press, 1958.

Harnwell, G.P., and Legge, G.J.F. : "Physics: Matter, Energy and the Universe", Reinhold Publishing Corporation, New York, 1967.

Harper, W.R. : "Contact and Frictional Electrification", Oxford University Press, 1967.

Harrap, M.J. : Physics Education, 1, 46 (1966).

Hatsopoulos, G.N., and Keenan, J.H. : "Principles of General Thermodynamics", John Wiley & Sons, Inc., New York, 1965.

Hazen, W., and Pidd, R.W. : "Physics", Addison-Wesley Publishing Company, Inc., Reading, Mass., 1965.

/Helliwell.....

Helliwell, T.M. : "Introduction to Special Relativity", Allyn and Bacon, Boston, 1966.

Helm, H. : "Teachers' Guide for Physics", Joint Matriculation Board, Pretoria, 1963.

: Spectrum, 4, 372 (1966).

: " 4, 470 (1967).

: " 6, 275 (1968).

Helm, H., Pelliesier, S.H., and Schutte, C.J. : "Fundamentele Fisika en Chemie vir Standerd 6", Afrikaanse Pers Boekhandel, Johannesburg, 1968.

: "Fundamentele Fisika en Chemie vir Standerds 7 en 8", Afrikaanse Pers Boekhandel, Johannesburg, 1969a.

: "Natuur- en Skeikunde vir Seniors", Afrikaanse Pers Boekhandel, Johannesburg, 1969b.

Helsdon, R.M. : School Sci. Rev., 48, 544 (1967a).

: " " " 49, 270 (1967b).

Hempel, C.G. : "A Logical Appraisal of Operationism". In: "The Validation of Scientific Theories", edited by P.G. Frank, Collier Books, New York, 1961.

Herzfeld, K.F., and Rice, F.O. : Phys. Rev., 31, 691 (1928).

Hesse, M.B. : British Journal for the Philosophy of Science, 2, 281 (1952).

: "Forces and Fields", Thomas Nelson & Sons Ltd., London, 1961.

: "Models and Matter". In: "Quanta and Reality", Hutchinson & Co. (Publishers) Ltd., London, 1962.

: Amer. J. Phys., 32, 905 (1964).

Hewitt, F.G. : Phys. Teacher, 1, 123 (1963).

/Hoag.....

Hoag, J.B., and Korff, S.A. : "Electron and Nuclear Physics", D. van Nostrand Company, Inc., Princeton, N.J., 1948.

Hoffmann, B. : "The Strange Story of the Quantum", Dover Publications, Inc., New York, 1959.

Holden, A. : "The Nature of Solids", Columbia University Press, New York, 1965.

: "Bonds Between Atoms", Bell Telephone Laboratories, Murray Hill, N.J., 1966.

Holden, A., and Singer, P. : "Crystals and Crystal Growing", Heinemann Educational Books Limited, London, 1961.

Holton, G. : "Introduction to Concepts and Theories in Physical Science", Addison-Wesley Publishing Company, Inc., Reading, Mass., 1952.

: Phys. Teacher, 5, 198 (1967).

Hopper, V.D. : Phys. Rev., 66, 93 (1944).

Houston, W.V. : Amer. J. Phys., 7, 373 (1939).

Hoyt, G.D. : Amer. J. Phys., 36, 368 (1968).

Huggins, E.R. : "Physics I", W.A. Benjamin Inc., New York, 1968.

Hughes, A.L., and du Bridge, L.A. : "Photoelectric Phenomena", McGraw-Hill Book Company, Inc., New York, 1932.

Hughes, V.W. : "The Lyttleton-Bondi Universe and Charge Equality."
In: "Gravitation and Relativity", edited by H.Y. Chiu and W.F. Hoffman, W.A. Benjamin Inc., New York, 1964.

International Union of Pure and Applied Physics:

: "Symbols, Units and Nomenclature in Physics", Document I.U.P. 11 (S.U.N. 65-3), 1965.

Iona, M. : Amer. J. Phys., 29, 708 (1961).
: " " " 30, 530 (1962).
: Phys. Teacher, 1, 182 (1963a).
: Amer. J. Phys., 31, 398 (1963b).
: " " " 34, 707 (1966).

Ives, H.E., and Stillwell, G.R. : Journal of the Optical Society of America, 28, 215 (1938).

