Enhancing the use of Large-Scale Assessment Data in South Africa: Multidimensional Item Response Theory

Submitted in fulfilment

of the requirements of the degree of

Master of Science

of Rhodes University

Tamlyn Ann Lahoud

Grahamstown, South Africa

March 3, 2023

Declaration of Authorship

I, Tamlyn Ann Lahoud, declare that Enhancing the use of Large-Scale Assessment Data: Multidimensional Item Response Theory is my own work and that it has not been submitted, in whole or in part, for any degree in any other University. I have indicated by reference and acknowledgement the areas that are not my own work.

Abstract

This research aims to enhance the use of large-scale assessment data in South Africa by evaluating assessment validity by means of multidimensional item response theory and its associated statistical techniques, which have been severely underutilised. Data from the 2014 administration of the grade 6 Mathematics annual national assessment was used in this study and all analyses were conducted using the mirt package in R. A two parameter logistic item response theory model was developed which indicated a clear alignment between the model parameters and difficulty specifications of the test. The test was found to favour learners within the central band on the ability scale. An exploratory five-dimensional item response theory model was then developed to investigate the alignment with the test specifications as evidence for construct validity. Significant discrepancies between the factor structure and the specifications of the test were identified. Notably, the results suggest that some items measured an ability that was not purely mathematical, such as reading ability, which would distort the test's representation of Mathematics ability, disadvantage learners with lower English literacy, and reduce the construct validity of the test. Further validity evidence was obtained by differential item functioning analyses which revealed that fourteen items function differently for learners from different provinces. Although possible reasons for the presence of differential item functioning among provinces were not discussed, its presence provided sufficient evidence against the validity of the test. In conclusion, multidimensional item response theory provided an effective and rigorous approach to establishing the validity of a large-scale assessment. To avoid the pitfalls of the annual national assessments, it is recommended that this multidimensional item response theory and differential item functioning techniques are utilised

for the development and evaluation of future national assessment instruments in South Africa.

General Terms: Annual National Assessments, Differential Item Functioning, Mathematics Assessment, mirt Package, Multidimensional Item Response Theory.

Acknowledgements

This research has been made possible by the support and contributions of several people. Firstly, to my supervisor: Obrigado, Dr Fábio Corrêa! Sua orientação neste processo de aprendizagem é muito apreciada. Prof Anil Kanjee, thank you for sharing your valuable insights and expertise as my co-supervisor. To Mr Jeremy Baxter, Head of Department, and the Statistics Department staff at Rhodes University, thank you for your dedication to our learning and success.

To my wonderful parents and sister, thank you for sparking my joy for learning, encouraging me to persevere and keeping my tea cup full!

Above all, thank you, Jesus, for Your wisdom and strength. Proverbs 11:1 states: "The Lord detests the use of dishonest scales, but he delights in accurate weights" (NLT). May this study be a small step towards ensuring that future large-scale assessments in South Africa produce inferences that are accurate, valid and reliable – the kind that delight the Lord.

I am grateful to have my studies supported financially by Rhodes University and the Muirhead Scholarship. The data used in this research was collected by the Department of Education in 2014.

Contents

List of Figures

List of Tables

Glossary

The list of acronyms used in this thesis is given below.

1PL	One Parameter Logistic
2PL	Two Parameter Logistic
3PL	Three Parameter Logistic
4PL	Four Parameter Logistic
AIC	Akaike Information Criterion
ANA	Annual National Assessment
ANOVA	Analysis of Variance
BIC	Bayesian Information Criterion
CFA	Confirmatory Factor Analysis
CTT	Classical Test Theory
DBE	Department of Basic Education
DIF	Differential Item Functioning
df	Degrees of Freedom
EC	Eastern Cape
EFA	Exploratory Factor Analysis
EM	Expectation-Maximization
FET	Further Education and Training
FS	Free State
GET	General Education and Training
GMH	Generalized Mantel-Haenszel
GET	General Education and Training

GP	Gauteng
HSRC	Human Sciences Research Council
ICC	Item Chracteristic Curve
IIC	Item Information Curve
IRT	Item Response Theory
KZN	Kwa-Zulu Natal
LP	Limpopo
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MH	Mantel-Haenszel
MIMIC	Multiple Indicators Multiple Causes
MIRT	Multidimensional Item Response Theory
MML	Marginal Maximum Likelihood
MP	Mpumalanga
NC	Northern Cape
NIAF	National Integrated Assessment Framework
NW	North West
PISA	Programme for International Student Assessment
PISA	Programme for International Student Assessment
SAT	Scholastic Aptitude Test
QMC	Quasi-Monte Carlo
QMCEM	Quasi-Monte Carlo Estimation-Maximisation
PISA	Programme for International Student Assessment
RMSEA	Root Mean Square Error of Approximation
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
SAT	Scholastic Aptitude Test
TIMSS	Trends in International Mathematics and Science Study
WC	Western Cape

Chapter 1

Introduction

1.1 Context of Research

1.1.1 The Annual National Assessments

The Annual National Assessments (ANAs) were nationally standardised large-scale assessments of mathematics and literacy achievement instituted annually by the Department of Basic Education (DBE) for learners in grades 1-6 and 9 of the General Education and Training (GET) phase (Spaull, 2013). The ANA results claimed to serve as a proxy for the quality of education at the GET phase at a national level (DBE, 2014a). Furthermore, these assessments were intended to identify learning deficits and use the outcomes to develop remedial responses at a school level (Spaull, 2013). When assessed longitudinally, annual assessments would help track improvements and provide information to develop and adjust targeted educational interventions.

The GET phase is the first of two phases that make up the South African schooling system as shown in Table 1.1 below. The GET phase is completed at the end of grade 9. The optional Further Education and Training (FET) phase is the subsequent phase which ends with the Grade 12 exit examination commonly referred to as the Matric Examination (Bansilal, 2012).

Table 1.1: Structure of the South African Education System, adapted from Reddy *et al.* (2019)

Phase	Subphase	Grades	Structure
GET	Foundation Phase	Reception (Grade R) to Grade 3	Primary school
	Intermediate Phase	Grade 4 to Grade 6	Primary school
	Intermediate Phase	Grade 7 to Grade 9	7: Primary school; 8 & 9: Secondary school
FET		Grade 10 to Grade 12	Secondary school

South Africa is progressing towards meeting the national goal of all children being in school for the GET phase, and the natural next phase is to ensure that quality education is provided at all schools (Bansilal, 2012). Therefore, the ANAs were one of the most important policy developments of its time (Spaull, 2013).

The South African Assessment Policy defined assessment as "the process of identifying, gathering and interpreting information about a learner's achievement, as measured against nationally agreed outcomes for a particular phase of learning" (DBE, 1998). This process comprises four steps:

1. Generating and collecting evidence of achievement;
2. Evaluating this evidence against the outcomes;
3. Recording the findings of this evaluation; and
4. Using this information to assist the learner's development and improve the process of learning and teaching.

It is critical to translate findings into usable information to practically improve the learning and teaching processes in the classroom. As a result, this has been a focal point of research (e.g. Popham, 2009, Kanjee and Sayed, 2013, Kanjee and Moloi, 2014). However, one of the major elements of consideration in measurement theory is foundational to this as the findings of these assessments will influence activities at all levels of the school system; test scores must provide meaningful and accurate inferences (Sheng and

Wikle, 2008). This requires using assessments that are equated to be comparable across years, produce reliable results, and are valid in terms of their content and construct. The evaluation of test validity from an IRT perspecive is discussed in Section 2.5 on page 34. The DBE (2014b) claimed that the piloting of the tests ensured that they were suitable for the target grade, used appropriate language, eliminated item biases, and established validity and reliability of the assessments.

Since the ANAs were intended to serve as a diagnostic tool that would be useful for highlighting strengths and weaknesses in teaching and learning, Bansilal (2012) predicted that they would be a step in the right direction towards the actualisation of quality education. However, the ANAs have been criticised extensively in literature because of the incomparability of the results, the inaccessibility of items and the extra load it put on teachers (Spaull, 2013, Graven and Venkat, 2014, Van der Berg, 2015). These critiques are explicated below.

1.1.2 Critique of the Annual National Assessments

In 2011, independent moderation and verification processes were conducted externally by the Human Sciences Research Council (HSRC) for the grade 3 and 6 ANA results only. The HSRC was also involved in the training of provincial ANA coordinators who then provided training to district officials and principals responsible for the ANA administration at their respective schools (DBE, 2011). The subsequent administrations, however, used internal moderation processes. Spaull (2013) pointed out that the lack of external verification for some grades would lead to insufficient data on baseline anchor items for those grades, and thus, the difficulty of the assessments was not equated across years.

Failure to equate the tests in terms of difficulty would lead to internal inconsistency. This is revealed by the magnitude and unpredictability of the changes between grades and years presented in the annual reports (DBE, 2012, 2013b, 2014b). Specifically considering the improvements from 2011 to 2012, Spaull (2013) acknowledged the inclusion of anchor items but criticised the lack of use of Rasch analyses which would provide a method for equating

the assessments across years. The DBE (2014b) acknowledges the need for different tests to be administered each year and the problem that this poses for comparisons across years. Based on the information available, it is unclear whether the correct procedures were followed to ensure a constant difficulty level of the tests administered for each grade from year to year (Spaull, 2013).

Even though these errors were highlighted after the second round of the ANAs, erroneous comparisons continued to be made without addressing the issues. The (DBE, 2014b) report on the ANAs provided explicit comparisons of the aggregated provincial and national results from 2012 to 2014 indicating improvements in performance each year for most grades, including grade 6 which is shown in Figure 1.1 below.

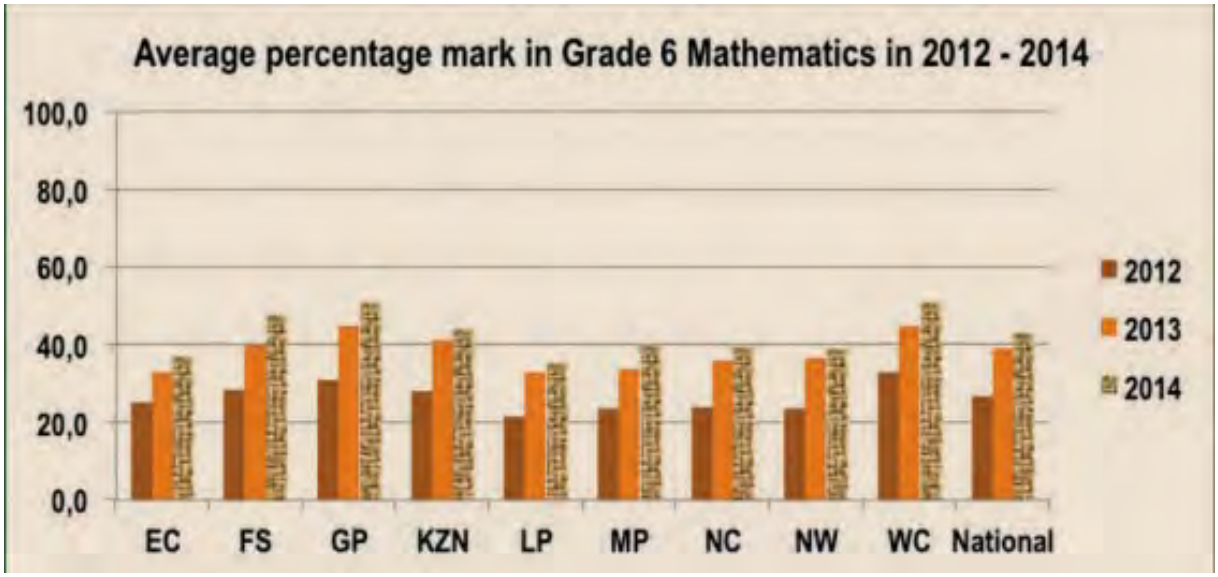


Figure 1.1: Provincial average percentage marks for Grade 6 Mathematics in 2012, 2013 and 2014 (DBE, 2014b).

Spaull (2013) stated that the DBE's misrepresentation of the real changes within the system undermined its own technical credibility as well as that of the entire ANA process going forward. In an interview, Van der Berg and Spaull (2012) explained how this lack of comparability – that results in erroneous feedback for teachers and parents – actually makes it more challenging for learning outcomes to be improved at the classroom level.

This unfortunately meant that the ANAs were unreliable for providing an accurate reflection of underlying learning. For this to be rectified, the tests would first need to be

well aligned with the curriculum and secondly, the test items calibrated to ensure that the difficulty levels would be grade–appropriate (Spaull, 2013).

In an unpublished report by Kanjee *et al.* (2013), the members of the Assessment Advisory Committee detailed some initial findings and key recommendations for the ANAs going forward from 2013. Kanjee *et al.* (2013) explained that it is critical to ensure that the reliability and validity of the current assessment instruments, including their items and sub–domains, are based on the instruments' intended uses. In addition, it is crucial that evidence is provided to confirm that items are not biased against any sub–group of learners. Furthermore, since there was no valid baseline for comparing learner performance trends across years, they suggested that advanced statistical techniques be applied for the analysis and reporting of results. Kanjee *et al.* (2013) advocated for item response theory (IRT) to be applied as it is a statistical technique that can be used to ensure the instruments' validity and reliability, and identify bias using differential item functioning analyses.

Interestingly, in a study of teachers' experiences of administrating the Mathematics ANAs conducted with a small sample (N=54) of participants in the Eastern Cape and Gauteng, Graven and Venkat (2014) found that the teachers generally considered the ANAs to be valid in terms of standard, format, scope and purpose. However, there were concerns raised about the difficulty of the language used in the assessments – preventing learners from accessing questions that they would be able to solve mathematically. Teachers were only allowed to read the questions to the learners in grades 1 and 2 but stated that this should be allowed for grade 3 and possibly even higher grades. Therefore, the teachers considered the validity of the assessment to be problematic as it pertained to learners from different linguistic backgrounds. Other concerns related to the burden of the administrative process on school resources and teacher time (Graven and Venkat, 2014).

After pressure from teacher unions in 2015, the ANAs were not written nationally and have not been continued since. In 2018, the ANAs were replaced by the National Integrated Assessment Framework (NIAF).

1.2 Motivation

The flaws in the design and implementation of the ANAs prevented them from being a meaningful diagnostic tool for characterising education quality and progress in South Africa. However, there is an abundant reserve of data that was collected when the ANAs were operative that could still be used to generate useful insights by using more rigorous statistical methods, and provide a veracious account of the GET phase of the schooling system during this period. By selecting IRT techniques, an example could be set to enhance the use of large-scale assessment data in South Africa and promote the establishment of assessment validity using IRT techniques for future large-scale assessments.

1.3 Problem Statement

This study aims to extend the use of large-scale assessment data by utilising IRT to investigate the validity and bias of an ANA instrument.

1.4 Research Question

The findings of this paper respond to the following research questions. Based on a multidimensional item response theory analysis of the assessment (MIRT), what is the utility value of the test in terms of its validity?

By responding to the central research question, the following secondary questions will be answered. To what extent does the test align with the theoretical framework prescribed by the DBE? And, for which groups of learners may the test function differently?

1.5 Research Objective

The objective of this study is to evaluate the utility value of an ANA test as verified by MIRT . The incidental objectives are to promote the use of IRT in the South African context and contribute to the body of research in this field.

1.6 Approach

The approach taken to fulfil the research objective of this study is to utilise IRT techniques to:

1. Identify the critical points of the assessment and the characteristics of the items using their parameter estimates.
2. Extrapolate the dimensions of the test and compare them to the sub-domains prescribed by the DBE (Algebra, Patterns, Space Shape, Measurement, Data Handling).
3. Identify for which covariate groups (Quintile[1], Geographic Type, Province, Gender) the test items may be biased, and evaluate significant items.

1.7 Limitations

A few limitations have been identified that may restrict the scope and/or generalizability of the findings of this study.

The author was not involved in any stage of the data collection process, which limits the scope of the differential item functioning analysis to the covariate information that was collected and recorded. Furthermore, many of the observations were empty, and the

[1]South African public schools are ranked in five quintiles of which Quintile one represents the poorest 20% of schools and Quintile five the 20% most affluent (White and Van Dyk, 2019)

reasons for the high proportion of missing values are unknown. The missing data present a limitation to the generalizability of the findings of this research.

The focus of this paper is limited to logistic IRT; therefore, polytomous items were removed during the data cleaning process. A subset of the original instrument items, excluding all polytomous items, was involved in the analyses. In the context of this study, this subset of dichotomous items is referred to as the "test". The exclusion of the 14 polytomous items limits the generalizability of the results to the full assessment instrument; however, the included test items represent all sub-domains and provide meaningful information.

1.8 Thesis Outline

The remainder of this thesis is arranged as follows:

Chapter 2: *Literature Review*

This chapter details the development and application of IRT, including multilevel IRT, MIRT and differential item functioning in the context of educational assessment.

Chapter 3: *Methodology*

The methodology chapter describes the process of conducting the data analysis.

Chapter 4: *Results*

This chapter includes the results produced during the study.

Chapter 5: *Discussion*

The discussion chapter reports on the implications of the results of the study and highlights important findings in relation to previous research.

Chapter 6: *Conclusions*

This chapter concludes the thesis, highlights the contributions to the body of research, and provides directions for future work.

Chapter 2

Literature Review

This review of literature investigates the development and utility of item response theory (IRT). First, classical test theory (CTT) is discussed to understand how limitations necessitated the development of an alternative psychometric model. Subsequently, IRT is introduced and shown to be an alternate model for assessing the validity and reliability of tests. The review of literature covers unidimensional IRT and its extension to the multidimensional context. This is followed by a review of differential item functioning and its utility in the context of IRT.

2.1 Classical Test Theory

2.1.1 Introduction

Before introducing IRT, it is helpful to understand how CTT provides a model for assessing the validity and reliability of tests.

CTT assumes that examinee i, where $i = 1, ..., N$, has a true score, T_i, and this true score can be obtained if and only if traits are constant and there are no random errors which can affect the result (Adedoyin and Mokobi, 2013). Under this approach, the number of

items answered correctly during one administration of the test is known as the raw score Y_i, and consists of examinee i's true ability T_i with some degree of measurement error ϵ_i (Crocker and Algina, 1986, de Ayala, 2009, Cappelleri *et al.*, 2014). It is important to note that "ability" is interchangeable with "trait" or "construct", each of which refers to the underlying or latent variable that is being measured by an assessment tool or instrument (Chen *et al.*, 2021).

It is assumed that T_i and ϵ_i are uncorrelated and (T_i, ϵ_i) are independent identically distributed samples from a population. The total observed score for examinee i is given as

$$Y_i = T_i + \epsilon_i \tag{2.1}$$

For equally weighted items, the total observed score for examinee i can also be expressed as

$$Y_i = \sum_{j=1}^{J} y_{ij} \tag{2.2}$$

where $y_{ij} \in \{0, 1\}$ represents the response of examinee i to dichotomous item j of J items, such that $j = 1, ..., J$. Dichotomous items have only two score categories; usually these are correct and incorrect. Polytomous items have more than two score categories.

The true score represents the mean of a theoretical distribution of observed scores (Y_i) that would be formed in an infinite number of independent assessments of a person on the same test (Finch and French, 2019). Since the distribution of random errors is assumed to be standard normal, $\mathrm{E}(\epsilon_i) = 0$ (Cappelleri *et al.*, 2014). The standard deviation of the distribution of random errors is known as the standard error of measurement (Kline, 2005). Smaller values of standard error of measurement indicate that the attribute is measured more accurately (Magno, 2009).

A major contribution of CTT was to formally take the effect of measurement error into account in the modelling of testing data (Chen *et al.*, 2021). DeVellis (2003) defines scale

reliability as the proportion of variance attributable to the true score of the latent variable. This lead to the concept of test reliability, defined as

$$\rho^2_{TY} := \frac{\text{Var}\,(T_i)}{\text{Var}\,(Y_i)} = 1 - \frac{\text{Var}\,(\epsilon_i)}{\text{Var}\,(Y_i)}, \tag{2.3}$$

which quantifies the proportion of influence of the true and error scores on attained test scores (DeVellis, 2003).

2.1.2 Limitations of Classical Test Theory

2.1.2.1 Item and Group Dependence

There are a number of limitations of CTT and its accompanying testing methods and measurement procedures. Hambleton *et al.* (1991) identified the most prominent shortcoming as the mutual dependence of examinee characteristics and test characteristics on each other.

Under this paradigm, item difficulty is defined as "the proportion of examinees in a group of interest who answer the item correctly" (Hambleton *et al.*, 1991, Reckase, 2009). This means that estimates of item difficulty are group dependent, such that a test item functions as easy or difficult given a sample of examinees and changes when the test is completed by a different sample of examinees (Magno, 2009, An and Yung, 2014). Similarly, the ability of the examinees is defined as "the expected value of the observed performance on the test of interest" and expressed by the true score T_i, which would be dependent on the difficulty of the items included in the test (Hambleton *et al.*, 1991).

An assumption of CTT is that a test is comprised of a random sample of items from the broad domain of knowledge being assessed, therefore making no inherent assumptions about the difficulty of the items that are sampled for a test (Stemler and Naples, 2021). Thus, according to the CTT paradigm, if two people score 60/100, we would conclude that their knowledge of the domain was at an equivalent level. However, they would

likely have different response patterns which would indicate different levels of proficiency being measured by the test (Adedoyin and Mokobi, 2013). Furthermore, there is poor consistency of a single test because the ability scores of examinees fluctuate depending on different occasions they take the test (Magno, 2009).

In summary, test and item characteristics fluctuate in response to the examinee context, and examinee characteristics differ according to the test contexts. The practical implications of this limitation is that, due to the difficulty in obtaining a sample that sufficiently represents the target population, it is challenging to construct tests that effectively determine the ability of the examinees (Hambleton *et al.*, 1991).

2.1.2.2 Reliability

In general, the utility of a scale is determined by examining its reliability, because this gives an indication of its measurement precision (Govender *et al.*, 2016). The internal reliability of a scale is determined based on the concept that if items have a strong relationship to their latent variable, they will have a strong relationship with each other. As a result, an internally consistent scale will also be highly intercorrelated (DeVellis, 2003). This is useful for evaluating the unidimensionality of tests because each item of a unidimensional test should be parallel (measuring the same latent trait) and therefore, correlated with one another (DeVellis, 2003).