Jammer, M. : "Concepts of Space", Harvard University Press, 1954.
: "Concepts of Force", Harvard University Press, 1957.
: "Concepts of Mass", Harvard University Press, 1960.
: The Conceptual Development of Quantum Mechanics", McGraw-Hill Book Company Inc., New York, 1966.

Jardine, J. : "Physics is Fun", Heinemann Educational Books Limited, London.
Book One, 1964;
Book Two, 1965;
Book Three, 1966;
Book Four, 1967.

Jaseja, T.S., Javan, A., Murray, J., and Townes, C.H. : Phys. Rev., 133, A1221 (1964).

Jeans, J. : "An Introduction to the Kinetic Theory of Gases", Cambridge University Press, 1940.

Jefimenko, C. : Amer. J. Phys., 30, 19 (1962).
: "Electricity and Magnetism", Appleton-Century-Crofts, Inc., New York, 1966.

Joint Committee of the South African Chemical Institute and Institute of Physics: Comments on Draft Common Basic Syllabus in Physical Science,

Prétoria, 1965.

Joint Matriculation Board: : Matriculation Examination Handbook, 1962.
: " " " 1963.

Kaempffer, F.J. : "Concepts in Quantum Mechanics", Academic Press, Inc., New York, 1965.
: "The Elements of Physics", Blaisdell Publishing Company, Waltham, Mass., 1967.

Kane, G. : Phys. Rev., 71, 458 (1947).

Kane, P.P. : Amer. J. Phys., 30, 385 (1962).

Katz, A. : Amer. J. Phys., 35, 882 (1967).

Katz, R. : "Introduction to the Special Theory of Relativity", D. van Nostrand Company, Inc., Princeton, N.J., 1964.

Kaye, G.W.C., and Laby, T.H. : "Tables of Physical and Chemical Constants", Longmans, Green & Co. Ltd., London; Thirteenth Edition, 1966.

Kelly, M.J. : Phys. Rev., 16, 260 (1920).

Kenworthy, R.W. : Amer. J. Phys., 9, 380 (1941).

King, A.L. : Amer. J. Phys., 30, 387 (1962).
: Phys. Teacher, 1, 122 (1963).
: " " 5, 169 (1967).

King, J.G. : Physical Review Letters, 5, 562 (1960).

Kip, A.F. : "Fundamentals of Electricity and Magnetism", McGraw-Hill Book Company, Inc., New York, 1962.

Kitaigorodskiy, A.I. : "Order and Disorder in the World of Atoms", Longmans, Green & Co. Ltd., London, 1967.

Kittel, C. : "Introduction to Solid State Physics", John Wiley & Sons, Inc., New York; Third Edition, 1966.

Kittel, C., Knight,
W.A., and Ruderman, M. : "Mechanics", Volume I of the Berkeley Physics Course, McGraw-Hill Book Company, Inc. New York, 1965.

Kleeman, D.R. : "A Kinetic Theory of Gases and Liquids", John Wiley & Sons, Inc., New York, 1920.

Knight, C.A. : "The Freezing of Supercooled Liquids", D. van Nostrand Company, Inc., Princeton, N.J., 1967.

Kolb, K.B. : Amer. J. Phys., 30, 929 (1962).

Kolm, H.H. : Science Journal, 4, 61 (September, 1968).

Landé, A. : "From Dualism to Unity in Quantum Physics", Cambridge University Press, 1960.

: Amer. J. Phys., 29, 503 (1961).

: "New Foundations of Quantum Mechanics", Cambridge University Press, 1965a.

: Amer. J. Phys., 33, 123 (1965b).

Lange, L. : Berliner königliche Gesellschaft der Wissenchaften, Mathematisch - physikalische Klasse, 333 (1885).

: "Die geschichtliche Entwicklung des Bewegunggriffes", Leipzig, 1886.

Lanni, R.P. : Phys. Teacher, 6, 35 (1968).

Lee, E.W. : "Magnetism", Penguin Books Limited, London, 1963.

Lewis, J. : "Modern Physical Science Reports No.1" and Section 11 of "Modern Physical Science Reports No.2", Association for Science Education and Nuffield Foundation, 1964.

: Physics Bulletin, 16, 81 (1965).

Lewis, G.N., and
Tolman, R.C. : Phil. Mag., 18, 510 (1909).

/Lindsay.....

Lindsay, R.B. : "Physical Mechanics", D. van Nostrand Company, Inc., Princeton, N.J., 1933.

: Amer. J. Phys., 16, 371 (1948).