Although the concept of test reliability is closely related to the coefficient of determination from linear regression, the true score is not directly observed like the dependent variables in regression. Therefore, a single administration of the test without additional assumptions is not sufficient to differentiate the effects of the true score and the measurement error from the observed total scores (Chen *et al.*, 2021). Cronbach (1951)'s alpha and the split-half, test-retest, and parallel-form reliability coefficients (Lord and Novick, 1968) all require additional assumptions which effectively create repeated measurements of the true score (Chen *et al.*, 2021). Furthermore, the reliability of tests is based on the correlation of scores between parallel tests; however, parallel tests are difficult to achieve in practice

(Hambleton *et al.*, 1991, An and Yung, 2014). Usually, to attain parallel tests, a single test is split in half and the correlation between the two halves is determined. This is not a perfect method: there are various ways the test could be split (DeVellis, 2003) and longer tests are usually more reliable than shorter tests (An and Yung, 2014) because an increased number of test items acts similarly to an increased sample size in terms of statistical power (DeVellis, 2003). Splitting the test halves the number of items and leads to an underestimate of the reliability of the full set of items (DeVellis, 2003). In addition, the method used to split the test items could lead to unknown biases (Hambleton and Van der Linden, 1982, DeVellis, 2003).

2.1.2.3 Test-Orientation

The fact that the test, rather than each item, is the focus of CTT leads us to the final shortfall of CTT. Since CTT is test-oriented rather than item-oriented; the total score of the test is required for measurement. Imputation is required for examinees with missing responses to be scored, which makes test development and subject scoring more difficult (An and Yung, 2014). For the same reason, CTT methods are less powerful and sometimes unfeasible when examinees answer different test items (Chen *et al.*, 2021). As a result of the test-orientation of CTT, there is also no consideration of how an examinee or a group of examinees will perform on a given item which limits the usefulness of CTT in educational testing contexts (Hambleton *et al.*, 1991).

2.1.3 An Improved model: Item Response Theory

These limitations of CTT contributed to the desire of psychometricians to develop theories and models that were able to overcome the shortfalls of this method. As a result, advances in a new technique IRT began. The development of IRT will be addressed in detail in the next section, however, the prominent advantages of IRT over CTT, as highlighted by Immekus *et al.* (2019), are summarised below.

1. Item parameter estimates and the group of examinees are independent of each other. As a result, two samples with with different distributions of ability should, given the IRT models are well-fit, provide equal item parameter estimates for the same test (Emberston and Reise, 2000).

2. The examinee's estimated ability is independent of the administered items. This allows for the estimation of examinee ability using a subset of previously calibrated items, which provides the basis for computer adaptive testing (Immekus *et al.*, 2019).

3. IRT is item-orientated rather than test-oriented, providing a measure of error for each trait estimation rather than for the entire score distribution like CTT (Crocker and Algina, 1986).

2.2 Unidimensional Item Response Theory

2.2.1 History of Item Response Theory

The development of IRT has grown from concepts originating in 19th century mathematical and psychological concepts to modern day statistical estimation principles (Bock, 1997).

In the early 1900s, Spearman (1904) created a latent factor model for intelligence using correlations between general ability tests and school examination scores. Later, Thurstone (1925) developed an age-graded scale on which to place the items of the Binet and Simon (1905) children's mental development test. The Binet and Simon test of children's mental development was typically administered such that the child would not complete all possible tasks, but rather a few items at each point for which the child has some capacity to respond correctly until the child can no longer respond correctly (Bock, 1997). The premise is that a child's mental age is judged, not from the number of items successfully completed, but by the highest age-graded items the child can complete successfully (Bock, 1997).

Using his scaling procedure, Thurstone (1925) offered a method of placing the items of this test on an age-graded scale which is shown in Figure 2.1.

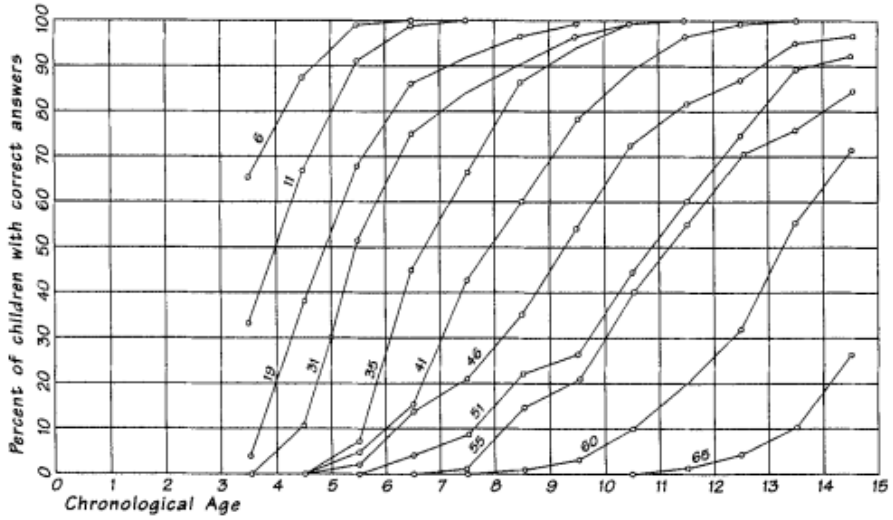


Figure 2.1: Proportions of correct response to selected items from the Binet-Simon test among children in successive age groups (Thurstone, 1925).

In Figure 2.1, points have been plotted for the percentage of correct answers for a specific item from the Binet-Simon test (y-axis) for each age group (x-axis). For each item, the points were connected by lines to form curves that give an indication of the probability of a correct answer to an item based on their age. For example, 60% of 8 year old learners responded correctly to item 41. By age 10, this proportion increases to 90%. Bock (1997) notes that Thurstone's analysis shares some basic features with IRT. Both theorise a response model to estimate the probability of success on a given item as a function of a continuous variable that represents an attribute of the examinee. In addition, they both include parameters in the model to characterise the items. Finally, the objectives of the analyses are alike as they both seek to represent the item locations and the examinees' scores as points on the continuous variable scale (Bock, 1997).

Bock (1997) also pointed out some contrasts in the ways the concepts from Thurstone (1925) appear in IRT. The continuous variable is latent (ability) in IRT, rather than manifest (age) within Thurstone's framework (Bock, 1997). Furthermore, an IRT model describes the probability of a correct response for a single person responding to a specific

item, not the probability for a proportion of correct responses for a sample of people who fall within a specific range on the continuous variable (Bock, 1997).

Although there were some early developments in the measurement of ability, the work of Lord, Novick and Rasch – beginning in the 1950s – is known to have launched the growth in popularity of IRT, as discussed below. Their goal was to move away from the dominant CTT methods and develop a technique to be able to evaluate respondents without depending on the same items included in the test (Hambleton and Jodoin, 2003).

Lord (1952) introduced and defined many of the now common IRT terms such as item characteristic curves, test characteristic curves, and standard errors conditional on latent ability (Carlson and Davier, 2017) which are explained in Section 2.2.2.2. Lord and Novick (1968) provided a general framework of IRT models, which has good statistical properties due to their natural exponential family form. The one parameter logistic model (Section 2.2.2.2) was generalised to form the two parameter and three parameter logistic models (Sections 2.2.2.3 and 2.2.2.4 respectively). The two parameter and three parameter logistic models are conventionally used in educational testing today. Around the same time, Rasch (1960) proposed the Rasch Model, which has some parallels to the 1PL model. During the next two decades, Lord expanded on his earlier work to develop IRT more completely, and also demonstrated its use on operational test scores alongside the utilisation of early software to estimate the model parameters (Lord, 1980, 2012).

Following these pioneer works, more flexible models and more powerful statistical tools have been developed to better measure human responses, promoting IRT to become one of the dominant paradigms for measurement in education (Carlson and Davier, 2017, Chen *et al.*, 2021). Most national and international large-scale assessments for monitoring education quality, such as the Programme for International Student Assessment (PISA), the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) and the Trends in International Mathematics and Science Study (TIMSS), are analysed and reported under the IRT framework. TIMSS has been using IRT increasingly since the first round in 1995 (Rasch, 1960, Yan *et al.*, 2016) and started to utilise more complex IRT models to estimate proficiency from 1999 (Lord, 2012, Yan *et al.*, 2016).

2.2.2 Item Response Theory Models

Dichotomous items are characterised by their two response options – correct or incorrect – and are fitted to logistic models. Although there are many possible models for dichotomous items, only a few are commonly applied. These models are labelled according to the number of parameters used to explain the characteristics of the test items (Reckase, 2009). Depending on the purpose of the analysis and some assumptions about the test items, different models can be selected, and these models can be verified according to how well they explain the observed test results (Hambleton *et al.*, 1991). The scope of this study is limited to dichotomous items which are modelled using logistic IRT models, and hence this text excludes information pertaining to polytomous items (items with more than two possible score categories) and their analysis. Nering and Ostini (2011) is a good resource for information about polytomous IRT models.

2.2.2.1 Basic Model Assumptions

Although IRT can be utilised in various contexts, the model assumptions are discussed in terms of their application to educational assessment.

For a sample of N examinees answering J test items;

let $Y_{ij} \in \{0, 1\}$ denote a random variable representing examinee i's dichotomous response to item j, where $Y_{ij} = 1$ indicates a correct response by examinee i to item j and $Y_{ij} = 0$ otherwise. We further denote y_{ij} as a realization of Y_{ij}, and denote $\boldsymbol{Y_i} = (Y_{i1}, ..., Y_{iJ})'$ as the random vector of binary item outcomes for examinee i, and $\boldsymbol{y_i} = (y_{i1}, ..., y_{iJ})'$ the corresponding vector of observations.

IRT assumes that $\boldsymbol{Y_i}$, $i = 1, ..., N$ are independent and models the joint distribution of $\boldsymbol{Y_i}$. IRT model assumes one latent variable, θ_i for each examinee i which is interpreted as the examinee's ability or trait level measured by the test (Chen *et al.*, 2021). An assumption of IRT is that examinees' response patterns are completely characterized by their levels on

the latent trait (Chen *et al.*, 2021). In general, θ_i is assumed to be normally distributed, with $\mu = 0, \sigma = 1$ (Fox and Glas, 2001).

The distribution of $\boldsymbol{Y_i}$ is conditional on the examinee's ability (θ_i). The specification of the conditional distribution relies on the following two assumptions: The assumption of local independence specifies that for examinee i, $Y_{i1}, ..., Y_{iJ}$ are conditionally independent given θ_i. Secondly, an assumption on the item characteristic curve.

IRT models can be plotted to visually represent the characteristics of the test items (Hambleton *et al.*, 1991). The graph of the probability of a correct response as a function of θ is typically called an item characteristic curve (ICC) and is sometimes referred to as an item probability function (Reckase, 2009). An ICC is the mathematical expression, defined as $g_j(\theta \mid \boldsymbol{\pi}_j) := P(Y_{ij} = 1 \mid \theta_i = \theta)$, where $\boldsymbol{\pi}_j$ is a generic notation for the parameters of item j, which vary according to which model is being utilised. ICCs are useful as they can be plotted to visually represent the characteristics of each item. In educational assessment, ICCs should be monotonically increasing, such that a higher ability never corresponds with a lower probability of a correct response (Chen *et al.*, 2021). The ICCs differ slightly for each model (see Sections 2.2.2.2, 2.2.2.3 and 2.2.2.4) such that each time a parameter is added, another piece of information is provided on the function. The item information curve (IIC) is a function of the first derivative of the ICC equation with respect to the ability parameter (Frey, 2018). This function provides a visualisation of the amount of information that the item provides for a given ability. It is preferable for the curve to peak close to the examinees ability to provide the most information about the ability.

Let p_i denote the proportion of items correctly answered by examinee i. As shown in Equation 2.4, the proportions undergo a logit transformation,

$$\theta_i = \ln\left(\frac{p_i}{1 - p_i}\right). \tag{2.4}$$

This transformation yields θ_i, which is the ability parameter for examinee i (Rasch, 1960, Stemler and Naples, 2021). This transformation allows the distribution of the latent trait

(ability) underlying the raw score to follow a normal distribution in a population. Most parameter estimation methods assume

2.2.2.2 The One Parameter Logistic Model

The one parameter logistic (1PL) model is one of the simplest and most widely used IRT models in various contexts (An and Yung, 2014). In the 1PL model, the difficulty parameter, β_j, which is an estimate of the difficulty level of item j, is the only parameter used to characterise the items.

This model follows a simple logistic function which uses the difference between examinee ability and item difficulty to estimate the probability of a correct response (S.J. Howie, 2009) as follows.

$$P(Y_{ij} = 1|\theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}} \tag{2.5}$$

where θ_i is the ability of examinee i (Equation 2.4) and β_j is the difficulty parameter of item j.

Items with higher β values are more difficult, such that a value of β greater than 1 indicates a difficult item. Lower ability examinees are less likely to answer these items correctly. A summary of the interpretation of item difficulty values, as shown by Bichi and Talib (2018) is presented in Table 2.1.

Table 2.1: Interpretation of the difficulty parameter estimates in an IRT Model

Difficulty Value	Interpretation
β>2	Very Difficult
1<$\beta \leq$ 2	Difficult
-1<$\beta \leq$ 1	Moderately Difficult
-2<$\beta \leq$-1	Easy
-3 < $\beta \leq$ -2	Very Easy

The horizontal axis in Figure 2.2 represents the underlying ability or latent trait, the vertical axis represents the probability of a correct response. On the ICC, item difficulty is located at the point on the horizontal axis where the curve has the steepest slope. An easier way of determining the difficulty from an ICC is by using the fact that the probability of a correct response is 0.5 for any subject whose ability is equal to the value of the difficulty parameter for that item (An and Yung, 2014). First, identify the point on the ICC that aligns with a 0.5 probability. Then, the item difficulty is the location on the x-axis that corresponds to this point. Figure 2.2 shows three items with different difficulty levels. The item represented by the blue curve has a difficulty of negative two, while the difficulty of the red and blue items are zero and two respectively. An increase in the difficulty of an item is represented on the ICC by a shift to the right. When all items from a test are plotted on a single graph, the distribution of the difficulty levels can be identified. Areas along the ability scale where the ICCs are more concentrated indicate that the test consists of more items suited to providing information on examinees from the corresponding ability level.

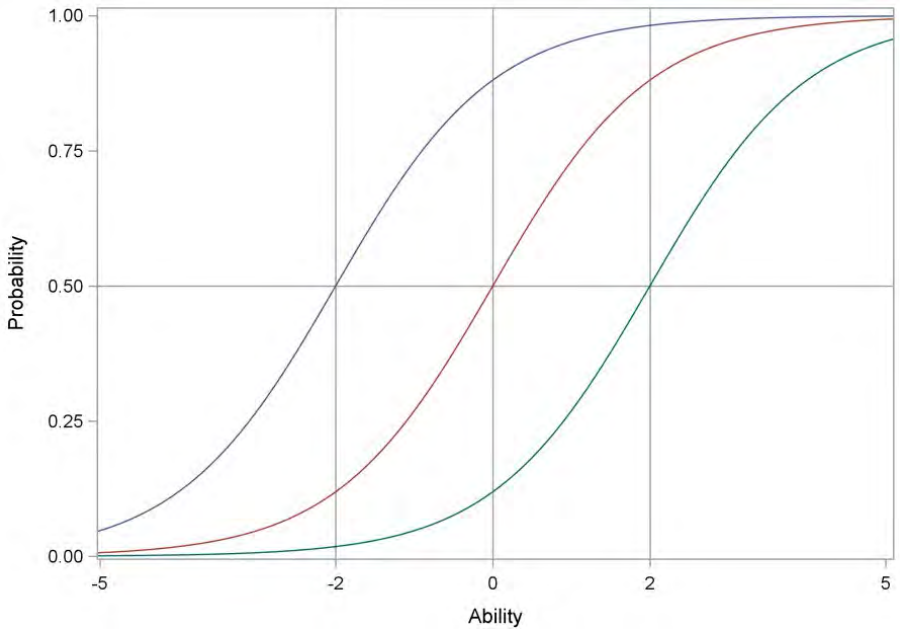


Figure 2.2: ICCs of three 1PL IRT model items showing how difficulty affects the location of an ICC (An and Yung, 2014)

The ICC in Figure 2.2 shows that the model clearly satisfies the IRT assumption of

monotonicity as a higher level of ability indicates a greater probability of answering the item correctly. It also shows that this model has a lower asymptote of zero and an upper asymptote of one.

Several characteristics of this model can be confirmed through some relatively simple visual analyses. A cursory look at the graph in Figure 2.2 shows that the curves are steeper (i.e., have a greater slope) for some values of θ than for others. When the slope is steep, the probability of a correct response to the item is quite different even for individuals with θ values that are relatively close to each other. However, in regions where the ICC is fairly flat, the θ values must be relatively far apart before the probability of correct response is noticeably different (Reckase, 2009). This means that an item can differentiate between abilities within a certain region better than other ability regions. In other words, an item provides more information about the ability of examinees within a certain region where the slope is steeper (Reckase, 2009).

The slope of the ICC can be determined at each point on the θ scale so that the steepness of the ICC can be determined for any value of θ (Reckase, 2009). The first derivative of the function, in Equation 2.5, describing the interaction of the persons and the item, provides the slope of the ICC as a function of θ.

For simplicity, let $P = P\left(Y_{ij} = 1 \mid \theta_j, \beta_i\right)$. Then, for the 1PL model (Equation 2.5), the first derivative of P with respect to θ is given by the following expression:

$$\frac{\partial P}{\partial \theta} = \frac{\partial}{\partial \theta}\left\{\frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}\right\} = P - P^2 = P(1 - P)$$

From this expression, it is clear that the slope of the ICC is zero only when the probability of correct response is zero or one. This confirms the asymptote values (of zero and one) mentioned previously.

The 1PL model has a number of advantages over more complex models including simplicity in form and mathematical properties that make the estimation of the parameters of the model particularly convenient (Reckase, 2009). One convenient mathematical property is

that there is a direct relationship between the number-correct scores for a set of test items and the estimates of θ. All persons with the same test score have the same maximum-likelihood estimate of θ (Reckase, 2009). A similar relationship exists between the number of correct responses on a test and the maximum-likelihood estimate of the β parameter for a test item. All items with the same proportion of correct responses for a sample of examinees have the same maximum likelihood β parameter estimate based on that sample (Reckase, 2009). These relationships allow the θ parameters for the examinees to be estimated independently of the β parameters for the test items (Reckase, 2009). A significant contingent of the psychometric community maintains that these properties of the 1PL model are so desirable that models that do not have these properties (such as CTT methods) should not be considered (Reckase, 2009). The perspective of this group is that only when person and item parameters can be estimated independently of each other do θ estimates result in numbers that can be called measurements (Reckase, 2009).

Another advantage is that although IRT estimates individual item locations (difficulties) and test-taker locations (abilities) independently, they are estimated on the same scale (Carlson and Davier, 2017). This allows for defining important cut points on an assessment scale and provides mechanisms for placing different test forms on the same scale (linking and equating) (Carlson and Davier, 2017).

The one parameter model considers item difficulty as the only parameter responsible for the probability of a correct response (S.J. Howie, 2009). As a result, all the items are assumed to have the same shape, however, this is not always the case in reality (An and Yung, 2014). This lead to the extension of the 1PL Model by Lord and Novick (1968) to form a general framework of IRT models, including the two parameter logistic (2PL) and three parameter logistic (3PL) models which are still conventionally used in educational testing.

2.2.2.3 The Two Parameter Logistic Model

To avoid the inaccurate assumption that all items have the same shape, the discrimination parameter (α), also known as the slope parameter, was introduced by Lord and Novick

(1968). The item discrimination parameter expresses how well an item can differentiate among examinees with different abilities. The resulting model is called the two parameter logistic (2PL) model, where the probability of a correct response is given by:

$$P(Y_{ij} = 1|\theta_i, \alpha_j, \beta_j) = \frac{e^{\alpha_j(\theta_i - \beta_j)}}{1 + e^{\alpha_j(\theta_i - \beta_j)}} \tag{2.6}$$

where θ_i is the ability of examinee i, β_j is the difficulty parameter of item j and α_j is the discrimination parameter for item j (Lord and Novick, 1968).

A feature of a good test item is that examinees with higher ability will answer it correctly more often than examinees at the lower end of the ability scale (Adedoyin and Mokobi, 2013). In other words, a good item discriminates effectively between learners of different abilities. This is measured by the discrimination parameter. A positive discrimination value indicates that ability level and probability of a correct answer are positively related and a negative discrimination value indicates a negative relationship. Among positive discrimination values, a higher discrimination level indicates that the item discriminates better between examinees of different abilities and is graphically expressed by a steeper ICC.

Figure 2.3 provides the ICCs of two items from An and Yung (2014). The lines intersect at (0; 0.5) indicating that a 0.5 probability of a correct answer is reached at the same difficulty level of zero for the three items. The curves have the same location, but they differ in shape. The item represented by the green line has the steepest slope which indicates that it has the highest discrimination. The blue curve represents the item with the lowest discrimination level as illustrated by the flattest slope.

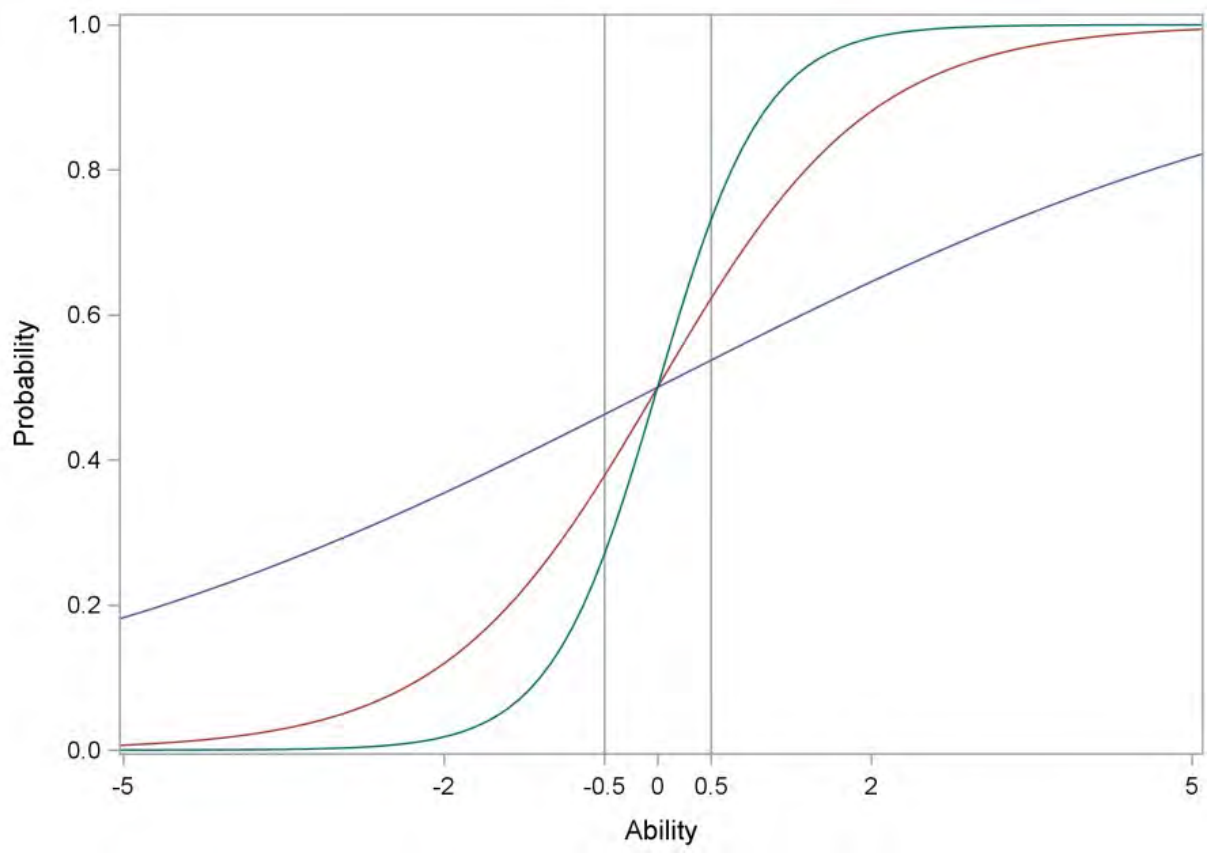


Figure 2.3: Item characteristic curves for two items from a 2PL model (Immekus *et al.*, 2019).

The interaction between difficulty and discrimination is seen by the fact that the items have a higher discriminatory power (steeper slopes) around the difficulty value on the ability continuum, and limited discriminatory power (flatter slopes) at the upper and lower ends of the continuum (Immekus *et al.*, 2019). In the context of Figure 2.3, this indicates that the items provide the most information for examinees whose ability is close to zero (the difficulty estimate for the items).

Good item discrimination is priority in test construction such that items are sometimes only considered if they have a discrimination value above a certain threshold (Reckase, 1986). Adedoyin and Mokobi (2013) notes that, although discrimination values of items in a good test range between 0.5 and 2, values above 1 are desirable and α values above 0.75 can also be acceptable. Bichi and Talib (2018) also acknowledge these general guidelines. However, Bichi and Talib (2018) also present specific interpretations of discrimination

values of a test item are shown in Table 2.2.

Table 2.2: Interpretation of the Discrimination parameter of an IRT Model

Discrimination Value (α)	Item Quality
$\alpha \geq 1.70$	Satisfactory
$\alpha \geq 1.35$	Good
$\alpha \geq 0.65$	Moderate
$\alpha \geq 0.35$	Marginal
$\alpha < 0.35$	Poor

2.2.2.4 The Three Parameter Logistic Model

In addition to the difficulty and discrimination parameters, the 3PL model includes the guessing parameter (c) which represents the probability that an examinee with a very low ability is able to answer an item correctly by and, therefore, has a greater-than-zero probability of answering an item correctly in a test (Bichi and Talib, 2018). This model generally applies to multiple-choice items.

Let c_j denote the guessing value of item j. Consistent with Equations 2.5 and 2.6, α_j and β_j denote the discrimination and difficulty parameters of item j respectively, and θ_i is the ability level of examinee i. The equation for the probability of examinee i providing a correct answer to dichotomous item j is given as

$$P(Y_{ij} = 1|\theta_i, \alpha_j, \beta_j, c_j) = c_j + (1 - c_j)\frac{\exp(\alpha_j(\theta_i - \beta_j))}{1 + \exp(\alpha_j(\theta_i - \beta_j))}. \tag{2.7}$$

Since it is a probability, the numerical value of the guessing parameter is interpreted directly (Baker, 2001). For example, if $c = 0.18$, then the probability of answering the question correctly by guessing alone is 0.18.

The guessing parameter c is the lowest value that an ICC attains. Therefore, the lower tail of each item characteristic curve approaches the guessing parameter value (Bichi and

Talib, 2018). As a result, in the 3PL model, the lower tail of the test characteristic curve approaches the sum of the guessing parameters for the test items rather than zero (Baker, 2001). This reflects the fact that under this model, very low-ability examinees can get a test score simply by guessing (Baker, 2001). The upper tail of the test characteristic curve will still approach the number of items in the test(Baker, 2001).

Warm (1978) suggested a categorisation of guessing parameters which is tabulated below.

Table 2.3: Interpretation of the guessing values in an IRT Model (Warm, 1978)

Guessing Value (c)	Interpretation
$c > 0.30$	Unacceptable
$c \leq 0.30$	Acceptable
$c \leq 0.20$	Desirable

2.2.3 Multilevel Models

There is a growing interest in educational research that necessitates the description of relations between variables of different aggregation levels (Fox and Glas, 2001). A classic example of this is where the effects of both school- and learner-level factors are involved in the assessment of learner performance.

All school systems are characterised by their hierarchical structure, with pupils grouped, nested or clustered within schools, which themselves are clustered under educational authorities (Adams *et al.*, 1997, Goldstein, 2010). In South Africa, schools are nested within municipalities which then fall under a district, and each province consists of multiple districts.

Using aggregate-level data based on school means could create problems in modelling the dependencies between the factors which could lead to unstable and misleading conclusions (Fox and Glas, 2001). The factors at a school level could be modelled separately to those at a learner level, however, this would make it impossible to study the extent to which

school and learner characteristics interact to influence learner responses and performance (Goldstein, 2010).

Thus, the mathematical modelling of learner achievement should reflect the hierarchical nature of the data structure (Clarke *et al.*, 2010). Although the multilevel models are most commonly applied in analyses involving regression and analyses of variance (ANOVAs), multilevel modelling can be applied to any statistical models involving units that are nested within aggregates (Fox and Glas, 2001).

2.2.3.1 A Multilevel Item Response Theory Model

Sulis and Toland (2017) extended the 2PL IRT model from Equation 2.6 to suit the multilevel context. Consider a population from which K schools indexed $k = 1, ..., K$ are drawn. The multilevel 2PL model is essentially a multilevel logistic regression model. Examinees $i = 1, ..., n_k$ are nested within school k where $k = 1, ..., K$. $\boldsymbol{\theta}$ represents the ability vector at the learner level $\left(\theta_{ik} \sim N\left[0, \sigma^2_{\theta^{(l)}}\right]\right)$ and a school level $\left(\theta_{0k} \sim N\left[0, \sigma^2_{\theta^{(s)}}\right]\right)$. Note that for the variances, the superscript of (l) and (s) in the subscript α indicate that it is at the learner level and school level respectively. $\boldsymbol{\lambda}$ is the vector of loadings [1] at the learner level $\left(\lambda_i^{(l)}\right)$ and at the school level $\left(\lambda_i^{(s)}\right)$.The probability of examinee i, from school k responding correctly to item j is expressed in Equation 2.8.

$$P\left(Y_{ijk} = 1 \mid \boldsymbol{\theta}, \alpha, \lambda\right) = \frac{e^{\alpha_i x_i + \lambda_i^{(s)}\theta_{jk} + \lambda_i^{(c)}\theta_{.k}}}{1 + e^{\alpha\alpha_{iji} + \lambda_i^{(s)}\theta jk + \lambda_i^{(c)}\theta_{.k}}} \tag{2.8}$$

The learner level loading measures item j's capacity to differentiate across students (between-students within-classes) with different abilities conditional upon the school level of the latent ability. The loading at the school level provides information on item j's capability to discriminate between classes on the basis of learners' abilities (Sulis and Toland, 2017).

[1] loadings are correlation coefficients between observed variables and latent common factors

2.3 Limitations of Unidimensional Item Response Theory

Unidimensional IRT models are highly useful in various contexts while possessing a simple mathematical form and maintaining robustness even when assumptions are violated. However, the application of unidimensional IRT is limited in some contexts. Unidimensional models are limited in that they may not be appropriate to multidimensional instruments i.e. instruments which measure multiple latent abilities (Immekus *et al.*, 2019). This is because the assumption of a single trait or ability being the determinant of a correct response may oversimplify the interactions between examinees and test items (Reckase, 2009). Although the utility of unidimensional IRT models is not a contentious issue in many contexts, there need for the complexity of IRT models to reflect the complexity of the interactions between examinees and test items necessitated the extension of unidimensional IRT (Reckase, 2009).

Multidimensional item response theory (MIRT) is an extension of unidimensional IRT models that provides an ideal foundation for modelling learner performance more accurately in complex domains (Hartig and Höhler, 2009, Reckase, 2009). MIRT takes multiple basic abilities into account at the same time and displays how different mixtures of the abilities are necessary for responding correctly to different test items (Hartig and Höhler, 2009). MIRT will be discussed next in Section 2.4.

2.4 Multidimensional Item Response Theory

2.4.1 Introduction to Multidimensional Item Response Theory

In practice, examinees may need to apply a variety of skills and abilities to determine the correct response to test items (de Ayala, 2009, Reckase, 2009). Naturally, these examinees would vary on a range of abilities, and specific subsets of those abilities would critical to

their performance on a test. From an IRT perspective in these contexts, it is reasonable to hypothesize that a person's responses to a set of test items or even a single item is due to their locations on multiple latent abilities (de Ayala, 2009). The number of abilities involved would depend on the breadth and complexity of the subject being tested. Due to the importance of choosing a model that provides the most complete description of the data (Sheng and Wikle, 2008), there is a need to represent the multiple latent dimensions in IRT models. Based on this premise, the development of MIRT models began.

2.4.2 Origins of Multidimensional Item Response Theory

The work of Reckase (1985) and Reckase (1986) laid the foundations for evaluating the interaction between multidimensional items and multidimensional ability distributions –associated with samples of examinees – by formally defining multidimensional IRT characteristics.

First, Reckase (1985) developed a multidimensional index of item difficulty which described the direction and distance to the most informative point in the multidimensional space for each item.

He defined multidimensional item difficulty (MID) based on three general assumptions. Consistent with unidimensional IRT, Reckase (1985) assumed monotonicity such that the probability of answering an item correctly would increase with an increase in ability on any dimension. The second assumption made was that it is advantageous to locate an item at a single point in a multidimensional space. Since the MID had typically been determined based on the points of inflection of a multidimensional item response surface, the MID was characterized by a hypersurface in the multidimensional space (McKinley and Reckase, 1983). However, this definition made for cumbersome comparisons due to it being difficult to determine whether two items were measuring the same combination of abilities. The use of this concept was greatly simplified by locating each item at a single point in the multidimensional space. The third assumption is that the most sensible classification of MID is the point where the item provides the most information about the person being measured, defined as where the item is most discriminating (Reckase, 1985).

For unidimensional models, where the probability of correctly answering a test item depends on one underlying ability dimension, scalar parameters are used. However, since the probability of a correct answer in MIRT is modelled as a function of multiple ability dimensions, MIRT models describe the interaction of ability vectors with the characteristics of test items rather than representing a single trait parameter as a scalar (Hartig and Höhler, 2009). The way this is incorporated into the model depends on the type of multidimensionality selected.

2.4.3 Types of Multidimensionality

The type of multidimensionality present in a MIRT model depends on the design of the assessment. Items can be designed to depend on one ability dimension or require a combination of ability dimensions simultaneously.

Between-item multidimensionality is incorporated into the MIRT model when separate clusters of items are used to measure each dimension to represent the single ability dimension required to correctly respond to a given item. The structure of models with between-item multidimensionality represents a combination of several unidimensional IRT models into a single model due to the simple structure of loadings. Within-item multidimensionality, on the other hand, presents a mixture of the ability dimensions required to answer a given item correctly.

The different dimensionality types may be equivalent in terms of their fit to empirical data (Hartig and Höhler, 2008). Therefore, the most appropriate type of multidimensionality to incorporate into the model depends on the intended interpretation. Thus, the research focus is an important consideration (Hartig and Höhler, 2008, Hartig and Höhler, 2009).

2.4.4 Types of Multidimensional Item Response Theory Models

Reckase (2009) explains that there are two main types of MIRT models that differ in the way that the item characteristics are combined with the information from the coordinates

of $\boldsymbol{\theta_i}$ to estimate the probability of a correct response.

2.4.4.1 Compensatory Model

The first type is based on a linear combination of θ-coordinates. This linear combination is used with a normal ogive or logistic form to determine the probability of a correct response. Since linear combinations of θ-coordinates are additive in nature and can yield the same sum with different combinations of θ-values, a low value in one dimension can be compensated for in another dimension (Reckase, 2009, Immekus *et al.*, 2019). This characteristic of the model is what lead to it being known as a compensatory model.

The most commonly used compensatory models are the multidimensional logistic models (Reckase and McKinley, 1991) and the multidimensional normal ogive IRT model (Bock *et al.*, 1988).

The 2PL compensatory model is given by the equation

$$P(Y_{ij} = 1 \mid \boldsymbol{\theta_i}, \beta_j, \boldsymbol{\alpha_j}) = \frac{\exp\left(\sum_{m=1}^{M} \alpha_{\mathrm{jm}}\theta_{\mathrm{im}} + \beta_j\right)}{1 + \exp\left(\sum_{m=1}^{M} \alpha_{\mathrm{jm}}\theta_{\mathrm{im}} + \beta_j\right)} \tag{2.9}$$

where β_j is the difficulty intercept, θ_{im} is the ability score of examinee i on dimension m, and α_{jm} is the weight of dimension m on item i. Here, the probability of a correct answer is dependent on the difficulty of the item as well as a weighted combination of the abilities (Reckase and McKinley, 1991). Jun (2014) explains that a specific ability dimension can be more influential on an item which is indicated by a higher weight for that item. In response, an examinees probability of success increases as their ability level on that dimension increases.

Kruglova *et al.* (2021) provides equations to calculate multidimensional difficulty and discrimination measures A_j and B_j respectively

$$A_j = \sqrt{\sum_{m=1}^{M} a_{jm}^2} \tag{2.10}$$

$$B_j = \frac{-d_j}{sqrt\sum_{m=1}^{M} a_{jm}^2} \tag{2.11}$$

Apart from the fact that its function uses a cumulative normal distribution, the multidimensional normal ogive IRT model is quite similar to the logistic models.

These compensatory models can be categorised as exploratory or confirmatory depending on how the dimensions are determined. In confirmatory models (e.g. the bifactor method (Holzinger and Swineford, 1937)), the relationship between the dimensions and items is specified prior to the estimation of parameters (Reckase, 2009). Conversely, exploratory models do not impose these constraints on the estimation process, and the number of dimensions is contingent on the fit of different models to the data. Sometimes exploratory models resemble confirmatory models because the number of dimensions is selected prior to estimating the parameters (Reckase, 2009).

In exploratory analyses, a standard multivariate normal distribution of the population is generally assumed such that the mean is a vector of zeros $\boldsymbol{\mu} = [\mathbf{0}]_{m\times 1}$ and the variance matrix is the identity matrix[2] such that $\boldsymbol{\Sigma} = \boldsymbol{I}_{m\times m}$. For confirmatory analyses, the parameters of the multivariate normal distribution are estimated using confirmatory factor analysis techniques from structural equation modelling.

The appropriate type of model depends on the purpose of the analysis; Reckase (2009) notes that when investigating the validity of classifications, exploratory analyses should be used among other sources of validity evidence.

[2]An identity matrix is a square matrix in which all elements of the principal diagonal are ones and all other elements are zeros

2.4.4.2 Non-Compensatory Model

The non-compensatory model is a less commonly used model type. This type divides the item into parts that are each modelled using a unidimensional model. The probability of a correct answer is then determined using the product of the probabilities for each part (Reckase, 2009). There is less compensation in this model type because the overall probability will not be higher than any of the part probabilities. However, there is still some compensation because higher ability parameters on any dimension will result in a higher probability of a correct response to the item. As a result, it is sometimes referred to as the partially compensatory model.

2.4.5 Multidimensional Item Response Theory Models

MIRT represents a broad class of probabilistic models designed to characterize an individual's likelihood of an item response based on item parameters and multiple latent traits.

In particular, MIRT locates an examinee's ability as a point in a multidimensional space which is represented in vector form. Consistent with the unidimensional models, let $i = 1, ..., N$ represent the examinees, and $j = 1, ..., J$ the test items. Now, suppose there are M latent factors (dimensions) involved, such that $k = 1, ..., M$.

Then examinee i's ability is

$$\theta_{\mathbf{i}} = \begin{bmatrix} \theta_{i1}, & \theta_{i2}, & \theta_{i3}, & \dots \theta_{iM} \end{bmatrix}',$$

with the corresponding item slopes

$$\alpha_{\mathbf{j}} = \begin{bmatrix} \alpha_{j1}, & \alpha_{j2}, & \alpha_{j3}, & \dots \alpha_{jM} \end{bmatrix}',$$

Furthermore, d_i is a scalar item intercept parameter somewhat resembling the difficulty parameter of the unidimensional model.

Consistent with the unidimensional section, this study focuses on logistic MIRT models that apply to dichotomous items only. Only the 2PL model is presented below. However, this model can easily be restricted to form the 1PL MIRT model or generalised to form the 3PL MIRT model where applicable.

2.4.5.1 The Two Parameter Logistic Multidimensional Item Response Theory Model

In the context of the 2 parameter logistic M-dimensional model, the equation for the probability of a correct response to item j by examinee i is then given as

$$P\left(y_{ij}=1 \mid \boldsymbol{\alpha}_j, d_j, \boldsymbol{\theta}_i\right)=\frac{e^{\left(\boldsymbol{\alpha}_j' \boldsymbol{\theta}_i+d_j\right)}}{1+e^{\left(\boldsymbol{\alpha}_j' \boldsymbol{\theta}_i+d_j\right)}}.$$

By letting $d_j=-\sum_{k=1}^{M} \alpha_{jk}\beta_{jk}$, where α_{jk} is an element of $\boldsymbol{\alpha}_j$, and θ_{ik} is the k^{th} element of $\boldsymbol{\theta}_i$, the exponent can also be expressed as follows to resemble the usual expression for the 2PL model

$$\sum_{k=1}^{M} \alpha_{jk}\left(\theta_{ik}-b_{jk}\right),$$

where M is the number of dimensions.

2.5 Evaluation of Validity

The utility of obtained scores to yield meaningful information is directly dependent on the quality of their psychometric properties of reliability and validity (Immekus *et al.*, 2019). Validation in assessment involves evaluating logical arguments and empirical evidence to determine whether they support the proposed interpretations of assessment results and their consequent inferences (Taylor, 2013).

Considine *et al.* (2005) provided a process of the design, analysis and selection of multiple choice questions for use in nursing research. This is presented in Figure 2.4. Their process,

although specific to one context, can easily be adapted to suit a range of contexts such as a large-scale Mathematics assessment.

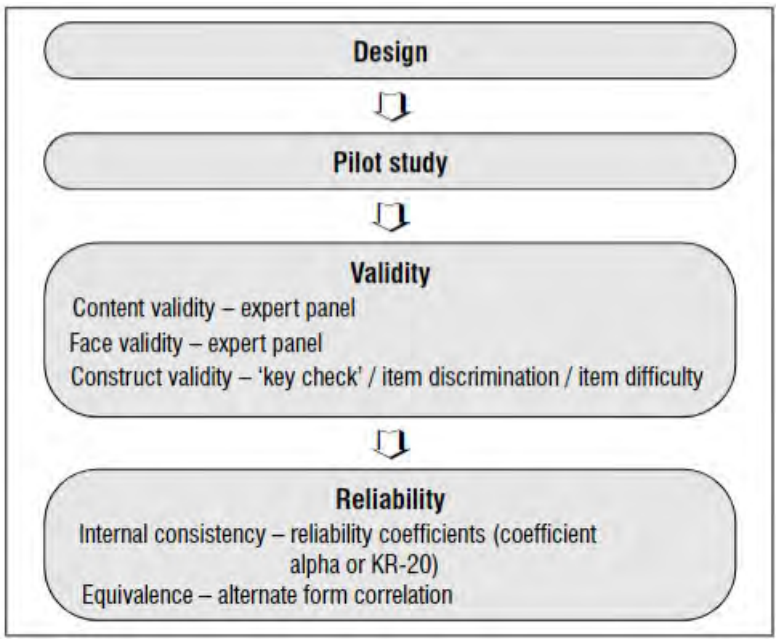


Figure 2.4: Process summary: design, analysis and selection of items (Considine *et al.*, 2005).

After the initial assessment design is finalised, Considine *et al.* (2005) indicate that the data used for this evaluation of reliability and validity would, ideally, be obtained during a pilot study so that improvements can be made before official administrations. Considine *et al.* (2005) note that the sample for the pilot study should consist of participants who would not be part of the true research sample. However, the pilot study sample would need to be representative of the target population in terms of range and level of ability and complete the test under conditions that are similar to those of the intended use of the test.

As seen in Figure 2.4, the validity analysis is conducted prior to the assessment of reliability. For an instrument to be valid, it must be reliable. On the other hand, instruments may be deemed reliable even when they are not valid (Beanland *et al.*, 1999).

Test score validity is defined as the degree to which evidence and theory support the interpretations of test scores for the proposed uses of tests (AERA, 2014). Simply put, the

validity of a test or research instrument is the extent to which the instrument measures what it is supposed to measure (Beanland *et al.*, 1999). Therefore, the first goal in validation research involves assessing the scientific rationale for the substantive and structural components of the assessment tool (Taylor, 2013). However, Taylor (2013) states that an important aspect of validity theory is that assessment tools themselves are not deemed "valid" or "invalid". Instead, these tools produce scores from which inferences about the examinees in relation to a construct performance are made, and these inferences require validation (Taylor, 2013). Although the validation process is conducted in terms of the assessment tool, it is interpreted with regard to the inferences deduced from the assessment tool scores.