: "Operationism in Physics". In: "The Validation of Scientific Theories", edited by P.G. Frank, Collier Books, New York, 1961.

: Phys. Teacher, 1, 157 (1963).

: "The Nature of Physics", Brown University Press, Rhode Island, 1968.

Lindsay, R.B., and Margenau, H. : "Foundations of Physics", McGraw-Hill Book Company, Inc., New York, 1936.

Lipson, H. : Physics Bulletin, 8, 13 (1957a).

: " " 8, 205, 209 (1957b).

Livesey, D.L., Cannon, G.H., and Ryniak, T. : "A Laboratory Course in Physics", Books One and Two, Sir Isaac Pitman & Sons, Ltd., London, 1967.

Lord, B.E., and Ritchie, W. : Physics Bulletin, 13, 3 (1962).

Losee, J. : Amer. J. Phys., 34, 430 (1966).

Lovell, K. : "The Growth of Basic Mathematical and Scientific Concepts in Children", Routledge and Kegan Paul, London 1962.

: "The Attainment and Abilities of African Children - A Brood Survey", paper read at the Leverhulme Conference on School Science Education, University of Malawi, Blantyre, March 1968 (Lovell 1968a).

: Oral communication to the author (1968b).

Lustig, H. : Amer. J. Phys., 28, 820 (1960).

/Lyngesen,.....

Lyngesen, J. : Physics Education, 2, 229 (1967).

McCaig, M. : Physics Education, 4, 185 (1969).

McCormick, W.W.: "Fundamentals of College Physics", The Macmillan Company, New York, 1965.

Macfadyen, K.A.: School Sci. Rev., 35, 371 (1954).

: Physics Bulletin, 9, 34 (1958).

McGlashan, M.L.: Physics Education, 4, 1 (1969).

Mach, E. : "The Science of Mechanics", translated by T.J. McCormack, The Open Court Publishing Company, Chicago; Second Edition, 1902.

Mackay, R.S. ; Amer. J. Phys., 17, 444 (1949).

McKeehan, L.W. : Amer. J. Phys., 16, 109 (1948).

Maclachlan, J., and McNeill, K.G.: "Matter and Energy", Clarke Irwin & Co., Toronto, 1964a.

: "Matter and Energy" - Notes for Teachers, Clarke Irwin & Co., Toronto, 1964b.

Macleod, R.R. : Physics Education, 1, 282 (1966).

MacPhail, A.L. : "Chemistry for Schools - A New Approach", Part II, Juta & Co. Ltd., Cape Town, 1964.

Magie, W.F. : "A Source Book in Physics", McGraw-Hill Book Company, Inc., New York, 1935.

Major, S.S.,Jnr: Amer. J. Phys., 28, 816 (1960).

Mara, R.T. : "Electricity and Magnetism III: Circulation Laws and Their Consequences," University of Washington Press, Seattle, 1966.

Margenau, H. : Amer. J. Phys., 10, 224 (1942).

: "The Nature of Physical Reality", McGraw-Hill Book Company, Inc., New York, 1950.

: "Interpretations and Misinterpretations of Operationalism."
In: "The Validation of Scientific Theories", edited by P.G. Frank, Collier Books, New York, 1961.

: Scientific Research, 3, 5 (September2, 1968).

Margenau, H. and
Cohen, L. : "Probabilities in Quantum Mechanics". In: "Quantum Theory and Reality", edited by M. Bunge, Springer-Verlag, Berlin, 19[illegible]

Mark, H.F. : Scientific American, 217, 149 (September, [illegible]

Massey, H.S.W. : "Negative Ions", Cambridge University Press, 19[illegible]

: "The New Age in Physics", Elek Books Limited, London; First Edition 1960; Second Edition, 1965.

Maupertuis, P.L.M.de : "Essai de cosmologie", Deuvaux, J. [illegible], Lyons, 1756.

Maxwell, J.C. : Philosophical Transactions of the Royal Society of London, 158 (1868).

: "Matter and Motion", Society for the Promotion of Christian Knowledge, London, 18[illegible]

: "A Treatise on Electricity and Magnetism," Oxford University Press; Second Edition, 1881.

: "Theory of Heat", Longmans, Green & Co., London; Tenth Edition, 1891.

Meiring, J.A. : "Kinetic Theory Model", paper read at the Michaelhouse Conference of Science Teachers, January, 1968.

Mellor, J.W. : Phil. Mag., 3, 423 (1902).