The types of validity, as well as the possible approaches to assessing them, are discussed in Section 2.5.1.

2.5.1 Types of Validity

Validity encompasses multiple elements including content validity, face validity and construct validity (Considine *et al.*, 2005). Each type of validity is important for establishing overall validity of the instrument.

2.5.1.1 Content and Face Validity

Content validity addresses whether the items are relevant, appropriate and representative of the construct or cognitive processes that they are designed to test (Beanland *et al.*, 1999). Face validity involves the appearance of an assessment in terms of its clarity, readability and ease of administration, and can be considered a sub-type of content validity (Beanland *et al.*, 1999). Face validity is generally established by an editorial review and pilot study, where grammar, spelling, clarity and consistency issues are corrected (Haladyna, 1999).

In summary, face validity involves a surface level assurance that the test appears to measure what it is supposed to measure, and content validity ensures that the test represents

all aspects of the sub-domain being assessed. These analyses should be undertaken by multiple experts in the domain being examined who also have some expertise in tool development Beanland *et al.* (1999). For example, in the context of a national assessment of Mathematics, a board of experts in Mathematics and test development would need to be involved in establishing the content and face validity of the assessment tool. Since the evaluation of content and face validity would require the insight of Mathematics experts, it was beyond the scope of this study. Therefore the focus of the validity analysis for the purpose of this study is the construct validity of the test.

2.5.1.2 Construct Validity

Construct validity is the extent to which an instrument measures the theoretical construct or ability it claims to measure (Beanland *et al.*, 1999). Evaluating evidence for construct validity involves evaluating the logical arguments and empirical evidence supporting the central claim in assessment: that the scores from an assessment can be interpreted and used in a particular way (Taylor, 2013). Therefore, construct validity is related to whether or not the items measure the domain of knowledge that the assessment intends to test.

This type of validity can be established effectively using IRT methods (Ackerman, 1992). The evaluation of construct validity is but one stage of the process of designing a psychometrically sound assessment. However, due to its reliance on IRT methods, is a focus of this study. In Figure 2.4, Considine *et al.* (2005) specifies that the construct validity of a test can be established by conducting a key check and item response analyses that consider item difficulty and discrimination. Ackerman (1992)and Immekus *et al.* (2019) extended this for the evaluation of multidimensional assessments. Most validity theorists have agreed that construct-related evidence for validity is the cornerstone of validation research (Taylor, 2013).

The key check process is the process of ensuring that there is one possible correct answer when considering multiple choice items (Considine *et al.*, 2005). For open ended questions, it would likely ensure all possible correct answers are included in the memorandum to be

marked as correct. This process should be conducted by experts in the domain who review the items until a consensus is reached (Beanland *et al.*, 1999).

Generally, for multidimensional instruments, internal structure – a form of validity evidence – addresses the degree to which the relationship between items and latent dimensions align with theoretical expectations (Immekus *et al.*, 2019). This can be done by assessing the factor structure of the test. Confirmatory factor analysis (CFA) is utilised most visibly in literature for the purpose of extracting factorial validity evidence. However, exploratory factor analysis (EFA) is also used (Immekus *et al.*, 2019). The aim of CFA is to explain the covariance among the items based on a specified number of latent factors. For J items and M latent traits, the linear relationship between the items and latent variable is expressed in the factor analytic model as follows.

$$\mathbf{Y} = \Lambda_Y \xi + \epsilon \tag{2.12}$$

where $\boldsymbol{Y}$ is a $J \times 1$ vector of items, $\Lambda_Y Y$ is a J x M matrix of regression coefficients that represents the relationship between the observed, Y, and latent, ξ, variables. The disturbance term associated with each item is represented by ϵ. A set of matrices of the model parameters are used to produce an estimated covariance matrix.

A well-fitting CFA model's estimated covariace matrix will closely approximate the actual, observed covariance matrix (Immekus *et al.*, 2019). The primary aim of CFA is to construct a model that is theoretically supported and minimises the difference between the estimated and observed covariance matrix. Immekus *et al.* (2019) demonstrated how MIRT can be used to investigate the factor structure of an instrument with results that are comparable to those of factor analysis.

IRT differs from CFA in that the primary aim is to make statements about how people respond to individual items on a scale or test rather than seeking to reproduce the covariance among those items (Immekus *et al.*, 2019). In addition, CFA is based on a linear model while IRT models the probability of a specified response using a non-linear model (which is evident when visually inspecting ICCs). However, the two techniques

are similar due to the fact that they both describe the relationship between observed and latent variables using a model-based approach (Immekus *et al.*, 2019).

Although IRT methods are underrepresented in literature compared to CFA, IRT is quickly becoming the preferred method in psychological research (Immekus *et al.*, 2019).

In establishing construct validity of a multidimensional test, the test creators must specify what the test intends to measure and what the reported scores mean. The extent to which the test measures supplemental abilities (e.g. reading ability) is the extent to which the construct validity decreases (Ackerman, 1992). The completeness of construct representation is a characteristic obtained by identifying all relevant constructs that a test measures and limiting the impact of irrelevant constructs (Jun, 2014).

Another important procedure for establishing the validity of tests is the assessment of test bias, since biased items reduce the validity of a test. The first step in this process is assessing differential item functioning (DIF) (Abedlazeez, 2010, Chen *et al.*, 2021). The presence of DIF is defined as examinees from different groups having different probabilities of success on an item, after controlling for overall ability level (Abedlazeez, 2010). This would indicate that the items may function differently for different groups of individuals, or they could measure different traits for members of different groups (Chen *et al.*, 2021). This is addressed in Section 2.6.

2.6 Differential Item Functioning

The basic idea of DIF analysis is to compare groups for their performance on test items, taking into consideration that the groups may have different ability distributions (Chen *et al.*, 2021). In general, comparisons are made between a reference group and a focal group of interest. Usually, the focal group is compared against the standard of the reference group in terms of their scored item responses, so as to identify items that function differently between the two groups (Holland and Thayer, 1986, Swanson *et al.*, 2002).

DIF analysis is a very important in educational research because ignoring this phenomenon could result in the distortion of results in terms of individual and even population characteristics (Gamerman *et al.*, 2017). It is imperative to confirm that items are not biased against any group of learners (Kanjee *et al.*, 2013). This is echoed by Chen *et al.* (2021) who noted that when the implementation of interventions rely on prediction results, fairness is a significant issue to consider.

DIF is central in determining the extent to which educational and psychological instruments provide fair, comparable, and valid information about the individuals within the population of learners. It is, therefore, of paramount importance that items on such assessment tools be assessed for the presence of DIF in order to ensure that they are providing appropriate information about all examinees (Finch and French, 2019).

The detection of DIF is particularly crucial to ensure that there are no groups of examinees at a disadvantage when the examinees are from different language or cultural backgrounds (Kanjee, 2007). The efforts towards adapting mono-cultural, Westernised tests for the many cultural groups in Africa have been minimal, which is mirrored by the lack of relevant bias studies (Foxcroft, 2011). Foxcroft (2011) stated that it is unacceptable, even unethical, to use tests before having undertaken bias studies, and thus called for a higher awareness among assessment practitioners in this regard.

Although DIF analyses are not sufficient to ensure fairness of test scores, it is a necessary and major component in building a fairness argument (Finch and French, 2019). The assessment of DIF is an essential step in the validation of educational and psychological tests. If an item or the whole test is biased, the inferences and decisions about the true ability of examinees based on the test would be incorrect (Chaimongkol, 2005).

When tests yield scores that result in different meanings for different groups of examinees, it is known as bias (Association *et al.*, 1999). Bias is often ascribed to construct-irrelevant (i.e. irrelevant to the latent ability) factors that affect the test scores differently for the members of different groups. On the other hand, when construct-relevant factors affect the tests scores differently for different groups of examinees, impact occurs (Gierl, 2004). In summary, where impact is concerned, the item is a relevant measure of the ability,

and the difference between the groups is indicative of a true difference in that ability. However, for bias, one group is disadvantaged for reasons irrelevant to the ability being measured by the test which is the focus of DIF analyses (Gierl, 2004).

If the difference in the performance on an item is measured between reference and focal group members who are unmatched in terms of ability, the result is considered a measure of impact rather than of DIF (Holland and Thayer, 1986). Therefore, it is important that the groups are comparable in terms of ability.

2.6.1 The History of Differential Item Functioning

Early work investigating DIF began with Cardall and Coffman (1964) who conducted ANOVAs to identify differences between three groups taking the Scholastic Aptitude Test (SAT) and compared the relative difficulty of the items across the groups of examinees using independent correlations of difficulty within and between the groups. Later, Angoff and Ford (1973) conducted a similar study with Preliminary SAT examinees, but matched examinees according to performance, and noticed a decrease in the size of the interaction between the group and item. This implied that the difference in performance levels between the groups contributed to the interaction.

DIF has historically been referred to as item bias. Lord (1980) defined a biased item as one that has a different item response function for one group than for another. This indicates that examinees of the same ability would have a different chance of answering the item correctly, depending on their group membership.

Since then, a range of parametric and nonparametric methods for analyzing uniform and non-uniform DIF have been developed, such as the Mantel-Haenszel (MH) procedure (Holland and Thayer, 1988), logistic regression (Swaminathan and Rogers, 1990), and by visual inspection of the ICCs (Kanjee, 2010).

2.6.2 Methods of Identifying Differential Item Functioning

2.6.2.1 Mantel-Haenszel Procedure

The MH procedure (Holland and Thayer, 1988) is a popular DIF analysis procedure that compares the performance of a reference and focal group for an item based on the MH statistic (Mantel and Haenszel, 1959). When examining DIF, examinees in the reference and focal groups should be comparable. Therefore, the group difference in the latent ability distributions is accounted for in this comparison by using the total score as a matching variable to stratify the data.

As explained in Holland and Thayer (1988), the MH procedure is an item-by-item detection method. 2 x 2 x K contingency tables are constructed for each item $j = 1, ..., J$ at each score level (see Table 2.4). The contingency table data is based on the binary outcome for the item (correct or incorrect), group membership (focal or reference), and score on an overall proficiency measure (with K levels).

Table 2.4: 2 x 2 contingency table for a particular item at the jth score level

	Correct	Incorrect	Total
Reference group	A_j	B_j	n_{Rj}
Focal group	C_j	D_j	n_{Fj}
Total	m_{1j}	m_{0j}	T_j

The MH Chi-square statistic, MH_{χ^2}, is used to test for uniform DIF (Holland and Thayer, 1988) and under the null hypothesis of no DIF it has approximately a chi-squared distribution with one degree of freedom. The MH_{χ^2} statistic is given by the formula:

$$MH_{\chi^2} = \frac{\left[\left|\sum_{j=1}^{S}\left[A_j - E\left(A_j\right)\right]\right| - .5\right]^2}{\sum_{j=1}^{S}\mathrm{Var}\left(A_j\right)}, \tag{2.13}$$

where

$$\mathrm{Var}\left(A_j\right) = \frac{n_{Rj}n_{Fj}m_{1j}m_{0j}}{T_j^2\left(T_j - 1\right)}, \tag{2.14}$$

and $E(A_j) = n_{Rj}m_{1j}/T_j$.

Chi square procedures test a hypothesis, but do not produce a parametric measure of the amount of DIF exhibited by the studied item. In contrast, the MH chi square test produces both a test of statistical significance and an estimate of effect size for DIF, which indicates the size of the departure of the data from the null hypothesis (Holland and Thayer 1986; Chaimongkol 2005). This is done using the MH odds ratio estimate (α_{MH}) (Mantel and Haenszel, 1959).

This odds ratio is an estimate of DIF effect size, and is given by the formula (Holland and Thayer, 1988)

$$\alpha_{MH} = \frac{\sum_{j=1}^{S} A_j D_j / T_j}{\sum_{j=1}^{S} B_j C_j / T_j}. \tag{2.15}$$

This value represents the ratio of the odds that a person from the reference group will answer correctly to the odds for a matched person from the focal group. Therefore, α_{MH} values less than one would indicate that the item favours the focal group, and values greater than one indicate that the item favours the reference group. If α_{MH} is equal to 1, then it indicates that there is no difference in the performance of the two groups on the item at j^{th} score level. This value is easily combined across all score levels or matching variables measuring the DIF effect size.

In order to quantify the DIF effect size, Holland and Thayer (1988) proposed a logarithmic transformation of α_{MH} as follows

$$\Delta\alpha_{MH} = -2.35 \ln(\alpha_{MH}). \tag{2.16}$$

This transformation is negative for items that favour the reference group and positive for items that favour the focal group (Zwick and Ercikan, 1989).

Zwick and Ercikan (1989) then proposed the categorisation of these values into three

categories: A, B and C representing negligible, moderate and large DIF respectively as presented in Table 2.5.

Table 2.5: Interpretation of DIF effect sizes

Value	Type	Interpretation
$\|\Delta\alpha_{MH}\|<1$	A	Negligible
$1\leq\|\Delta\alpha_{MH}\|< 1.5$	B	Moderate
$1.5\leq\|\Delta\alpha_{MH}\|$	C	Large

A significant benefit of the MH procedure is that it provides a practical, inexpensive, and powerful way to detect test items that function differently between examinees from two different groups (Holland and Thayer, 1986).

For comparisons between more than two groups, the traditional approach is to conduct multiple pair-wise comparisons using the two-group comparison techniques (Kanjee, 2007). However, this process has several disadvantages including an increase in Type 1 error rates, reduction of the power of the analysis and the tendancy to be quite time consuming and costly to apply (Penfield, 2001). Penfield (2001) compared the performance of three MH procedures used for DIF across multiple groups including the MH chi-square statistic (with and without a Bonferroni adjusted alpha level) and the Generalized Mantel-Haenszel statistic (GMH) which provides a single test of significance across all groups. He found that the GMH performed the best, consistently having the highest power, with type I errors at the nominal level of 0.05.

However, since this procedure is only applicable in the context of uniform DIF, there was still a need for further developments in DIF analyses.

2.6.2.2 Logistic Regression Model of Differential Item Functioning

Although the easy implementation and test of significance associated with the MH procedure proposed by Holland and Thayer (1988) makes it particularly attractive, this procedure is limited to detecting only uniform DIF (Swaminathan and Rogers, 1990). Swaminathan and Rogers (1990) presented a logistic regression model for characterizing

DIF between two groups. By design, this model is able to detect and distinguish between uniform and non-uniform DIF. Not only is the logistic regression procedure more powerful than the MH procedure for detecting non-uniform DIF, it is also just as effective in detecting uniform DIF (Swaminathan and Rogers, 1990). It also accounts for the continuous nature of ability like item response models.

We know that for probability p,

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right). \tag{2.17}$$

By substituting p in the above expression with the probability of a correct response under the 2PL model (2.6), it is clear that the logit of the 2PL equation for person i on item j simplifies to

$$\text{logit}\left[\text{Prob}\left(Y_{ij}=1\right)\right] = \alpha_j\left(\theta_i - \beta_j\right) = -\alpha_j\beta_j + \alpha_j\theta_i. \tag{2.18}$$

Now, assuming 2PL θ estimates are representative of proficiency, and setting equations 2.20 and 2.18 equal to each other, we get

$$b_{0j} = -\alpha_j\beta_j \quad \text{and} \quad \beta_{1j} = \alpha_j. \tag{2.19}$$

The basic level-1 equation without the DIF term is

$$\text{logit}\left[\text{P}\left(Y_{ij}=1\right)\right] = b_{0j} + b_{1j}\text{x proficiency}_i. \tag{2.20}$$

Logistic regression is modelled based on the equation,

$$P(Y_{ij}=1 \mid X) = \frac{e^{(\beta_{0j}+\beta_{1j}X_{1j})}}{1+e^{(\beta_{0j}+\beta_{1j}X_{1j})}}, \tag{2.21}$$

where Y_{ij} is examinee i from group j's response, θ_i is the observed ability of examinee i, β_{0j} is the intercept parameter and β_{1j} is the slope parameter. Zimkowski *et al.* (1996)

hierarchical logistic regression as an alternate perspective for the parameterization of the 2PL IRT model. This model can be used to test for DIF and the results can be interpreted as follows. If the resulting ICCs are equivalent, that is they have equal β_0 and β_1 parameters, then there is no DIF present. If the curves are parallel, indicated by an equal β_1 parameter but different intercepts (β_0), then there is uniform DIF. Lastly, differing β_1 parameters indicate non-uniform DIF irrespective of the β_0 parameters.

However this only detects DIF rather than providing an effect size measure to quantify the magnitude of DIF that is detected. An alternate method was proposed as

$$P(Y = 1) = \frac{e^z}{1 + e^z}, \tag{2.22}$$

where $z = \tau_0 + \tau_1 X + \tau_2 g + \tau_3 X_g$. If the examinee is in the reference group then $g = 1$ and $g = 0$ indicates the examinee is in the focal group. X_g is the product of g and θ, τ_2 is the group difference, and τ_3 is the interaction between group and ability. Therefore, $\tau_2 = \beta_{01} - \beta_{02}$ and $\tau_3 = \beta_{11} - \beta_{12}$.

Following the same line of reasoning as previously, the DIF is uniform if $\tau_2 \neq 0$ and $\tau_3 = 0$, and non-uniform if $\tau_3 \neq 0$ for any τ_2.

In order to simultaneously test these hypotheses, a matrix C is introduced:

$$C_r = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \tag{2.23}$$

Here, $H_0 : C_\tau = 0$ and $H_A : C_\tau \neq 0$.

To evaluate the presence of uniform and non-uniform DIF on the item of interest, the hypothesis is tested for the reference group against two focal groups using a chi square (Ukanda, 2019). The test statistic is calculated as follows

$$\chi^2 = \hat{\tau}' C' \left(C \sum C' \right)^{-1} C \hat{\tau}' \tag{2.24}$$

which has a χ^2 distribution with 2 degrees of freedom (Zimkowski *et al.*, 1996).

Therefore, we reject H_0 when χ^2_{obs} from equation **??** $> \chi^2_{\alpha,2}$.

To test for the presence of uniform DIF, the improvement in chi-square model fit associated with adding a term for group membership (G) against the baseline model is examined. Computationally, Model 2 (i.e. $z = \beta_0 + \beta_1 X + \beta_2 G$) is subtracted from Model 1.

The presence of non-uniform DIF is tested by examining the improvement in chi-square model fit associated with adding a term for group membership (G) and a term for the interaction between test score and group membership (X_G) against model 2. In other words, Model 3 (i.e.$z = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 X_G$) subtracted from Model 2.

Large sample sizes have always been problematic for significance tests based on chi-square statistics. This is because the increased sample size increased the power so much that a miniscule diference is detected as a significant misfit between the model and data. As a result, chi-square tests become almost redundant for large samples as the outcome is already known (Martin-Löf, 1974)

Raykov (1998) – who conducted research on personality – defended that in this context, the appropriate reference distribution for the chi-square value is the non-central chi-square (with positive non-centrality) rather than a central chi-square distribution. Based on this, the large sample size and model complexity can be accounted for by dividing by the sample size and the model degrees of freedom to get the square root of the non-centrality parameter. This would yield the root mean square error of approximation (RMSEA). The RMSEA has been found to be relatively unaffected by sample size by Browne and Cudeck (1993). Raykov (1998) suggests that it may be more widely applicable to research in this field.

Various procedures are available to calculate effect size based on different formulae for the logistic regression coefficient of determination (Zumbo, 1999). To determine if either

uniform or non-uniform DIF is present, an omnibus test of the hypothesis $H_0 : \tau_2 = \tau_3 = 0$ is recommended by Zumbo and Thomas (1996).

With similarities to simple linear regression, the R^2 statistic can be partitioned into components to represent the effects unrelated to DIF (τ_0 and τ_1), those for uniform DIF (τ_2) and those related to non-uniform DIF (τ_3). As a result, three R^2 values are produced: R_1^2 is derived from the model with only τ_0 and τ_1, R_2^2 is from the model with τ_2 also included and R_3^2 is from the full model including τ_3 too.

These values are then used to calculate the various DIF effect sizes. Overall DIF effect size is indicated by $R^2\Delta = R_3^2 - R_1^2$. $R^2\Delta - U = R_2^2 - R_1^2$ which reflects uniform DIF effect size while $R^2\Delta - NU = R_3^2 - R_2^2$ represents the effect size for non-uniform DIF.

(Zumbo and Thomas, 1996) proposed $R^2\Delta$, a weighted least squares effect size measure when logistic regression is used in DIF detection to quantify the magnitude of uniform or non-uniform DIF. However, there was a high Type I error rate when using their categorisation. Jodoin and Gierl (2001) improved the categorisation to reduce Type I error rates as follows. This improved categorisation has been used in other studies (e.g. Kanjee (2007)).

1. Negligible DIF: $R^2\Delta < 0.035$
2. Moderate DIF: $0.035 \leq R^2\Delta \leq 0.070$
3. Large DIF: $R^2 > 0.070$

2.6.2.3 Differential Item Functioning in Item Response Theory

IRT models have been enhanced by the incorporation of DIF to identify items that behave differently for these groups within the population, and for examining relationships between the latent trait and other variables (Finch and French, 2019). DIF manifests as examinees who have the same ability from different groups having a different probability of correctly answering the item (Gamerman *et al.*, 2017).

Gamerman *et al.* (2017) outlines two general approaches to dealing with DIF items: either to discard them from the analysis or incorporate them into the analysis. Since items with DIF may be informative of factors, such as teaching and learning issues or cultural aspects, that affect assessment, the second approach is generally preferred, and DIF detection, quantification, and sometimes explanation is included in the main analysis (Gamerman *et al.*, 2017).

The integration of covariates into the IRT model differs from the test responses (von Davier and Sinharay, 2014). The method depends how the covariate affects the distributions of the latent traits and the responses to the test items as illustrated in Figure 2.5 (Chen *et al.*, 2021).