Mendelssohn, K. : "The Quest for the Absolute Zero", World University Library, London, 1966.

Mendoza, E. : School Sci. Rev., 34, 216 (1953).

: Contemp. Phys., 6, 217 (1965).

Michaelhouse Conference of Science Teachers, January, 19[illegible]

: Recommendations to the Joint Matriculation Board.

Miller, D.C. : "The Science of Musical Sounds", The Macmillan Company, New York: Second Edition, 1926.

Miller, R.C., and
Kusch, P. : Phys. Rev., 99, 1314 (1955).

Moelwyn-Hughes, E.A.: "States of Matter", Oliver and Boyd Ltd., London, 1961a.

: "Physical Chemistry", Pergamon Press Ltd., London, 1961b.

Morgan, G.D. : School Sci. Rev., 34, 241 (1953).

Morse, P.M. : "Thermal Physics", W.A. Benjamin Inc., New York, 1964.

Mossop, G.M. : "Advanced Level Atomic Physics", University of London Press Ltd., 1968.

Mott, N.F. : Contemp. Phys., 5, 401 (1964).

: Scientific American, 217, 80 (September, 1967).

Mueller, M.M. : Amer. J. Phys., 36, 166 (1968).

Nedelsky, L. : "Science Teaching and Testing", Harcourt, Brace and World, Inc., New York, 1965.

Neher, H.V., and :
Leighton, R.B: Amer. J. Phys., 31, 255 (1963).

Newton, I., translated A. Motte, revised F. Cajori:

: "Mathematical Principles of Natural Philosophy", University of California Press, Berkeley, 1947.

Noakes, G.R. : Physics Education, 4, 65 (1969).

Nuffield Physics: "Teachers' Guide, I-V", Longmans Green & Co. Ltd., and Penguin Books Ltd., London, 1966.

Ogborn, J.M. : School Sci. Rev., 47, 53 (1965).

O'Leary, A.J. : Amer. J. Phys., 15, 280 (1947a).

: " " " 15, 146 (1947b).

: " " " 15, 336 (1947c).

Orange Free State Department of Education:

: "Teachers' Guide for the Implementation of the New Syllabus in General Science for the Middle School, 'In Service' Course", Bloemfontein, 1967.

Orear, J. : "Fundamental Physics", John Wiley & Sons., Inc., New York; First Edition, 1961; Second Edition, 1967.

Osborn, J.M : "Semiconduction - a Simple Approach for Use in Schools." In: "Modern Physical Science Reports No.2", Association for Science Education, 1964.

Owen, J.B. : Spectrum, J. 11 (May, 1963).

Page, L., and Adams,
N.I., Jnr. : "Principles of Electricity", D. van Nostrand Company, Inc., Princeton, N.J., 1931.
: Amer. J. Phys., 3, 51 (1935).
: " " " 13, 141 (1945).

Paldy, L.G. : Phys. Teacher, 1, 126 (1963).

Panofsky, W.K.H., and
Phillips, M.: "Classical Electricity and Magnetism", Addison-Wesley Publishing Company, Inc., Reading, Mass.; Second Edition, 1962.

Parry, L.G. : Contemp. Phys., 6, 68 (1964).

Parry, R. : School Sci. Rev., 49, 268 (1967).

Partington, J.R: "An Advanced Treatise on Physical Chemistry", Volume I, Longmans, Green & Co. Ltd., London, 1949.

Peierls, R.E. : Contemp. Phys., 6, 129 (1964).
: Scientific American, 216, 137 (January, 1967).

Pendse, C.G. : Phil. Mag., 24, 1012 (1937).
: " " 27, 51 (1939).
: " " 29, 477 (1940).

Phillips, M. : Amer. J. Phys., 30, 309 (1962).
: Phys. Teacher, 1, 155 (1963).
: "Electricity and Magnetism II: Magnetostatics", University of Washington Press, Seattle, 1965.

Physical Science Study Committee:
: "Physics", D.C. Heath & Co., Boston; First Edition, 1960; Second Edition, 1965.

: "Laboratory Guide", D.C. Heath & Co., Boston, 1960.

: Teachers' Resource Book and Guide, Parts One to Four, D.C. Heath & Co., Boston, 1960.

: "College Physics", Raytheon Educational Corporation, Boston, 1968.

Pierce, J.R : "Electrons and Waves", Heinemann Educational Books Ltd., London, 1965.

Pippard, A.B. : "Elements of Classical Thermodynamics", Cambridge University Press, 1957.

Plimpton, S.J., & Lawton, W.E. : Phys. Rev., 50, 1066 (1936).

Pollock, R.E. : Amer. J. Phys., 31, 901 (1963).

Popper, K.R. : British Journal for the Philosophy of Science, 4, 21 (1953).

: "Philosophy of Science: a Personal Report". In: "British Philosophy in Mid-Century", edited by G.A. Macc, George Allen & Unwin Ltd., London, 1957.

: "The Logic of Scientific Discovery", Hutchinson & Co., (Publishers) Ltd., London, 1959.

: "The Propensity Interpretation of the Calculus of Probability, and the Quantum Theory." In: "Observation and Interpretation in the Philosophy of Physics", edited by S. Körner, Dover Publications, Inc., New York, 1962.

: Same as Popper (1953), reprinted in "Conjectures and Refutations", Routledge and Kegan Paul Limited, London, 1963.

: "Quantum Mechanics without 'The Observer'". In: "Quantum Theory and Reality", edited by M. Bunge, Springer-Verlag, Berlin, 1967.

Porter, A. : Transactions of the Faraday Society, 24, 108 (1928).

Pound, R.V., and Rebka, G.A.Jnr.: Physical Review Letters, 4, 337 (1960).

Pound, R.V., and Snider, J.L. : Phys. Rev., 140, B788 (1965).

Present, R.D. : "Kinetic Theory of Gases", McGraw-Hill Book Company, Inc., New York, 1958.

Prinsloo, L.A., Kottingn, C., and Koen, J.G. : "Natuur- en Skeikunde vir Seniors", Voortrekkerpers Bpk., Johannesburg, 1965.

Prinsloo, L.A., Helm, H., and otners : Teachers' Guide for Physics and Chemistry (First Part), Joint Matriculation Board, Pretoria, 1968.

Prokhovnik, S.J: "The Logic of Special Relativity", Cambridge University Press, 1967.

Pryde, J.A. : "The Liquid State", Hutchinson & Co. (Publishers) Ltd., London, 1966.

Pugh, E.M. : Amer. J. Phys., 29, 484 (1961).
: " " " 32, 879 (1964).

Pugh, E.M., and Pugh, E.W. : "Principles of Electricity and Magnetism", Addison-Wesley Publishing Company, Inc., Reading, Mass., 1960.

Purcell, E.M. : "Electricity and Magnetism", Volume 2 of the Berkeley Physics Course, McGraw-Hill Book Company, Inc., New York, 1965.

Rabi, I.I. ; Phys. Teacher, 5, 197 (1967).

Ramsey, F.P. : "Foundations of Mathematics", Kegan Paul Limited, London, 1931.

Redlich, O. : Rev. Mod. Phys., 40, 556 (1968).

Reif, F. : "Fundamentals of Statistical and Thermal Physics", McGraw-Hill Book Company, Inc., New York, 1965.

: "Statistical Physics", Volume 5 of the Berkeley Physics Course, McGraw-Hill Book Company, Inc., New York, 1967.

Reines, F., and Cowan, C.L. : Nature, 178, 446 (1956).

Rohrveld, J. : Amer. J. Phys., 26, 139 (1958).

: " " " 26, 159 (1958).

: "Relativity", In: "Teaching Physics Today", Organization for Economic Co-operation and Development, Geneva, 1965.

: Amer. J. Phys., 37, 716 (1969).

Resnick, R. : "Introduction to Special Relativity", John Wiley & Sons, Inc., New York, 1968.

Ritchie, W.R. : Physics Education, 1, 185 (1966).

Roberts, J.K., and Miller, A.R. : "Heat and Thermodynamics", Blackie & Son Ltd, London; Fifth Edition, 1960.

Robertson, H.P.: Rev. Mod. Phys., 21, 378, (1949).

Robinson, G.F. : Amer. J. Phys., 33, 746 (1965).

Rogers, E.M. : "Physics for the Inquiring Mind", Oxford University Press, 1960.

: "Teaching Physics for the Inquiring Mind", Princeton University Press, 1962.

: Public Lecture: "Why Should We Believe What Scientists Tell Us", Michaelhouse Conference of Science Teachers, 1968a.

: Oral communication to the author (1968b).

Rogers, M.M., McReynolds, A.W., and Rogers, F.T.: Phys. Rev., 57, 379 (1940).