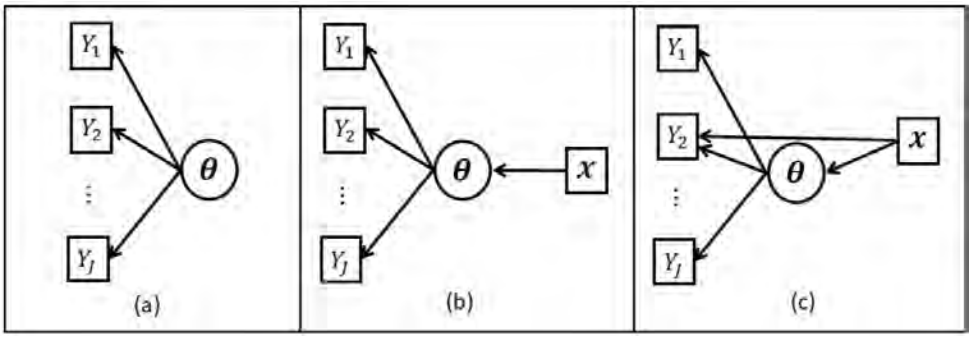


Figure 2.5: Path diagrams for three IRT models, including (a) an IRT model without covariates, (b) a latent regression model, and (c) a multiple indicators multiple cause model. The individual subscript i is omitted here for simplicity. (Chen *et al.*, 2021).

If the covariates in the model affect the latent trait distribution, but not the item responses directly, DIF is said to be uniform (Mislevy, 1985), and the model is a latent regression model as shown in Figure 2.5. Here, the probability of correctly answering an item is uniformly greater for one group, independent of ability (Gamerman *et al.*, 2017) because only the dificulty parameters differ. This type of model is particularly useful for the analysis of large-scale assessments where the distributions of the latent traits at a group level are relevant to policy (Chen *et al.*, 2021).

Chen *et al.* (2021) shows how covariate information would be included in a unidimensional model using this method by assuming θ_i to follow a normal distribution $N\left(\mathbf{x}_i'\boldsymbol{\beta}, 1\right)$ instead

of a standard normal distribution, where $\boldsymbol{\beta}$ is a p-dimensional vector of the coefficients and x_i indicates the covariate group. In this context, the covariates do not directly affect the distribution of item responses, and the remaining model assumptions, including local independence and ICC, stay the same. However, the covariates do influence the mean of the latent trait distribution. Graphically, this means that the item characteristic curves from the two groups are parallel because the difference is only found in the difficulty parameter (Gamerman *et al.*, 2017). This is visualised in Figure 2.6.

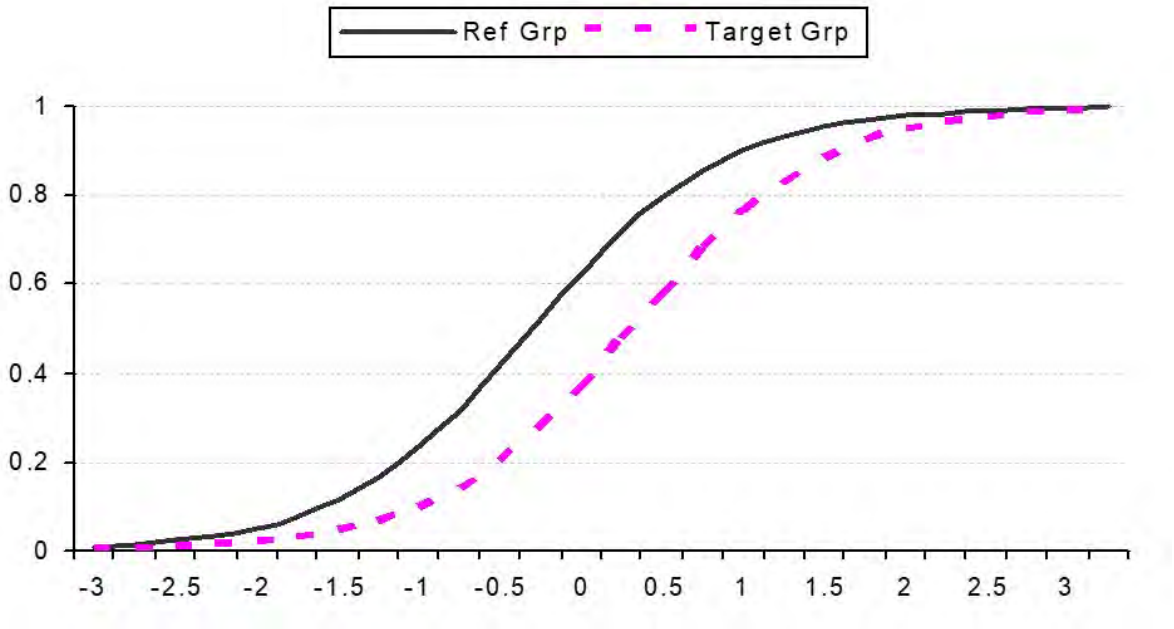


Figure 2.6: ICCs of an item presenting uniform DIF (Kanjee, 2010).

Multiple indicators multiple cause (MIMIC) models, represented in Figure 2.5, were introduced for contexts where the covariates affect the distribution of the latent trait as well as the responses to some items (Chen *et al.*, 2021). A simple MIMIC model for DIF can be set as follows.

Let $x_i \in 0, 1$ be a single binary covariate that indicates the group membership as reference or focal group respectively. Suppose that item j is the only DIF item among J items, then the ICC for item j is:

$$P(Y_{ij} = 1) = \frac{e^{\beta_j + \alpha_j \theta_i + \gamma_j x_i}}{1 + e^{\beta_j + \alpha_j \theta_i + \gamma_j x_i}} \tag{2.25}$$

which takes a 2PL model framework with γ_i being the parameter characterising the group effect on the ICC. To ensure identifiability, some items are assumed to be DIF-free. These items keep the usual 2PL form by setting $\gamma_j = 0$ (Chen *et al.*, 2021). The latent trait distribution can be modelled by setting θ_i given x_i to follow a normal distribution $N(\boldsymbol{\beta}'\mathbf{x}_i, 1)$ (Chen *et al.*, 2021).

Non-uniform DIF occurs when there is an interaction between DIF and ability (Gamerman *et al.*, 2017) as illustrated by the MIMIC model in Figure 2.5. Chen *et al.* (2021) gives the example of a reading comprehension item in a language test that could show non-uniform DIF between the male and female groups if the item refers to a topic that is typically more familiar to one of the genders. The interaction between the construct and the covariate is indicated in the ICCs as a difference in discrimination as follows.

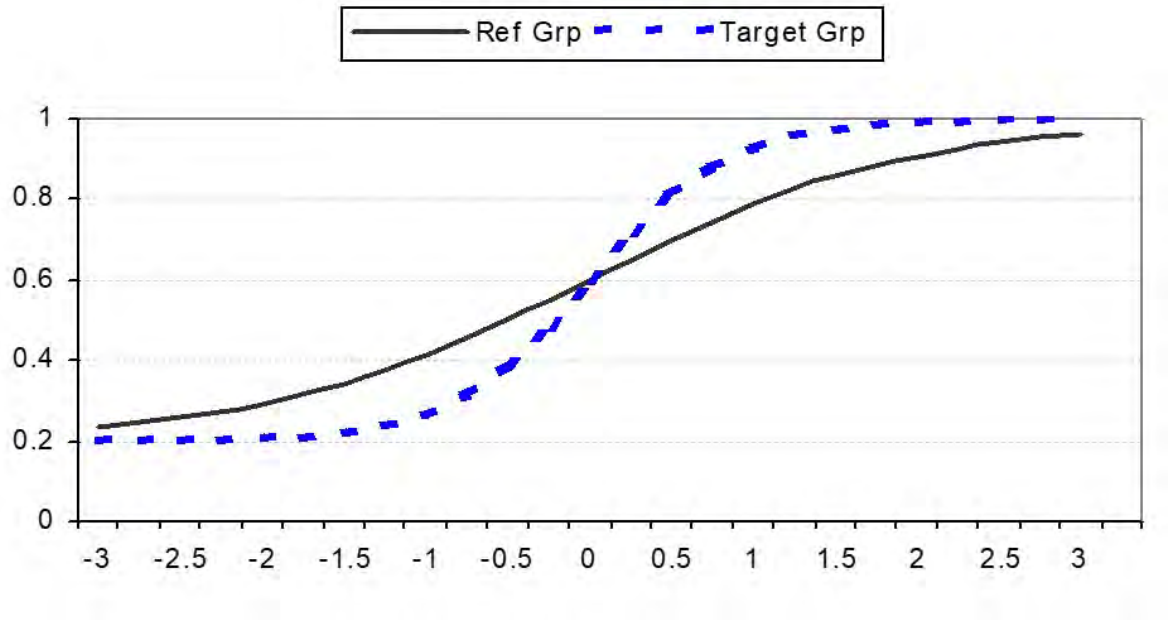


Figure 2.7: ICCs of an item presenting non-uniform DIF (Kanjee, 2010).

In Figure 2.7, it can be seen that the ICCs for the two groups intersect. Before the crossover point, the item favours the reference group, and after the ICCs cross, the item starts to favour the focal group. Because of this, non-uniform DIF can cancel itself out, such that the item would show no net DIF. This is likely to occur when the two ICCs for item j have similar β_j parameters and different α_j parameters (Abedlazeez, 2010).

2.7 Assessing Model-Data Fit

One of the major elements of consideration in measurement theory is that test scores can provide meaningful inferences, which requires the test data and theoretical framework to be well fit (Sheng and Wikle, 2008). This forms the basis for legitimizing any application of modern IRT models. However, it is well recognised by measurement experts that none of the theoretical models represent the complex reality perfectly (Van der Linden and Hambleton, 1997). As Box (1976) famously stated: "all models are wrong, some are useful". Models provide only a simplified approximation of the real world (Burnham and Anderson, 2002). Therefore, when applying IRT in the context of education, it is important to select a model that provides the most complete description of the data (Sheng and Wikle, 2008) in terms of both the model-data fit and the theoretical framework.

2.7.1 Likelihood-ratio Test

When comparing two competing IRT models for model-data fit, the likelihood-ratio chi-square statistic is usually used. This statistic is calculated by multiplying the likelihood-ratio statistic by -2, and approximates a chi-square distribution with degrees of freedom (df) equal to the difference between the number of parameters between the tested models (Cohen and Kang, 2007, Immekus *et al.*, 2019). This approximation allows for statistically significant differences between the models to be easily measured. For models that do not differ significantly, the more restrictive model (with fewer parameters) would be selected based on model parsimony. However, if models do differ significantly, then the less restrictive model (with more parameters) would be a better fit (Immekus *et al.*, 2019).

As discussed earlier, large sample sizes tend to be problematic for significance tests based on chi-square statistics. This is because the larger sample size increases the power so much that even small differences are detected as a significant misfit between the model and data. As a result, chi-square tests are almost redundant for large samples as the outcome is already known (Martin-Löf, 1974). In these cases, the log likelihood values are compared and models with values closer to zero indicating a better fit.

Generally, the log likelihood statistic is used in conjunction with Akaike information criterion (AIC), Bayesian information criterion (BIC), and root mean squared error of approximation (RMSEA), as discussed below.

2.7.2 Aitke Information Criterion

The AIC, developed by Akaike (1974), is used in model selection as it compares the estimated model with the data to indicate the amount of information lost due to estimation of a model. Since AIC provides an estimate of prediction error, the lowest value among a set of models being compared is indicative of the best fit. AIC is calculated as

$$\text{AIC} = 2k - 2\ln(L), \tag{2.26}$$

where k is the number of estimated parameters in the model and L is the maximum likelihood function of the model. The $2k$ in the equation acts as a penalty for overparameterization. Due to the lack of penalisation for sample size, there can be inconsistencies in the performance of AIC and it tends towards overfitting.

Since the AIC value for a model is arbitrary unless compared with other models' AIC values, Burnham and Anderson (2004) introduced a method of rescaling AIC values from a set of models to allow for a simpler interpretation. They defined Δ_{AIC} as

$$\Delta_{AIC} = AIC - AIC_{\min}, \tag{2.27}$$

where AIC_{min} is the minimum AIC value of the best model in terms of AIC, and Δ_{AIC} represents the loss of information caused by using the given model instead of the model with the minimum AIC. Burnham and Anderson (2004) gave rough guidelines for the interpretation of Δ_{AIC} values in terms of the relative merits of the models as follows: the models with $\Delta_{AIC} \leq 2$ have substantial support in comparison to the other models; when $4 \leq \Delta_{AIC} \leq 7$ the model has considerably less support; and models with a large $\Delta_{AIC} > 10$ have essentially no support. Naturally, the best model will have $\Delta_{AIC} = 0$.

2.7.3 Bayesian Information Criterion

The BIC was introduced by Schwarz (1978) as another model selection criterion that, like AIC, depends on the maximum likelihood function. The BIC is designed for dimension estimation and resolves the AIC's issue of overfitting models by introducing a larger penalty term that depends on the sample size. The BIC is defined as

$$\text{BIC} = k \ln(n) - 2 \ln(L), \tag{2.28}$$

where k is the number of parameters to be estimated, n is the sample size, and L is the maximum likelihood function. The BIC is interpreted the same way as the AIC where a lower value suggests better fit.

2.7.4 Root Mean Squared Error of Approximation

RMSEA (Maydeu-Olivares, 2013) is borrowed from structural equation modelling, but it is not directly generalizable to IRT. Thus, the RMSEA values merely offer a general framework and should be interpreted alongside more established measures (Immekus *et al.*, 2019) such as those previously discussed.

RMSEA is calculated using the X^2 statistic and df of the model, and N is the sample size.

$$RMSEA = \sqrt{\frac{x^2 - \text{df}}{df(-1)}} \tag{2.29}$$

where values below 0.08 and 0.05 are used to identify models that provide adequate and excellent fit respectively (Browne and Cudeck, 1993).

Chapter 3

Method

3.1 The Assessment Tool

The 2014 ANA was administered by the DBE countrywide in Literacy and Mathematics with learners in Grades 1-6 and 9 in September 2014. Both public and state-funded independent schools took part in the assessment (DBE, 2014a). DBE (2014b) reported that of the 7 376 334 learners who were registered for the 2014 ANAs, 896 939 were in grade 6. The test instrument and memorandum were retrieved from the DBE website and attached as Appendix C for ease of access. The 43 dichotomous items which were included in the analysis are highlighted on both documents for clarity.

3.1.1 Sub-domains

According to DBE (2014a), knowledge and skills from five sub-domains within the domain of Mathematics were measured by the assessment tool. These five sub-domains correspond to those prescribed in the National Curriculum Statement (NCS). The sub-domains of the test are an important consideration because construct validity of multidimensional tools can be investigated using IRT by comparing the sub-domains with the dimensions of the

MIRT model. The following five sub-domains with associated skills were designed to be assessed in the Grade 6 Mathematics test.

1. Numbers, Operations & Relationships:
    - Knowledge of place value(s) of numbers and expanded notations;
    - Knowledge of the technique of rounding off given numbers to the specified nearest number;
    - Ability to work with Operations (Addition, Subtraction, Multiplication and Division);
    - Solving money problems in given financial contexts;
    - Knowledge of and ability to work with properties of numbers;
    - Knowledge of and ability to do calculations based on ratio and rate; and,
    - Ability to solve problems involving common fractions, decimal fractions and percentages ability to identify and distinguish between multiples and factors of numbers.
2. Patterns, Functions & Algebra: Knowledge and skills that learners had to demonstrate in this area included:
    - Ability to write number sentences;
    - Demonstrating ability to solve problems that require knowledge of 'input and output values', and
    - Ability to identify Numeric and Geometric patterns and relationships.
3. Space & Shape:
    - Properties of 3-D objects;
    - Ability to locate position of objects in 2-D and 3-D space;
    - Skills in viewing and transformation of objects;
    - Recognising symmetry in shapes, and

- Solving non-routine problems.

4. Measurement:

   - Calculating quantities such as mass, length, perimeter, temperature and volume of objects;
   - Calculate time, and
   - Express measurement in different forms.

5. Data Handling:

   - Ability to read and analyse pie-charts, and
   - Calculate median and mode of a given set of numbers.

3.1.2 Cognitive Levels

In an unpublished document including the assessment specifications, Kanjee (2016) associated each item with one of four cognitive levels: Knowledge (K); Routine procedures (R); Complex procedures (C), and Problem-solving (P). The assessment specifications (Kanjee, 2016) including sub-domains, skills and cognitive levels associated with all polytomous and dichotomous items can be found in Appendix B. It should be noted that, for the purpose of the study, the skills and subdomains were allocated according to the official (published) document by DBE (2014a), and so the diction used to describe the skills in this study differs slightly from the attached assessment specifications by (Kanjee, 2016). Although there was alignment between the assessment specifications and the table of specifications used in the study, only the cognitive levels were directly retrieved from the assessment specifications. The table of specifications, including only items involved in the analyses, is provided in Table 3.1.

3.1.3 Table of Specifications

Table 3.1: Table of specifications for the dichotomous test items

Item	Content Area	Skill	Cognitive Level
1.1	Numbers, Operations & Relationships	knowledge of place value(s) of numbers	K
1.2	Numbers, Operations & Relationships	ability to work with properties of numbers	K
1.3	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	K
1.4	Numbers, Operations & Relationships	identify and distinguish between multiples and factors of numbers	K
1.5	Patterns, Functions & Algebra	identify Numeric and Geometric patterns and relationships	R
1.6	Patterns, Functions & Algebra	solve problems that require knowledge of 'input and output values'	C
1.7	Space & Shape	recognising symmetry in shapes,	K
1.8	Space & Shape	locate position of objects in 2-D and 3-D space	K
1.9	Data Handling	calculate the median of a given set of numbers	K
1.10	Measurement	calculating quantities such as „, temperature	R
2	Numbers, Operations & Relationships	knowledge of expanded notations	R
3	Numbers, Operations & Relationships	rounding off given numbers to the specified nearest number	K
5	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	P
6	Numbers, Operations & Relationships	ability to work with properties of numbers	P
7	Numbers, Operations & Relationships	identify and distinguish between multiples and factors of numbers	R
9	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	R
10.1	Numbers, operations & relationships	common fractions, decimal fractions & percentages	K
10.2	Numbers, operations & relationships	common fractions, decimal fractions & percentages	K
10.3	Numbers, operations & relationships	common fractions, decimal fractions & percentages	K
11	Numbers, Operations & Relationships	calculations based on ratio and rate	R
13	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	R
15.1	Patterns, Functions & Algebra	problems that require knowledge of 'input and output values'	R
15.2	Patterns, Functions & Algebra	problems that require knowledge of 'input and output values'	R
16	Patterns, Functions & Algebra	identify Numeric and Geometric patterns and relationships	C
17.1	Space & Shape	identifying types of angles	K
17.2	Space & Shape	identifying types of angles	K
18.1	Space & Shape	identifying 2-D shapes	K
18.2	Space & Shape	identifying 2-D shapes	K
18.3	Space & Shape	identifying 2-D shapes	K
19.1	Space & Shape	distinguishing the properties of polygons	K
19.2	Space & Shape	distinguishing the properties of polygons	K
22.1	Measurement	calculate time	P
22.2	Measurement	calculate time	P
23	Measurement	do conversions in expressing measurement in different forms	R
24	Measurement	do conversions in expressing measurement in different forms	R
25.1	Measurement	read the mass on the electronic scale	K
25.2	Measurement	do conversions in expressing measurement in different forms	R
26.1	Data Handling	read and analyse piecharts	R
26.2	Data Handling	read and analyse piecharts	R
26.3	Data Handling	read and analyse piecharts	R
26.5	Data Handling	read and analyse piecharts	R
27	Data Handling	calculate the mode of a given set of numbers	K
28	Patterns, Functions & Algebra	find a pattern (non-routine problem)	P

3.2 Data

The data was collected by the DBE in 2014 and was accessed by the authors for the purpose of this research. The DBE (2014b) indicated that 896 939 grade 6 learners from 17 326 schools were registered for the 2014 ANAs nationally. The dataset used in this study comprised 21 123 observations from 965 South African public schools.

3.2.1 Sampling Method

In 2013, randomly selected samples of Grade 6 learner scripts were utilised for the report (DBE, 2013a). The writers of the 2014 report considered the 2014 sample typical because it was representative of the population being studied and included schools that had performed poorly in the 2013 ANAs (DBE, 2014a). In 2014, all schools were identified from a sampling frame that had been supplied by the Department of Education. The data that was used in compiling the 2014 report was obtained from marked scripts collected from representative samples of schools and learners from all nine provinces that participated in Verification ANA in 2014 (DBE, 2014a). DBE (2014a) claims that the information from the 2013 ANA assisted with ensuring that the 2014 sample was typical and representative of the population. This study makes the assumption that the data used in this is typical representative of the population of grade 6 learners in South Africa in 2014. However, the author recognises that the lack of evidence to support this claim is a limitation of the research.

3.2.2 Data Structure

Each observation included the examinee's scored response to each test item. In addition, each observation included the examinee's responses to the ten multiple-choice items where answers A, B, C, and D were encoded as 1, 2, 3, and 4 respectively. Furthermore, information about the examinee's context was recorded, such as his/her Quintile (1; 2; 3; 4; 5), Geographic Type (Rural; Urban), Province (Western Cape; Northern Cape;

Free State; Eastern Cape; KwaZulu Natal; Mpumalanga; Gauteng; North West), Gender (Male; Female), and School Identification Number (a unique number assigned to each school).

3.2.3 Data Cleaning

The encoded answers to the multiple-choice items provided redundant information as the scores for those items were already recorded as scored responses with the rest of the test items. Therefore, this information was removed from the dataset. The dataset seemed to have a large proportion of missing values. However, upon closer inspection of the data, it was found that many of the missing values were from empty observations. Generally for large scale assessments, missing values are treated either as missing data or as incorrect responses, but the underlying assumptions of both these approaches can be problematic (Rose *et al.*, 2010). By treating these omissions as missing data, it is assumed that the reason (e.g. unknown response) for nonresponses can be ignored. On the other hand, the treatment of omissions as decidedly incorrect responses assumes that the correct response is unknown by the examinee, irrespective of their ability.

A combination of methods for treating missing values was applied to this study. Since it would be unreasonable to assume that completely empty observations were purely due to all the answers being unknown by these examinees, these were removed from the dataset. After the removal of all empty observations, there were 8 971 observations left to be used in the analysis. Therefore, there were N = 8971 examinees, such that $i = 1, 2, ..., 8971$. The remaining missing values were from non-empty observations. Since it is likely that these missing values were a result of the learners' inability to answer correctly, these nonresponses were scored as incorrect.

It should be noted that these missing values provide a limitation to the study and that the lack of involvement of the author in the data collection process prevented the use of more rigorous approaches to dealing with the missing values.

Due to the focus of this study on logistic IRT and MIRT, only the dichotomous items

(scored as correct or incorrect) were included in the analysis. Thus, 43 of 57 items were evaluated. It is important to note that when the author refers to the test, it is in reference to the the 43 items used in the study, thus, J = 43, and hence $j = 1, 2, ..., 43$. Although this could limit the generalizability of the results, all subdomains and cognitive levels were represented among the included items. Therefore, the model should still be able to identify the appropriate dimensions.

3.3 Bayesian Estimation

The Bayesian approach, established by Bayes (1763), combines prior beliefs or knowledge about the unknown parameters with data as it becomes available. Bayesian statistics applied to IRT (Bürkner, 2019) suggests that the posterior distribution $P(\theta, \xi \mid y)$ of the person (θ) and item parameters (ξ) is estimated given the observed data y. The point of interest may be the values that are computed based on the posterior distribution or the posterior distribution itself. The posterior distribution is defined based on the prior distribution $P(\theta, \xi)$, the likelihood function $P(y \mid \theta, \xi)$, and the marginal likelihood $P(y)$ (which serves as a normalising constant to ensure the posterior distribution is valid) as follows.

$$P(\theta, \xi \mid y) = \frac{P(y \mid \theta, \xi)P(\theta, \xi)}{P(y)}. \tag{3.1}$$

In the multidimensional context, the likelihood function is determined as follows (Chalmers, 2012).

Let $P\left(y_{ij} = k \mid \boldsymbol{\alpha}_j, d_j, \boldsymbol{\theta}_i\right) = \Phi\left(y_{ij} = k \mid \boldsymbol{\alpha}_j, d_j, \boldsymbol{\theta}_i\right)$ for k = {0, 1}.

Then, by letting the collection of all item parameters be $\boldsymbol{\psi}$, we can be more concise, putting the data in indicator form as

$$\chi\left(y_{ij}\right) = \begin{cases} 1, & \text{if } y_{ij} = k \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

The conditional distribution for the ith n x 1 response pattern vector $\boldsymbol{y}_i$ is

$$L_\ell\left(\mathbf{y}_i \mid \boldsymbol{\Psi}, \boldsymbol{\theta}\right) = \prod_{j=1}^{n} \prod_{k=0}^{1} \Phi\left(y_{ij} = k \mid \boldsymbol{\Psi}, \boldsymbol{\theta}_i\right)^{\chi(x_{ij})} \tag{3.3}$$

Assuming a distributional form $g(\boldsymbol{\theta})$, which is usually a multivariate normal distribution (Chalmers, 2012), the marginal distribution is

$$P_\ell(\boldsymbol{\Psi} \mid \mathbf{y}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} L_\ell\left(\mathbf{y}_i \mid \boldsymbol{\Psi}, \boldsymbol{\theta}\right) g(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{3.4}$$

where there are m integrals to be evaluated. Let $\boldsymbol{Y}$ denote the N x J matrix of observed data. The equation for the data likelihood function is

$$L(\boldsymbol{\Psi} \mid \mathbf{Y}) = \prod_{i=1}^{N} \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} L_\ell\left(\mathbf{y}_i \mid \boldsymbol{\Psi}, \boldsymbol{\theta}\right) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \right]. \tag{3.5}$$

The posterior distribution is, therefore

$$L(\boldsymbol{\Psi} \mid \mathbf{X}, \boldsymbol{\theta}) = \prod_{i=1}^{N} L_\ell\left(\mathbf{x}_i \mid \boldsymbol{\Psi}, \boldsymbol{\theta}\right) g(\boldsymbol{\theta} \mid \mu, \boldsymbol{\Sigma}). \tag{3.6}$$

3.3.1 Parameter Estimation

Chalmers (2012) allows for the estimation of parameters for exploratory and confirmatory MIRT models using the mirt package in R. The mirt package makes use of the following two estimation methods: the fixed quadrature Expectation-Maximization (EM) method for exploratory (Bock *et al.*, 1988) and bifactor (Gibbons and Hedeker, 1992) models, and the Metropolis-Hastings Robbins-Monro method for exploratory and confirmatory polytomous models (Cai, 2010). The EM method for exploratory models can be used in combination with quasi-Monte Carlo (QMC) or Monte Carlo (MC) methods for models with higher dimensionality.

Since the models in the study were exploratory and involved dichotomous rather than polytomoous items, the EM methods were utilised. For low to moderate factor solutions the EM algoritm is appropriate, as long as the number of quadratures per dimension decreases as the number of factors increases (Chalmers, 2012). Usually, the EM method works well for up to three factors, but when three or more dimensions are involved, the QMCEM and MCEM procedures are more applicable. The MCEM method is the default method for estimating the parameers of high-dimension exploratory models in the mirt package and was selected for the multidimensional models in this study (Chalmers, 2012). EM was selected as the method for parameter estimation in the unidimensional model.

3.4 Model Selection

The majority of the test items (43 out of 57 items) had a maximum score of one and therefore had a binary response pattern where each response was either correct or incorrect. Binary response patterns of the dichotomous items allow for logistic models to be utilised. As opposed to models for polytomous items, these models are less complex. This is advantageous, especially when multiple dimensions and covariates are involved. Therefore, a dichotomous model involving only the 43 binary items was selected as an appropriate model for this study.

The best-fitting models were determined by the model-data fit statistics in conjunction with the theoretical framework. Although chi-square tests are considered best practice for model comparisons, the samples were too large for this method to provide meaningful information (Martin-Löf, 1974). As a result, mode-data fit statistics which are less influenced by sample size were utilised. The following model-data fit statistics were employed: BIC, AIC, log likelihood and RMSEA. These statistics are detailed in Section 2.7.

Once the data cleaning process was complete, the following approach was taken to conduct the data analysis.

3.5 Assessing Construct Validity

3.5.1 Unidimensional Analysis

Four unidimensional models were developed (1PL, 2PL, 3PL, and 4PL) using the mirt package (Chalmers, 2012) in R. The best-fitting model was identified based on the model-data fit statistics with consideration of the functional form of the items. The item parameter estimates of the best-fitting unidimensional model were then extracted and categorised according to the specifications found in Section 2.2.2. As suggested by Considine *et al.* (2005), the difficulty and discrimination parameters were utilised as evidence for assessing construct validity.

3.5.2 Multidimensional Analysis

Since the outline of the assessment indicated multiple sub-domains being tested by the assessment, it was of interest to assess the relationship between the sub-domains defined by the DBE (2014a) and the dimensions defined by the best-fitting MIRT model. Exploratory MIRT models were fit to the data to identify if the table of specification could be replicated by the MIRT model to validate the test. A higher degree of similarity is associated with a tool producing inferences that fulfil the requirements of construct validity.

Three-, four- and five-dimensional exploratory models were developed using the mirt package in R and the item parameters were estimated by MCEM methods. For each number of dimensions, an additional model was developed which had a multilevel structure with random effects for schools to account for the variability attributed to the differences between schools. All six models were compared using the model-data fit statistics described in Section 2.7. The best-fitting MIRT model was used as evidence for determining construct validity.

3.5.3 Assessing Item Bias

3.5.3.1 Identifying Significant Covariates

Since test and item bias affect the validity, exploratory analyses were also conducted to identify any items that functioned differentially for different groups of examinees. First, all possible combinations of covariates were included in the model, and the model fit statistics were used to identify significant covariates for further analysis. DIF analyses helped provide information on whether the 2014 performance differences between the groups for the significant covariates were due to true difference in ability or a result of item biases.

3.5.3.2 Identifying Significant Items

Once significant covariates were identified, the IRT methods for identifying DIF in IRT models, outlined in Section 2.6.2.3, were utilised to identify DIF items based on the best-fitting IRT model. A likelihood-ratio test of DIF was conducted to identify significant differences in the item parameter estimates between groups, which would indicate that the items present DIF. When conducting a likelihood-ratio test for DIF based on the multipleGroup function in the mirt package, it must be assumed that some items are DIF-free to fix the latent variable distribution and ensure identifiability of DIF items (Chen *et al.*, 2021). Usually the anchor items which have previously been established as DIF-free would be selected for this purpose. Since the anchor items for the test were confidential (DBE, 2014b), another method of selecting DIF-free items was utilised. Lopez Rivas *et al.* (2009) and González-Betanzos and Abad (2012) suggest that items with higher discrimination are better suited as anchor items. Therefore, the items with discrimination parameters greater than the 75th percentile of discrimination ($\alpha = 1.44$) were selected as DIF-free items. Hence, items 1.1, 1.4, 1.5, 1.6, 1.10, 6, 16, 18.1, 18.3, 22.2, 24, 25.1, and 26.1 were selected as the anchor items. Although an evidence-based approach was taken for the selection of the anchor items, the lack of information on the official anchor items provided a limitation to the study.

3.5.3.3 Investigating Significant Items

The ICCs for the items presenting significant DIF were plotted for each group to identify for which groups the items functioned differentially. The ICCs were inspected visually and the notable features of the ICCs were discussed.

All analyses were conducted using the mirt package in R (Chalmers, 2012).

Chapter 4

Results

4.1 Data Cleaning

Once empty observations had been removed from the original dataset of 21 123 observations, there were 8 971 non-empty observations left for analysis. Once the polytomous items were removed from the dataset, there were 43 dichotomous items which were analysed for the purpose of this study.

The mean score for the test comprised of 43 items was 43.03%. The proportion of correct answers is tabulated for each item in Table A.1 of Appendix A.

4.2 Unidimensional Model

Four unidimensional logistic models were fit to the data and assessed using approximate fit statistics in the context of the theoretical framework of the data. The four models were the 1PL, 2PL, 3Pl and 4PL models.

4.2.1 Model Selection

4.2.1.1 Model-fit Statistics

Table 4.1: Assessing model fit of the Rasch, 2Pl, 3PL, and 4Pl models.

Model	AIC	BIC	RMSEA	logLik	p-value
1PL	409386.9	409699.4	0.053	-204649.5	0
2PL	402844.7	403455.5	0.046	-201336.4	0
3PL	400630.4	401546.5	0.032	-200186.2	0
4PL	399865.1	401086.6	0.030	-199760.5	0

4.2.1.2 Selection: the Two Parameter Logistic Model

As shown in Table 4.1, the AIC and BIC values decrease as the number of parameters in the model increases. The log-likelihood value increases as the number of parameters increases. Therefore, by considering only the model-fit statistics, it is clear that the 4PL model is the best fitting model. It should be noted that there was a larger improvement in model fit between the 1PL model and the 2Pl model than between the 2PL and 3PL or 3PL and 4PL models. Based on the RMSEA interpretation from Maydeu-Olivares (2013), the 2PL, 3PL and 4PL models were considered to have an excellent fit (RMSEA ≤ 0.05), with the 4PL model (RMSEA $= 0.030$) having the best fit.

The item types present in the data were majority (33) open-ended with only ten multiple-choice items. The 3PL and 4PL models fit the functional form of the multiple-choice items; however, the guessing parameter is redundant for items that have an open-ended functional form. Therefore, the 2PL model was best matched to the functional form of the majority of the items.

Due to the 2PL model being the simplest model to have an excellent fit while matching the theoretical framework of the data, it was selected to be the most suitable, parsimonious model for the purpose of this paper. The fit for each item was assessed and results in Table A.2 in Appendix A indicate that all items had an excellent fit.

4.2.2 2PL Model Parameters

Once the 2 parameter logistic model was selected, the difficulty (β_j) and discrimination (α_j) parameters were extracted for all $J = 43$ items.

It was found that the item estimates of discrimination ranged from 0.358 to 3.102, with a mean of 1.248 ($\pm$ 0.539). The difficulty parameters, with a mean of -0.434 ($\pm$ 1.227) lay within the range of -3.093 to 2.953. It is important to note that the mirt package in R produces difficulty estimates such that a higher difficulty parameter estimate correlates with a higher probability of a correct answer, and therefore, indicates an easier item. This is in contrast to the conventional scale found in literature where a higher difficulty parameter indicates a more difficult item. The parameters are presented as they were produced and the interpretation of the difficulty parameters was adjusted accordingly.

These parameters are presented in Table 4.2.

Table 4.2: Parameters of the unidimensional 2PL model

Item	α	β
1.1	0.879	-1.412
1.2	1.085	-1.510
1.3	1.149	-0.442
1.4	0.813	-0.886
1.5	0.915	0.557
1.6	0.567	-0.515
1.7	1.247	-0.211
1.8	1.062	-0.248
1.9	1.069	0.394
1.10	0.599	0.145
2	1.365	0.329
3	1.777	0.239
5	1.679	-0.607
6	0.492	0.151
7	1.389	-1.991
9	1.619	-0.316
10.1	3.102	-0.507
10.2	1.968	0.366
10.3	3.085	-0.642
11	1.113	0.303
13	1.156	-0.216
15.1	1.339	0.518
15.2	1.491	0.328
16	0.830	-0.393
17.1	1.275	-0.629
17.2	1.552	-0.047
18.1	0.982	0.249
18.2	1.077	0.181
18.3	0.713	-1.000
19.1	1.510	-3.019
19.2	1.487	-1.579
22.1	1.365	-3.093
22.2	0.978	-2.782
23	1.340	-1.843
24	0.358	-0.667
25.1	0.819	1.268
25.2	1.221	-2.502
26.1	0.907	0.582
26.2	1.395	2.953
26.3	1.104	1.060
26.5	1.531	-0.147
27	1.075	1.231
28	1.156	-2.290

The item characteristic curves (ICCs), also known as item probability functions, and item information curves (IICs) for all the test items are plotted together as shown in Figure 4.1. The inclusion of all the items on the same graph helps visualise the spread of the characteristics.

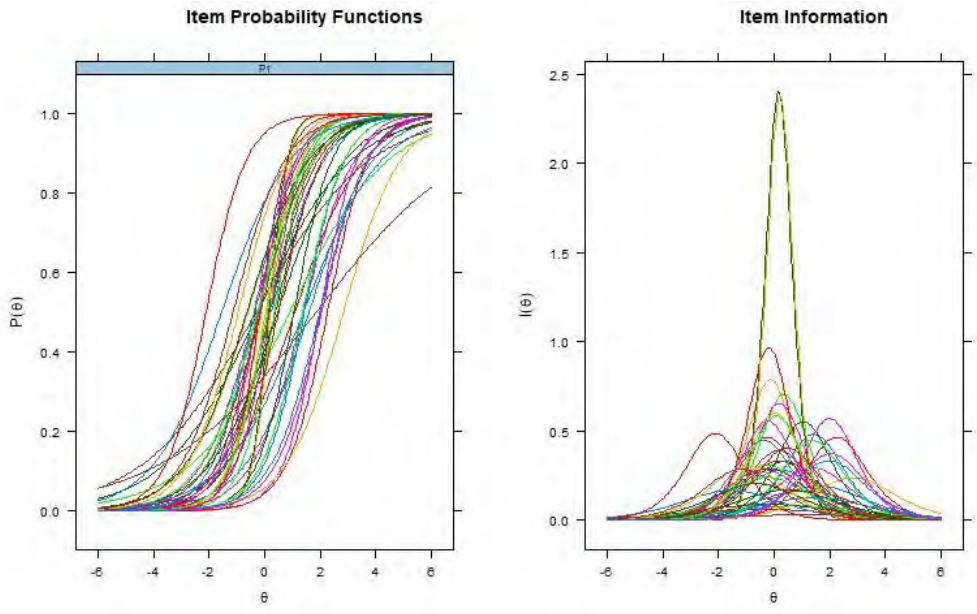


Figure 4.1: ICCs and IICs of all items plotted together

The ICCs graph indicates that there is a wide range of item difficulties present in the test, however, the curves are quite concentrated around $\theta = 0$. The IICs give a better graphical representation of the discrimination because they show how much information each item provides, as well as the range on the ability continuum for which the item provides this information. Here, it is confirmed that a substantial proportion of items provide information about examinees with ability levels close to zero. Furthermore, these items close to $\theta = 0$ vary in terms of how much information they provide about the examinees.

Although these graphs give a good indication of the parameters of the test items as a whole, they give little information on the individual item characteristics.

The item characteristic curves (ICCs) and item information curves (IICs) for all the test items are plotted individually below.

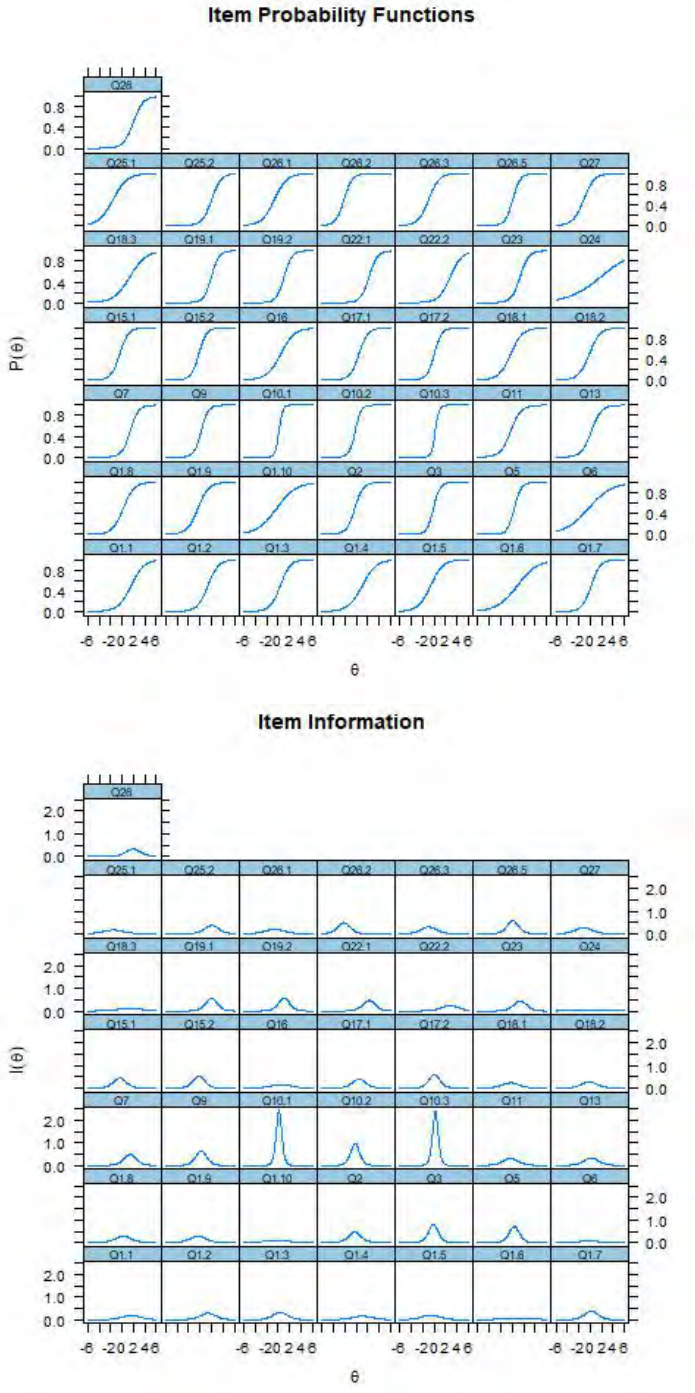


Figure 4.2: ICCs and IICs of all items plotted individually

Each individual curve in Figure 4.2 has a title above it to show which item the curve is for. This representation allows for the item characteristics and information to be linked to their respective items. For example, the item probability function with the most gradual slope (least discriminating) from 4.1 would be item 24 based on inspection of the ICCs

in Figure 4.2. In addition, the two items with largest information curves seen in Figure 4.1, can be identified as items 10.1 and 10.3 based on the individual information curves. These deductions can be confirmed by the parameter estimations presented in Table 4.2.

Although it is helpful to identify the properties of the test items based on their curves, it is also useful to be able to categorise the items distinctly according to their properties. Therefore, the items were categorised according to their parameters and interpreted as follows.

4.2.2.1 Difficulty Parameter

Based on the categorisation of difficulty by Bichi and Talib (2018), the percentage of items in each difficulty level were as follows.

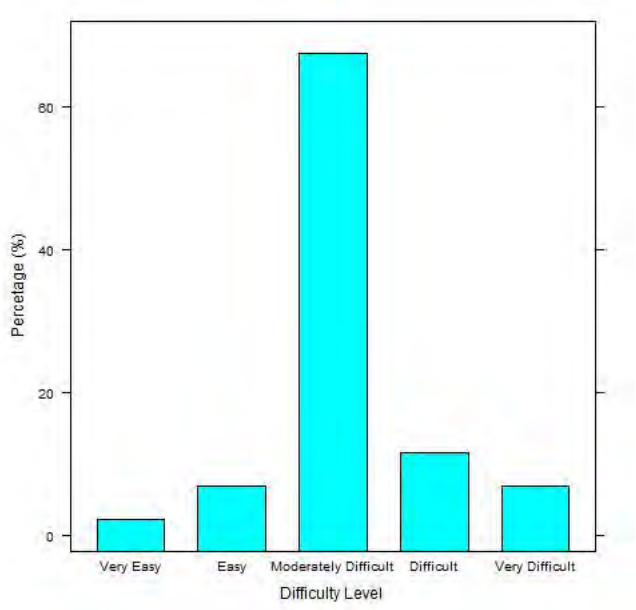


Figure 4.3: The percentage of test items in each difficulty level

The level of difficulty based on Figure 4.3 illustrates that, according to the categories specified in Table 2.1, most test items (67.44%) fit in the Moderately Difficult category. The Difficult and Very Difficult categories consisted of five items (11.63%) each. Three

items (6.98%) were considered easy while the Very Easy category consisted of only one item (2.33%).

To illustrate the range of difficulty values present in the test, the item characteristic curves (ICCs) and item information curves (IICs) were plotted for four items in different difficulty categories with comparable discrimination parameters.