Roll, P.G., Krotkov, R., and Dicke, R.H. : Annals of Physics, 26, 442 (1964).

Roller, D. : Amer. J. Phys., 4, 99 (1936).

Romer, A. : "The Restless Atom", Heinemann Educational Books Ltd., London, 1961.

Rose, A. : Advances in Biological and Medical Physics, 5, 211 (1957).

Rosser, W.G.V. : Amer. J. Phys., 31, 884 (1963a).

: " " " 31, 807 (1963b).

: "Introduction to the Theory of Relativity", Butterworth & Co. (Publishers) Ltd., London, 1964.

: "Introductory Relativity", Butterworths, London, 1967.

: "Classical Electromagnetism Via Relativity", Butterworths, London, 1968a.

: Physics Education, 3, 197 (1968b).

Rossi, B., and Hall, D.B. : Phys. Rev., 59, 223 (1941).

Ruchlis, H. : Phys. Teacher, 2, 175 (1964).

Rutherford, E. : "Radioactivity", Cambridge University Press; Second Edition, 1905.

: Phil. Mag., 25, 604 (1913).

: School Sci. Rev., 4, 107 (1923).

Satterly, J. : School Sci. Rev., 43, 111 (1961).

Sauter, F. : Annalen der Physik, S.6., 6, 60 (1949).

Scanlon, P.J., Henriksen, R.N., and Allen, J.R.: Amer. J. Phys., 37, 698 (1969).

Schiff, L.I. : "Quantum Mechanics", McGraw-Hill Book Company, Inc., New York; Second Edition, 1955.

Schoene, H. : "Essentials for a Minimum Course in Physics". In: "Teaching Physics Today", Organization for Economic Cooperation and Development, Geneva, 1965.

Schriever, W. : Amer. J. Phys., 5, 203 (1939).

Schrödinger, E. : "What is Matter?" Scientific American Reprint No. 241, September, 1953.

Schwartz, A.J.: Phys. Teacher, 6, 222 (1968).

Science Masters' Association:

: "The Teaching of Electricity With Special Reference to the M.K.S. System of Units", John Murray (Publishers) Ltd., London, 1954.

Science Masters' Association and Association for Women Science Teachers of Great Britain: "Physics for Grammar Schools", John Murray, London, 1961.

Scott, W.T. : "The Physics of Electricity and Magnetism", John Wiley & Sons, Inc., New York; First Edition, 1959.

Scroggie, M.G. : Wireless World, 61, 74 (February, 1955). Reprinted in "Essays on Electronics", Iliffe Publications, London, 1963.

Seamer, W.R. : School Sci. Rev., 39, 426 (1958).

Sears, F.W. : "Electricity and Magnetism", Addison-Wesley Publishing Company, Inc., Reading, Mass., 1946.

: Phys. Teacher, 1, 20 (1963a).

: Amer. J. Phys., 31, 299 (1963b).

: " " " 31, 439 (1963c).

: " " " 32, 646 (1964).

: Phys. Teacher, 6, 131, (1968).

Sears, F.W., and Brehme, R.W. : "Introduction to the Theory of Relativity", Addison-Wesley Publishing Company, Inc., Reading, Mass., 1968.

Sears, F.W., and Zemansky, M.W.: "College Physics", Addison-Wesley Publishing Company, Inc., Reading, Mass: Third Edition, 1960.

: "University Physics", Addison-Wesley Publishing Company, Inc., Reading, Mass: Third Edition, 1964.

Seeger, R.J. : Amer. J. Phys., 30, 930 (1962).

Segrè, E. : "Nuclei and Particles", W.A. Benjamin Inc., New York, 1964.

Semat, H. : "Introduction to Atomic and Nuclear Physics", Chapman & Hall, Ltd., London; Fourth Edition, 1962.

Shamos, M.H. : "Great Experiments in Physics", Holt, Rinehart & Winston, Inc., New York, 1959.

Shankland, R.S.: Amer. J. Phys., 31, 47 (1963).

: " " " 32, 16 (1964).

Sherwin, C.W. : "Basic Concepts of Physics", Holt, Rinehart & Winston Inc., New York, 1961.

Shire, E.S. : "Classical Electricity and Magnetism", Cambridge University Press, 1960.

Shortley, G., and Williams, D. : "Elements of Physics", Prentice-Hall Inc., Englewood Cliffs, N.J.; Fourth Edition, 1965.