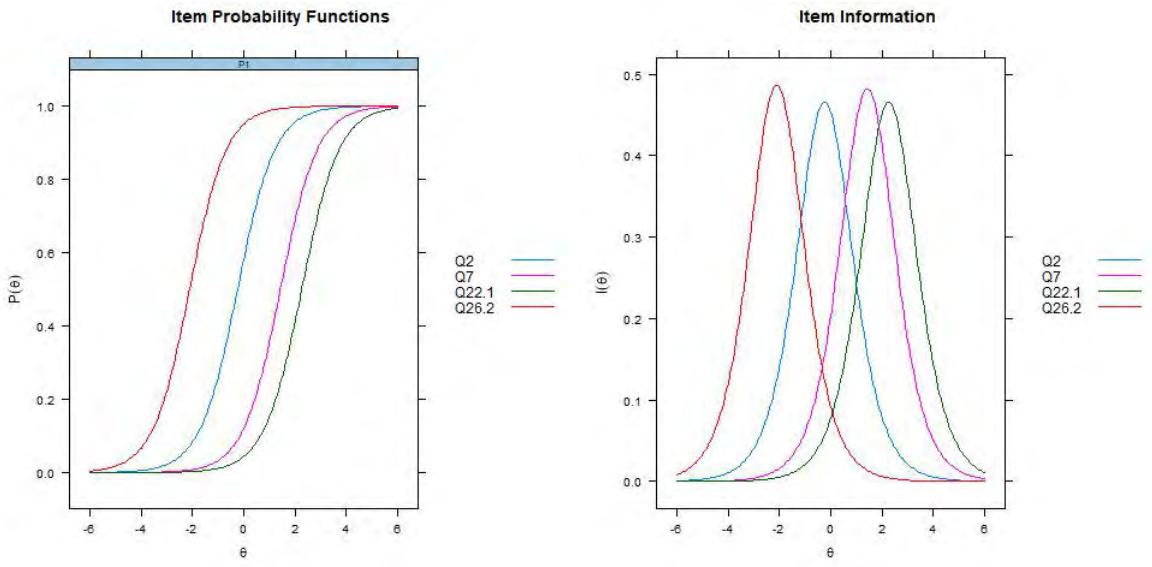


Figure 4.4: Graphical comparison of items of varying difficulty

To ensure comparability, all items selected were in the Good item quality category and lay within the narrow range of $1.365 \leq \alpha \leq 1.398$. This is illustrated by their similar item information curve heights. The difficulties selected for this illustration were very easy (Q26.2), moderately difficult (Q2), difficult (Q7), and very difficult (Q22.1).

4.2.2.2 Discrimination Parameter

In order to describe the item quality, the items were categorised according to specifications given by Adedoyin and Mokobi (2013) as well as the categorisation by Bichi and Talib (2018) presented in Table 2.2. As seen in Figure 4.5, more than half (55.81%) of test items were functioning moderately. A substantial proportion (25.58%) were considered

good quality. The marginal and satisfactory categories comprised only four items each (9.30%).

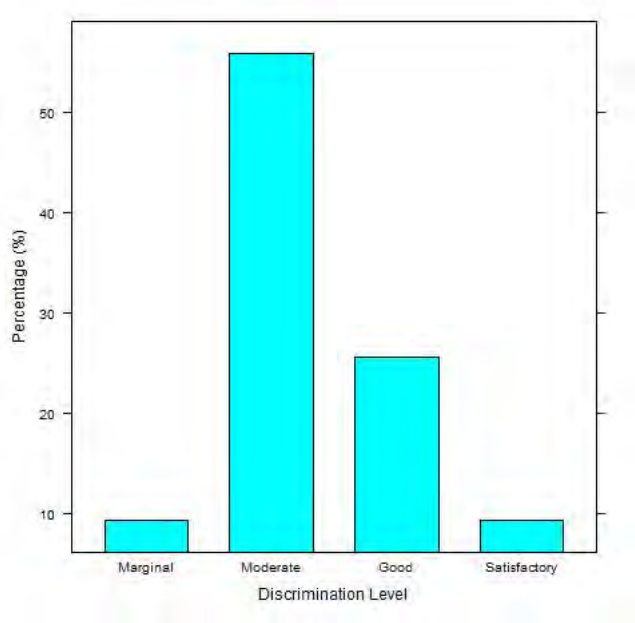


Figure 4.5: The percentage of items in each discrimination level defined in Bichi and Talib (2018)

A different categorization of discrimination parameters was presented by Adedoyin and Mokobi (2013), and items were also grouped according to this specification as follows.

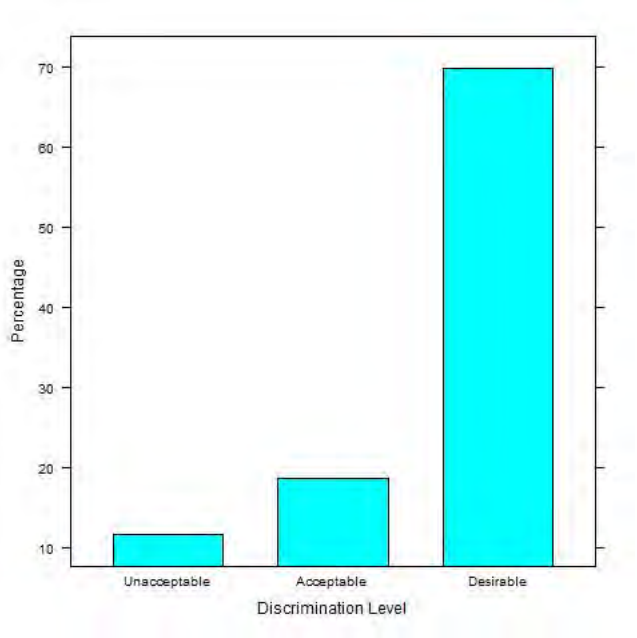


Figure 4.6: The percentage of items in each discrimination level defined in Adedoyin and Mokobi (2013)

Based on the results in Figure 4.6, the majority of items (69.77%) were of desirable quality, while 18.60% were considered acceptable and 11.63% unacceptable.

It is interesting to note that The moderate items in Figure 4.5 include items from all three categories in Figure 4.6. Therefore, the Figures can be used in combination to maximise the clarity of results.

To illustrate the range of discrimination values present in the test, the item characteristic curves (ICCs) and item information curves were plotted for four items, as shown in Figure 4.7, in different discrimination categories and the same difficulty category. The four items were all moderately difficult within the range of $-0.607 \leq \beta \leq -0.442$ to maintain comparability of the discrimination parameters. This is confirmed by their overlapping location on the ability continuum. Item 1.6 had marginal item quality while Q1.3 was considered moderate, Q5 was good and Q10.1, with the highest discrimination estimate, was satisfactory.

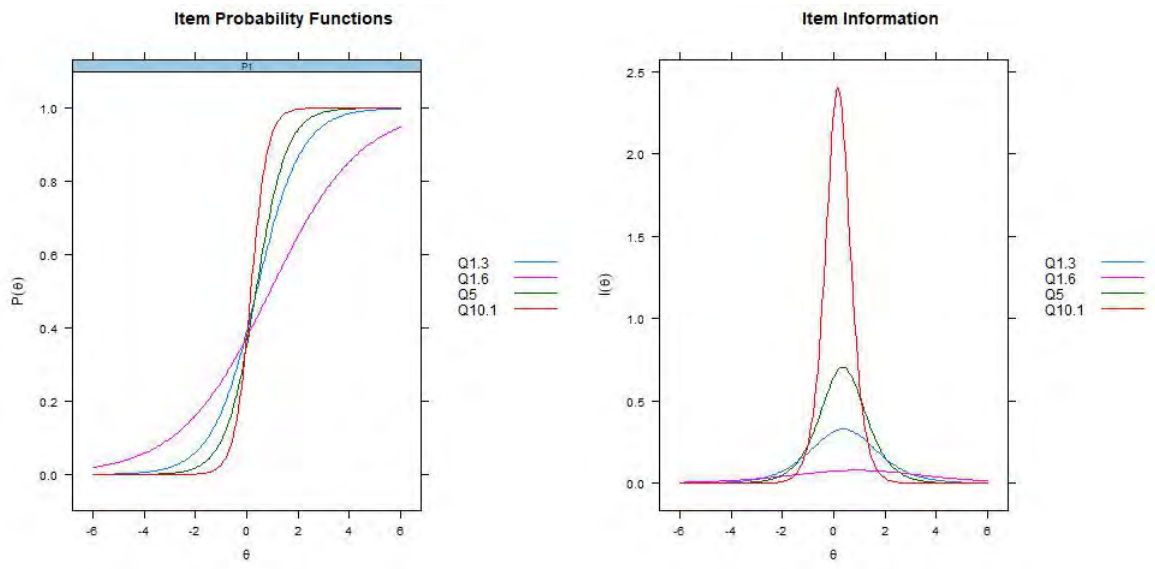


Figure 4.7: Graphical comparison of items of varying discrimination

Furthermore, the difficulty and discrimination parameters were plotted on a scatterplot in Figure 4.8 to show the relationship between the two parameters in the test items.

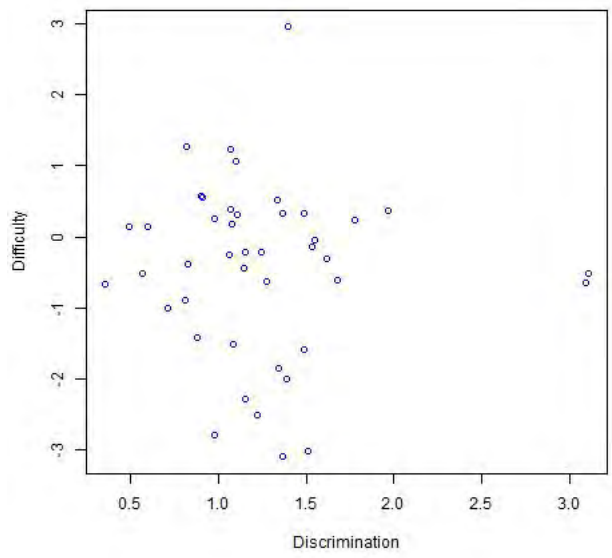


Figure 4.8: Discrimination vs difficulty parameters of the test items

Points located closer to the x-axis represent more difficult items. There is no appreciable

linear correlation between the discrimination and difficulty estimates of the items (r=-0.049). which indicates that the difficulty and quality of the items are not related.

4.2.3 Evaluating Test Specifications in terms of Difficulty and Discrimination

To identify and investigate any relationships between the subdomain, cognitive level or skill tested by an item and their difficulty or discrimination parameters, two tables were compiled: Table 4.3 lists items from easiest to most difficult while Table 4.4 orders them from least to most informative.

Table 4.3: Investigating sub-domain, skill and cognitive level in terms of item difficulty

Interpretation	Item	beta	Sub-Domain	Skill	Cognitive Level
Very Easy	26.2	2.953	Data Handling	read and analyse piecharts	R
Easy	25.1	1.268	Measurement	read the mass on the electronic scale	K
	27	1.231	Data Handling	calculate the mode of a given set of numbers	K
	26.3	1.060	Data Handling	read and analyse piecharts	R
Moderate	26.1	0.582	Data Handling	read and analyse piecharts	R
	1.5	0.557	Patterns, Functions & Algebra	identify Numeric and Geometric patterns and relationships	R
	15.1	0.518	Patterns, Functions & Algebra	problems that require knowledge of 'input and output values'	R
	1.9	0.394	Data Handling	calculate the median of a given set of numbers	K
	10.2	0.366	Numbers, operations & relationships	common fractions, decimal fractions & percentages	K
	2	0.329	Numbers, Operations & Relationships	knowledge of expanded notations	R
	15.2	0.328	Patterns, Functions & Algebra	problems that require knowledge of 'input and output values'	R
	11	0.303	Numbers, Operations & Relationships	calculations based on ratio and rate	R
	18.1	0.249	Space & Shape	identifying 2-D shapes	K
	3	0.239	Numbers, Operations & Relationships	rounding off given numbers to the specified nearest number	K
	18.2	0.181	Space & Shape	identifying 2-D shapes	K
	6	0.151	Numbers, Operations & Relationships	ability to work with properties of numbers	P
	1.10	0.145	Measurement	calculating quantities such as „, temperature	R
	17.2	-0.047	Space & Shape	identifying types of angles	K
	26.5	-0.147	Data Handling	read and analyse piecharts	R
	1.7	-0.211	Space & Shape	recognising symmetry in shapes,	K
	13	-0.216	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	R
	1.8	-0.248	Space & Shape	locate position of objects in 2-D and 3-D space	K
	9	-0.316	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	R
	16	-0.393	Patterns, Functions & Algebra	identify Numeric and Geometric patterns and relationships	C
	1.3	-0.442	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	K
	10.1	-0.507	Numbers, operations & relationships	common fractions, decimal fractions & percentages	K
	1.6	-0.515	Patterns, Functions & Algebra	solve problems that require knowledge of 'input and output values'	C
	5	-0.607	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	P
	17.1	-0.629	Space & Shape	identifying types of angles	K
	10.3	-0.642	Numbers, operations & relationships	common fractions, decimal fractions & percentages	K
	24	-0.667	Measurement	do conversions in expressing measurement in different forms	R
	1.4	-0.886	Numbers, Operations & Relationships	identify and distinguish between multiples and factors of numbers	K
	18.3	-1.000	Space & Shape	identifying 2-D shapes	K
Difficult	1.1	-1.412	Numbers, Operations & Relationships	knowledge of place value(s) of numbers	K
	1.2	-1.510	Numbers, Operations & Relationships	ability to work with properties of numbers	K
	19.2	-1.579	Space & Shape	distinguishing the properties of polygons	K
	23	-1.843	Measurement	do conversions in expressing measurement in different forms	R
	7	-1.991	Numbers, Operations & Relationships	identify and distinguish between multiples and factors of numbers	R
Very Difficult	28	-2.290	Patterns, Functions & Algebra	find a pattern (non-routine problem)	P
	25.2	-2.502	Measurement	do conversions in expressing measurement in different forms	R
	22.2	-2.782	Measurement	calculate time	P
	19.1	-3.019	Space & Shape	distinguishing the properties of polygons	K
	22.1	-3.093	Measurement	calculate time	P

Table 4.4: Investigating sub-domain and skills in terms of item discrimination power

Interpretation	Item	α	Sub-domain	Skill	Cognitive Level
Marginal	24	0,358	Measurement	do conversions in expressing measurement in different forms	R
	6	0,492	Numbers, Operations & Relationships	ability to work with properties of numbers	P
	1,5	0,567	Patterns, Functions & Algebra	identify Numeric and Geometric patterns and relationships	R
	1,9	0,599	Data Handling	calculate the median of a given set of numbers	K
Moderate	18,3	0,713	Space & Shape	identifying 2-D shapes	K
	1,3	0,813	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	K
	25,1	0,819	Measurement	read the mass on the electronic scale	K
	16	0,83	Patterns, Functions & Algebra	identify Numeric and Geometric patterns and relationships	C
	1,1	0,879	Numbers, Operations & Relationships	knowledge of place value(s) of numbers	K
	26,1	0,907	Data Handling	read and analyse piecharts	R
	1,4	0,915	Numbers, Operations & Relationships	identify and distinguish between multiples and factors of numbers	K
	22,2	0,978	Measurement	calculate time	P
	18,1	0,982	Space & Shape	identifying 2-D shapes	K
	1,7	1,062	Space & Shape	recognising symmetry in shapes,	K
	1,8	1,069	Space & Shape	locate position of objects in 2-D and 3-D space	K
	27	1,075	Data Handling	calculate the mode of a given set of numbers	K
	18,2	1,077	Space & Shape	identifying 2-D shapes	K
	1,91	1,085	Measurement	calculating quantities such as „, temperature	R
	26,3	1,104	Data Handling	read and analyse piecharts	R
	11	1,113	Numbers, Operations & Relationships	representing numbers on a number line	R
	1,2	1,149	Numbers, Operations & Relationships	ability to work with properties of numbers	K
	13	1,156	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	R
	28	1,156	Patterns, Functions & Algebra	find a pattern (non-routine problem)	P
	25,2	1,221	Measurement	do conversions in expressing measurement in different forms	R
	1,6	1,247	Patterns, Functions & Algebra	solve problems that require knowledge of 'input and output values'	C
	17,1	1,275	Space & Shape	identifying types of angles	K
	15,1	1,339	Patterns, Functions & Algebra	problems that require knowledge of 'input and output values'	R
	23	1,34	Measurement	do conversions in expressing measurement in different forms	R
Good	2	1,365	Numbers, Operations & Relationships	knowledge of expanded notations	R
	22,1	1,365	Measurement	calculate time	P
	7	1,389	Numbers, Operations & Relationships	identify and distinguish between multiples and factors of numbers	R
	26,2	1,395	Data Handling	read and analyse piecharts	R
	19,2	1,487	Space & Shape	distinguishing the properties of polygons	K
	15,2	1,491	Patterns, Functions & Algebra	problems that require knowledge of 'input and output values'	R
	19,1	1,51	Space & Shape	distinguishing the properties of polygons	K
	26,5	1,531	Data Handling	read and analyse piecharts	R
	17,2	1,552	Space & Shape	identifying types of angles	K
	9	1,619	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	R
	5	1,679	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	P
Satisfactory	3	1,777	Numbers, Operations & Relationships	rounding off given numbers to the specified nearest number	K
	10,2	1,968	Numbers, operations & relationships	common fractions, decimal fractions & percentages	K
	10,3	3,085	Numbers, operations & relationships	common fractions, decimal fractions & percentages	K
	10,1	3,102	Numbers, operations & relationships	common fractions, decimal fractions & percentages	K

4.3 Multidimensional Item Response Theory

4.3.1 Model Selection

The theoretical framework (3.1) of the test implied that there would be five dimensions representing each sub-domain present in the data. This was investigated by running three models for comparison. Since adding dimensions generally improves fit, the maximum number of dimensions was selected to match the theoretical framework. Therefore, three-, four- and five-dimensional models were assessed. Furthermore, since schools were expected to be a source of variance among examinees, another set of models with random effects for schools were run. These models are referred to as the schools models, and the models without random effects for schools were referred to as the non-schools models. The model fit statistics outlined in Section 2.7 were computed for all six models. These models were referred to as the non-schools models because there was.

4.3.1.1 Model-fit Statistics

	Dimensions	AIC	BIC	logLik	RMSEA
non-schools	3	358645.2	359929.9	-179133.6	0.043
	4	357065.0	358648.8	-178299.5	0.034*
	5	356351.0*	358233.9*	-177898.5*	0.037
schools	3	434256.2	437600.4	-216636.1	0.053
	4	433310.1	436953.4	-216119.0	0.049
	5	433166.3	437108.7	-216003.1	0.050

Table 4.5: Model-data fit statistics for the multidimensional IRT models

4.3.1.2 Selection: the 5-dimensional Non-schools Model

The 5-dimensional model that did not account for the variance between schools was found to be the best fitting model in terms of the model-data fit statistics and in terms of the theoretical framework.

The model parameters for each item are tabulated in Table 4.6.

Table 4.6: Parameters of the 5 dimensional IRT model

Item	α_1	α_2	α_3	α_4	α_5	β
1.1	0.598	1.167	0.740	1.673*	-1.686	-1.406
1.2	0.809	1.456	0.946	1.840*	-2.307	-1.530
1.3	0.747	1.046	1.160	1.724*	-0.995	-0.306
1.4	0.550	0.966	0.914	1.267*	-2.212	-0.826
1.5	0.550	0.552	0.915	1.747*	0.423	0.669
1.6	0.425	0.633	0.500	1.046*	-0.779	-0.414
1.7	0.672	1.186	1.057	1.521*	0.799	-0.128
1.8	0.624	0.910	0.849	1.467*	0.814	-0.215
1.9	0.495	1.086*	1.024	0.655	0.308	0.465
1.10	0.366	0.789	0.483	1.213*	-0.881	0.216
2	0.823	0.822	1.156	1.885*	1.274	0.414
3	1.093	1.242	1.642	1.901*	1.370	0.361
5	0.946	0.923	1.732	2.166*	0.755	-0.491
6	0.212	0.533	0.386	0.791*	0.745	0.144
7	0.755	1.155	1.164	1.825*	0.688	-1.827
9	0.958	1.229	1.366	1.566*	1.049	-0.198
10.1	9.434*	5.898	2.852	0.808	4.677	-0.869
10.2	3.252*	1.859	0.737	0.334	2.578	0.691
10.3	11.286*	6.697	3.111	1.040	5.834	-1.534
11	0.676	0.438	1.116	1.706*	1.259	0.379
13	0.667	0.595	1.309	1.604*	0.366	-0.127
15.1	1.537	-1.259	3.884*	-0.463	1.462	1.088
15.2	1.764	-1.039	4.757*	-0.371	1.206	0.939
16	0.415	0.615	0.848*	0.722	0.722	-0.351
17.1	0.559	3.142*	1.812	-4.082	2.791	-0.814
17.2	0.849	3.201	2.137	-3.760	3.722*	0.044
18.1	0.485	0.933	0.805	1.126	1.442*	0.306
18.2	0.526	0.878	0.933	1.054	1.933*	0.274
18.3	0.290	0.780	0.611	0.838*	0.805	-0.955
19.1	0.645	2.110*	1.154	1.973	0.141	-3.024
19.2	0.733	1.675	1.234	2.159*	0.865	-1.553
22.1	0.586	1.504	1.206	4.382*	-0.621	-3.222
22.2	0.307	1.221	0.825	4.118*	-0.609	-2.963
23	0.692	1.338	1.150	2.879*	0.069	-1.819
24	0.247	0.302	0.227	0.916*	0.577	-0.636
25.1	0.314	0.640	0.640	1.771	2.282*	1.440
25.2	0.606	1.299	1.037	2.524*	0.227	-2.601
26.1	0.400	0.696	0.671	2.871	3.702*	0.734
26.2	0.570	0.908	0.883	2.887	6.134*	3.536
26.3	0.436	0.838	0.934	2.383	3.695*	1.140
26.5	0.794	1.048	1.301	2.356	3.109*	-0.124
27	0.495	1.066	0.889	0.851	1.372*	1.325
28	0.564	1.153	1.031	2.825*	-0.762	-2.251

The predominant dimensions were obtained and tabulated in Table 4.7.