Shurcliff, W.A.: "Polarized Light", D. van Nostrand Company, Inc., New York, 1964.

Shuter, W.L.H. : Amer. J. Phys., 33, 766 (1965).

Simon, A.W. : Amer. J. Phys., 22, 318 (1954).

Slepian, J. : Amer. J. Phys., 19, 87 (1951).

Smith, C.J. : "The General Properties of Matter", Edward Arnold (Publishers) Ltd., London; First Edition, 1953.

: "Electricity and Magnetism", Edward Arnold (Publishers) Ltd., London; Second Edition, 1959.

Smith, J.H. : "Introduction to Special Relativity", W.A. Benjamin Inc., New York, 1965.

Sommerfeld, A. : "Thermodynamics and Statistical mechanics", Lectures on Theoretical Physics, Volume V, Academic Press, Inc., New York, 1964.

South African Bureau of Standards: "Basic Guide to the Metric System in South Africa", Pretoria, 1968.

Sparberg, E. : Amer. J. Phys., 34, 377 (1966).

Spiegel, M. : "Vector Analysis", Schaum Publishing Co., New York, 1959.

Sproull, R. : Scientific American, 207, 92 (December, 1962).

Spurgin, C.B. : Physics Education, 4, 12 (1969).

Stead, G. : "Elementary Physics", J. & A. Churchill Ltd., London, 1948.

Stickley, A.R. : Amer. J. Phys., 18, 382 (1950).

Stollberg, R., and Hill, F.F. : "Physics: Fundamentals and Frontiers", Houghton Mifflin Company, Boston, 1965.

Struthers, J.A.: Phys. Teacher, 1, 27 (1963).

Subudhi, K.S., and Tiwari, P. : Amer. J. Phys., 30, 333 (1962).

Summers, R.D. : Amer. J. Phys., 14, 311 (1946).
: " " " 17, 319 (1949).

Sutton, R.M. : "Demonstration Experiments in Physics", McGraw-Hill Book Company, Inc., 1938.

Swartz, C.E. : Phys. Teacher, 4, 280 (1966).

Tabor, D. : Contemp. Phys., 6, 112 (1965).
: "Gases, Liquids and Solids", Penguin Books Ltd., London, 1969.

Taylor, E.F. : "Introductory Mechanics", John Wiley & Sons, Inc., New York, 1963.

Taylor, E.F., and Wheeler, J. : "Spacetime Physics", W.H. Freeman and Company, San Francisco, 1966.

Taylor, G.I. : Proceedings of the Cambridge Philosophical Society, 15, 114 (1909).

Taylor, L.W. : "Physics, the Pioneer Science", Volume Two, Dover Publications Inc., New York, 1959.

Ter Haar, D. : "The Quantum Nature of Matter and Radiation". In: "Turning Points in Physics", North-Holland Publishing Company, Amsterdam, 1959.

Terrell, J. : Phys. Rev., 116, 1041 (1959).

Theobald, D.W.: "The Concept of Energy", E. & F.N. Spon Ltd., London, 1966.

Thomsen, J.S. : Amer. J. Phys., 30, 294 (1962).

Thomson, G.P. : Contemp. Phys., 9, 1 (1968).

Thomson, J.O. : Science, 77, 88 (1933).

Thomson, W., and Tait, P.G. : "Treatise on Natural Philosophy", Cambridge University Press, 1886.

Thumm, W. : Phys. Teacher, 6, 38 (1968).

Tolansky, S. : "Multiple-Beam Interferometry", Oxford University Press, 1949.

: "An Introduction to Interferometry", Longmans, Green & Co. Ltd., London, 1955.

Tomer, D.W. : Phys. Teacher, 4, 278 (1966).

Toulmin, S., and Goodfield, J: "The Fabric of the Heavens", Hutchinson & Co. (Publishers) Ltd., London, 1960.

Towne, D.H. : "Wave Phenomena", Addison-Wesley Publishing Company, Inc., Reading, Mass., 1967.

Townsend, J.S.: "Electricity in Gases", Oxford University Press, 1915.

Tricker, R.A.R: "On the Teaching of Electricity in Schools", John Murray (Publishers) Ltd, London, 1955.

: Contemp. Phys., 2, 21 (1960).

: Physics Bulletin, 13, 47 (1962).

: "Early Electrodynamics: The First Law of Circulation", Pergamon Press Ltd., London, 1965.

: "The Contributions of Faraday and Maxwell to Science", Pergamon Press Ltd., London, 1966.

: "The Contribution of Science to Education", Mills and Boon, London, 1967.

Tricker, R.A.R.: Contemp. Phys., 10, 187 (1969).

Trigg, G.L. : Amer. J. Phys., 34, 71 (1966a)

: " " " 34, 988 (1966b).

: "Crucial Experiments in Quantum Physics", University of Washington, Seattle, 1966c.

Truesdell, C.A.: "Foundations of Continuum Mechanics". In: "Delaware Seminar in the Foundations of Physics", edited by M. Bunge, Springer-Verlag, Berlin, 1967.

: "Essays in the History of Mechanics," Springer-Verlag, Berlin, 1968.

Turnbull, D. : Scientific American, 218, 98 (January, 1968).

Turner, L.A. : Amer. J. Phys., 29, 71 (1961).

: " " " 30, 804 (1962).

Tyler, F. : "Heat, Light and Sound", Edward Arnold (Publishers) Ltd., London, 1951.

Victor, L.R. : "An Outline for an Approach to Teaching the Ideal Gas Scale of Temperature", lecture delivered at the Michaelhouse Conference of Science Teachers, January, 1968.

Wannier, G.H. : "Statistical Physics", John Wiley & Sons, Inc., New York, 1966.

Warburton, F.W.: Amer. J. Phys., 2, 1 (1934).

: " " " 22, 299 (1954).

Warren, J.W. : Physics Bulletin, 12, 309 (1961).

: "The Teaching of Physics". Butterworth & Co. (Publishers) Ltd., London, 1965.

: Physics Education, 3, 219 (1968).

Weber, J. : "Gravitation and Light". In: "Gravitation and Relativity", edited by H.Y. Chiu and W.F. Hoffmann, W.A. Benjamin Inc., New York, 1964.

Weber, R.L., White, M.W., and
Manning, K.V.: "College Physics", McGraw-Hill Book Company, Inc., New York: Third Edition, 1959.

Weinstock, R. : Amer. J. Phys., 29, 698 (1961).

Weisskopf, V.F.: Scientific American, 218, 15 (May, 1968).

Wessner, J. : Amer. J. Phys., 35, 544 (1967).

Westaway, F.W. : "Science Teaching", Blackie & Son Limited, London, 1929.

Whelan, P.M. : Physics Education, 3, 48 (1968).

White, S. : Contemp. Phys., 2, 39 (1960).

Whiteman, M. : "Philosophy of Space and Time", George Allen and Unwin Ltd., London, 1967.

Whitmore, B.G. : Malawi Science Teacher, 3, 61 (1967).

Whittaker, E. : "A History of the Theories of Aether and Electricity", Volume II, Thomas Nelson & Sons Ltd., London, 1951.

Wichmann, E. : "Quantum Physics", Volume 4 of the Berkeley Physics Course, McGraw-Hill Book Company, Inc., New York, 1967.

Wilkinson, D.H.: "Towards New Concepts".
In: "Turning Points in Physics", North-Holland Publishing Company, Amsterdam, 1959.

Wilkinson, W.J.: Contemp. Phys., 7, 54 (1965).

Wilks, J. : "The Third Law of Thermodynamics", Oxford University Press, 1961.

Williamson, C. : Amer. J. Phys., 26, 204 (1958).

Witmer, E. : Amer. J. Phys., 35, 40 (1967).

Wood, A. : "Acoustics", Blackie & Son Ltd., London, 1940.

Woolfson, M.W. : School Sci. Rev., 49, 704 (1968).

Worrell, F.T. : Amer. J. Phys., 28, 406 (1960).

Yarnold, G.D. : Contemp. Phys., 1, 267 (1960).

Zahn, D.T., and
Spees, A.H. : Phys. Rev., 53, 353 (1938).

Zartman, I.F. : Phys. Rev., 37, 383 (1931).

Zemansky, M.W. : "Heat and Thermodynamics", McGraw-Hill Book Company, Inc., New York; Fourth Edition, 1957; Fifth Edition, 1968.

: Amer. J. Phys., 30, 163 (1962).

: "Temperatures Very Low and Very High", D. van Nostrand Company, Inc., Princeton, N.J., 1964.

Ziman, J.M. : Scientific American, 217, 180 (September, 1967).

Ziock, K. : "Basic Quantum Mechanics", John Wiley & Sons, Inc., New York, 1969.