Table 4.7: Predominant Item dimensions as derived by the MIRT model

	Dimension					
Item	1	2	3	4	5	Predominant Dimension
1.1	0.286	0.319	0.261	0.182	-0.194	2
1.2	0.352	0.362	0.304	0.182	-0.242	2
1.3	0.340	0.272	0.390	0.179	-0.109	3
1.4	0.262	0.263	0.322	0.138	-0.255	3
1.5	0.275	0.158	0.338	0.199	0.051	3
1.6	0.227	0.193	0.197	0.127	-0.100	1
1.7	0.312	0.315	0.362	0.161	0.089	3
1.8	0.306	0.255	0.307	0.164	0.096	3
1.9	0.242	0.303	0.369	0.073	0.036	3
1.10	0.194	0.239	0.189	0.147	-0.113	2
2	0.372	0.213	0.387	0.195	0.139	3
3	0.431	0.280	0.478	0.171	0.130	3
5	0.384	0.214	0.520	0.201	0.074	3
6	0.117	0.169	0.158	0.100	0.100	2
7	0.340	0.298	0.388	0.188	0.075	3
9	0.405	0.297	0.427	0.151	0.107	3
10.1	0.903	0.323	0.201	0.017	0.108	1
10.2	0.829	0.271	0.138	0.019	0.159	1
10.3	0.914	0.310	0.186	0.019	0.114	1
11	0.322	0.119	0.393	0.185	0.145	3
13	0.311	0.158	0.451	0.170	0.041	3
15.1	0.407	-0.191	0.761	-0.028	0.093	3
15.2	0.405	-0.136	0.809	-0.019	0.067	3
16	0.215	0.182	0.325	0.085	0.090	3
17.1	0.180	0.579	0.431	-0.300	0.217	2
17.2	0.255	0.550	0.474	-0.258	0.270	2
18.1	0.242	0.266	0.296	0.128	0.173	3
18.2	0.255	0.243	0.335	0.116	0.227	3
18.3	0.154	0.237	0.240	0.102	0.103	3
19.1	0.268	0.502	0.355	0.187	0.014	2
19.2	0.311	0.407	0.387	0.209	0.088	2
22.1	0.241	0.355	0.368	0.413	-0.062	4
22.2	0.139	0.318	0.277	0.428	-0.067	4
23	0.303	0.336	0.373	0.289	0.007	3
24	0.139	0.097	0.094	0.118	0.078	1
25.1	0.159	0.185	0.240	0.205	0.280	5
25.2	0.276	0.339	0.349	0.263	0.025	3
26.1	0.183	0.182	0.227	0.300	0.410	5
26.2	0.221	0.202	0.253	0.256	0.576	5
26.3	0.195	0.215	0.310	0.244	0.401	5
26.5	0.331	0.250	0.401	0.224	0.314	3
27	0.242	0.299	0.322	0.095	0.162	3
28	0.259	0.302	0.350	0.296	-0.084	3

4.3.2 Comparison of the Model Dimensions and Test Specifications

Table 4.8: Investigating model dimensions in terms of sub-domain, skill and cognitive level

Dimension	Item	Sub-domain	Skill	Cognitive Level
1	1.6	Patterns, Functions & Algebra	solve problems that require knowledge of 'input and output values'	C
	10.1	Numbers, operations & relationships	common fractions, decimal fractions & percentages	K
	10.2	Numbers, operations & relationships	common fractions, decimal fractions & percentages	K
	10.3	Numbers, operations & relationships	common fractions, decimal fractions & percentages	K
	24	Measurement	do conversions in expressing measurement in different forms	R
2	1.1	Numbers, Operations & Relationships	knowledge of place value(s) of numbers	K
	1.2	Numbers, Operations & Relationships	ability to work with properties of numbers	K
	1.10	Measurement	calculating quantities such as ... temperature	R
	6	Numbers, Operations & Relationships	ability to work with properties of numbers	P
	17.1	Space & Shape	identifying types of angles	K
	17.2	Space & Shape	identifying types of angles	K
	19.1	Space & Shape	distinguishing the properties of polygons	K
	19.2	Space & Shape	distinguishing the properties of polygons	K
3	1.3	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	K
	1.4	Numbers, Operations & Relationships	identify and distinguish between multiples and factors of numbers	K
	1.5	Patterns, Functions & Algebra	identify Numeric and Geometric patterns and relationships	R
	1.7	Space & Shape	recognising symmetry in shapes,	K
	1.8	Space & Shape	locate position of objects in 2-D and 3-D space	K
	1.9	Data Handling	calculate the median of a given set of numbers	K
	2	Numbers, Operations & Relationships	knowledge of expanded notations	R
	3	Numbers, Operations & Relationships	rounding off given numbers to the specified nearest number	K
	5	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	P
	7	Numbers, Operations & Relationships	identify and distinguish between multiples and factors of numbers	R
	9	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	R
	11	Numbers, Operations & Relationships	representing numbers on a number line	R
	13	Numbers, Operations & Relationships	Operations (Addition, Subtraction, Multiplication and Division)	R
	15.1	Patterns, Functions & Algebra	problems that require knowledge of 'input and output values'	R
	15.2	Patterns, Functions & Algebra	problems that require knowledge of 'input and output values'	R
	16	Patterns, Functions & Algebra	identify Numeric and Geometric patterns and relationships	C
	18.1	Space & Shape	identifying 2-D shapes	K
	18.2	Space & Shape	identifying 2-D shapes	K
	18.3	Space & Shape	identifying 2-D shapes	K
	23	Measurement	do conversions in expressing measurement in different forms	R
	25.2	Measurement	do conversions in expressing measurement in different forms	R
	26.5	Data Handling	read and analyse piecharts	R
	27	Data Handling	calculate the mode of a given set of numbers	K
	28	Patterns, Functions & Algebra	find a pattern (non-routine problem)	P
4	22.1	Measurement	calculate time	P
	22.2	Measurement	calculate time	P
5	25.1	Measurement	read the mass on the electronic scale	K
	26.1	Data Handling	read and analyse piecharts	R
	26.2	Data Handling	read and analyse piecharts	R
	26.3	Data Handling	read and analyse piecharts	R

Table 4.9: Contingency table of cognitive level and dimension of the items

		Cognitive Level				
		K	R	C	P	Total
Dimension	1	3	1	1	0	5
	2	6	1	0	1	8
	3	10	11	1	2	24
	4	0	0	0	2	2
	5	1	3	0	0	4
	Total	20	16	2	5	43

4.4 Differential Item Functioning

The covariates that were a source of DIF were investigated by running the MIRT model with all possible combinations of the four covariates: quintile, geotype, province and gender, and inspecting the model fit statistics to identify the significant covariates. These results are tabulated in Table 4.10.

Table 4.10: Model-fit statistics for models with various covariates

Covariates included in the model	AIC	BIC	Log Likelihood
Quintile + Geotype + Province + Gender	356351.0	358233.9	-177898.5
Quintile + Geotype + Province	347245.3	349121.3	-173346.6
Quintile + Geotype + Gender	320318.0	322139.7	-159891.0
Quintile + Province + Gender	347434.9	349310.9	-173441.4
Geotype + Province + Gender	335396.6	337252.2	-167425.3
Quintile + Geotype	308397.5	310198.8	-153933.8
Quintile + Province	338381.8	340251.0	-168915.9
Quintile + Gender	320532.3	322353.9	-159998.1
Geotype + Province	326239.6	328088.5	-162847.8
Geotype + Gender	308397.5	310198.8	-153933.8
Province + Gender	326525.1	328373.9	-162990.5
Quintile	311352.6	313167.4	-155409.3
Geotype	299370.6	301165.1	-149421.3
Province	290344.0*	292131.7*	-144909.0*
Gender	299447.6	301242.1	-149459.8

All model-fit statistics indicated that the model with only province as a covariate had the best fit, which indicates that, of the four covariates, province had the closest association with variance in the data. Therefore, a DIF analysis was conducted to investigate the effect of this covariate on learner responses. DIF analyses are conducted to identify potential group differences that may result in DIF. Therefore, the learner responses from learners in each of the nine provinces were compared. The provinces are as follows Western Cape (WC), Northern Cape (NC), Free State (FS), Eastern Cape (EC), Kwa-Zulu Natal (KZN), Mpumalanga (MP), Limpopo (LP), Gauteng (GP), North West (NW).

The likelihood-ratio test was utilised to identify items for which the item estimates differed significantly between the provinces, and the ICCs of significant items were plotted for visual inspection. The results of the likelihood-ratio test of DIF are presented in Table

D.1 in Appendix D. The significance level for the test was 5%, and so items with a p-value < 0.05 was identified as having significant differences between provinces, thus presenting with DIF. The significant items were confirmed by the adjusted p-values. The results indicated there were significant differences between the provinces for the following items: 1.2, 1.7, 1.8, 2, 6, 13, 16, 22.1, 22.2, 23, 25.2, 26.1, 26.2, and 28.

The IRT method of visual inspection of ICCs for DIF investigation was utilised. The ICCs for items with significant DIF are plotted in Section 4.4.1.

4.4.1 Item Characteristic Curves of DIF items by Province

Parallel ICCs indicate uniform DIF, intersecting ICCs reveal non-uniform DIF, and ICCs that overlap completely show that there is no DIF for that item.

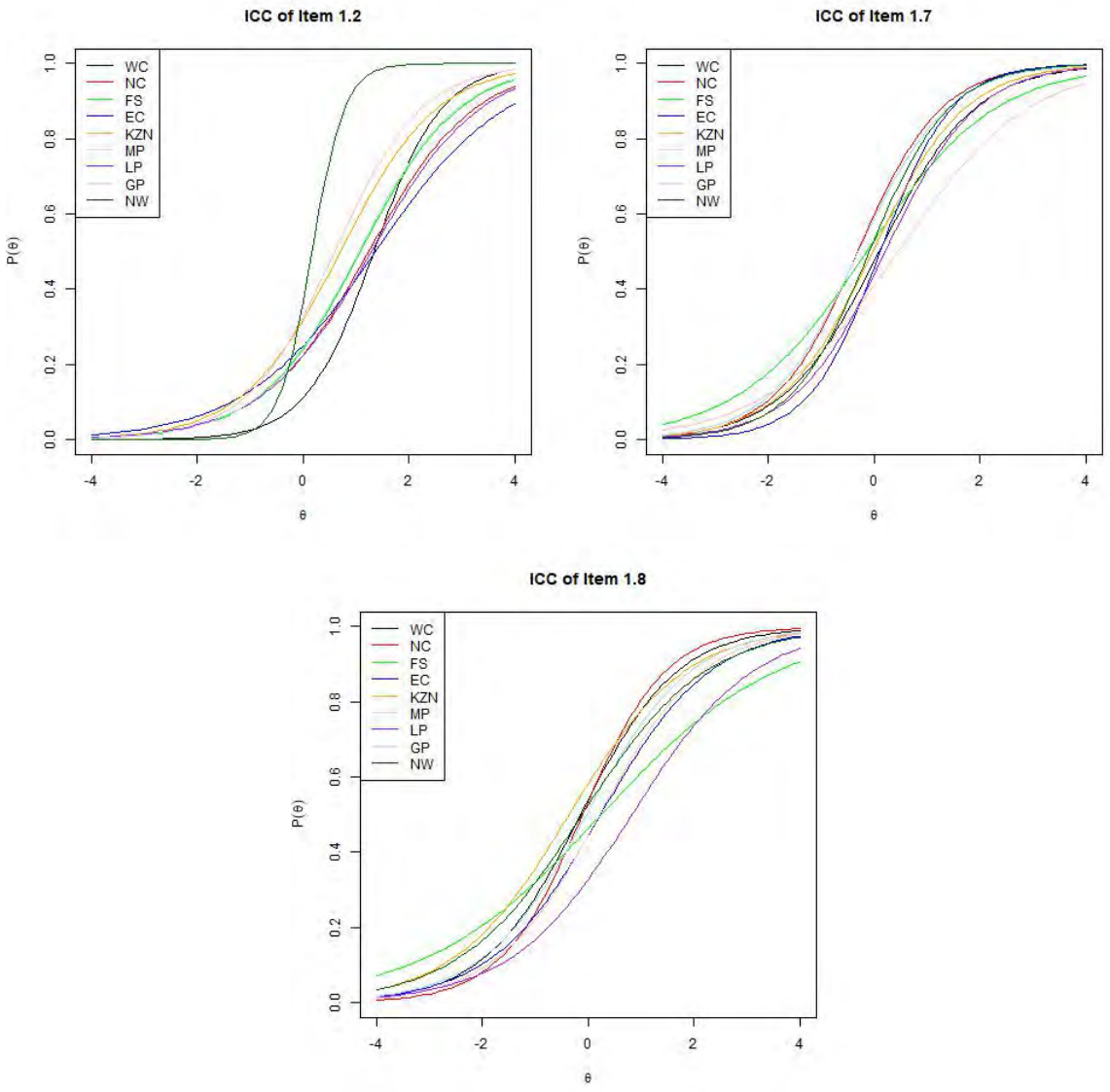


Figure 4.9: ICCs of items 1.2, 1.7 and 1.8 presenting DIF

For item 1.2, the plot (see Figure 4.9) shows that for most provinces, the item functioned the same for lower ability examinees and differently for higher ability examines. NW and WC were exceptions; the item discriminated against low ability learners from these two provinces. However, for NW learners, the item advantaged moderate and high ability learners. The ICCs for item 1.7 indicate slight non-uniform DIF can be identified when grouping FS and MP and comparing them to the rest of the provinces. Very mild uniform DIF can be seen among the rest of the provinces, based on the ICCs having a similar shape, but different locations. For item 1.8, it can also be seen that the shape and location of

FS and MP were different from the remaining provinces which had the same shape.

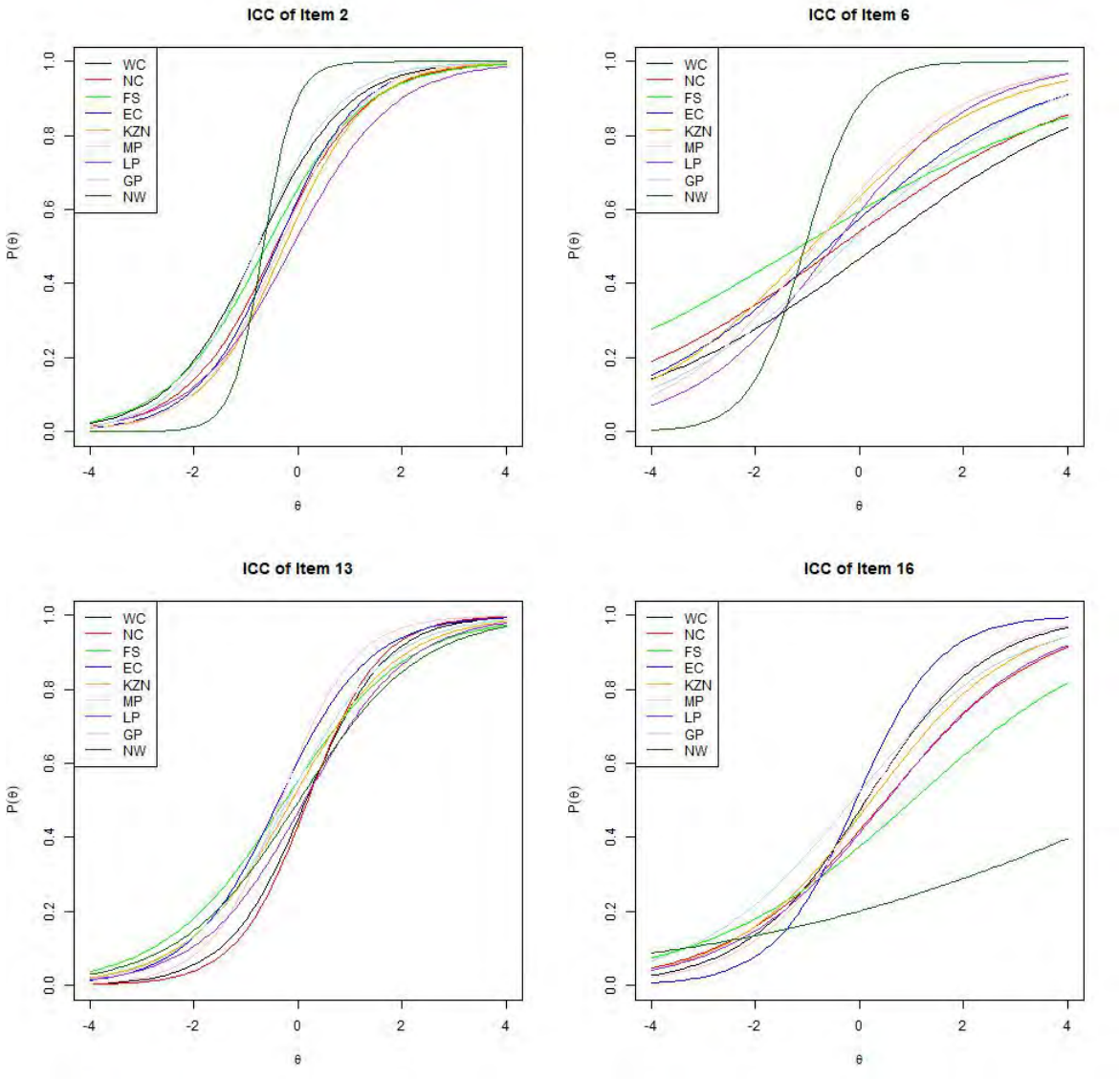


Figure 4.10: ICCs of items 2, 6, 13, and 16 presenting DIF

As seen in Figure 4.10, there is a clear difference between how item 2 functioned for NW compared to all the other provinces. Since the locations of the ICCs are close together, but the shape differs significantly between NW and the other provinces, non-uniform DIF is detected. Therefore, item 2 is biased towards high ability NW learners and against low ability learners from this province. A similar trend of non-uniform DIF concerning NW is noted for item 6, although, for the rest of the provinces, the curves are not as close to each other as for item 2. For item 13, there are no clear outliers, all provinces follow a similar

shape. Item 16 presents non-uniform, as seen by the clear intersection of the curves. For high ability learners, there are huge differences in their probabilities of a correct response to this item which is indicative of bias. The extremity of bias associated with this item is noted by the fact that, for a learner with an ability level of four, a learner from EC has an almost 100% probability of answering correctly, while a NW learner of the same ability has about a 40% probability of answering the same item correctly.

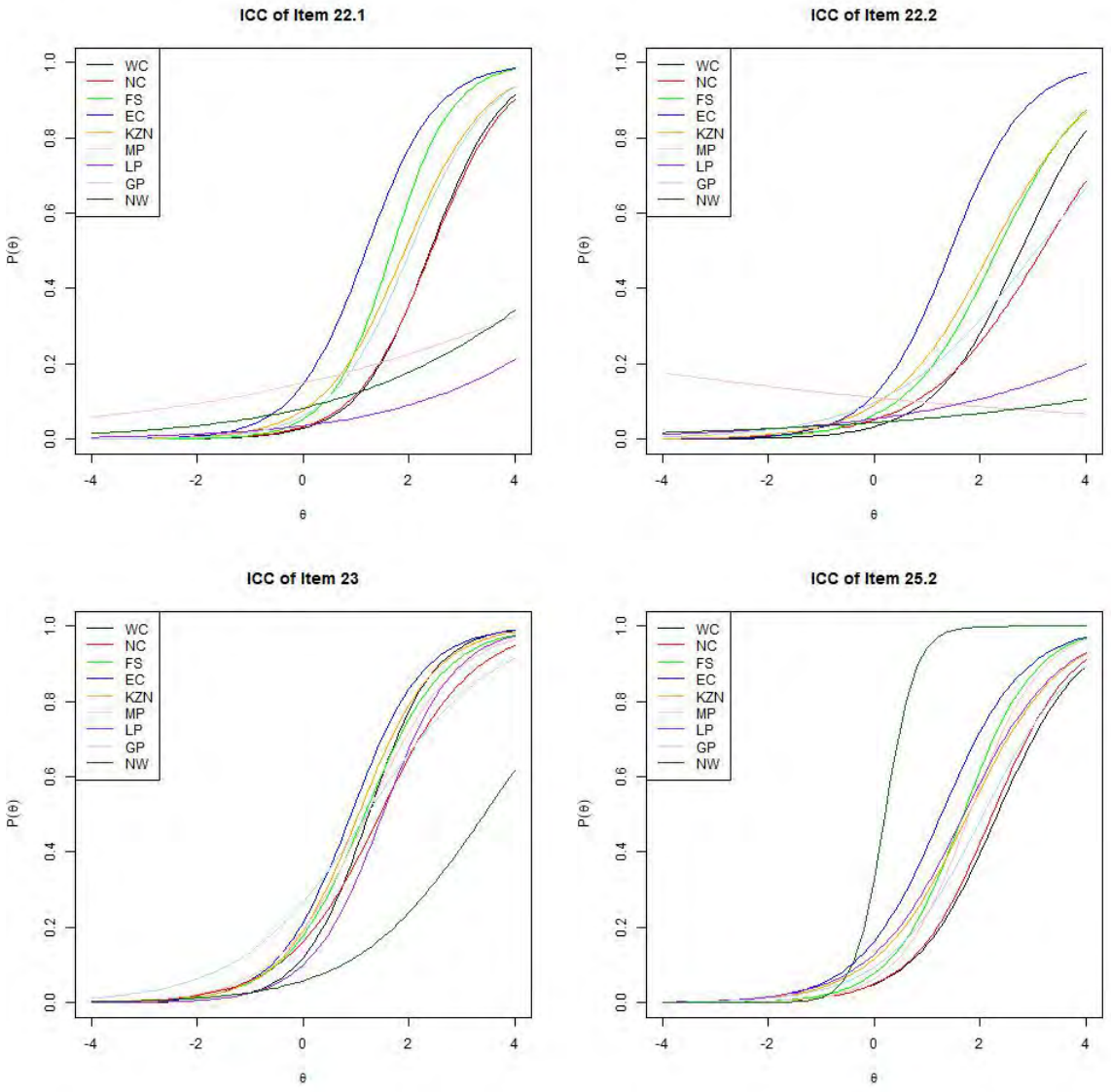


Figure 4.11: ICCs of items 22.1, 22.2, 23, and 25.2 presenting DIF

Items 22.1 and 22.2 (see Figure 4.11) were biased against the same groups. For both items, high ability learners from MP, NW, and LP had a much lower probability of answering the

items correctly. Uniform DIF can be seen among the remaining six provinces. For item 23, GP and NW had a similar shape but vastly different locations indicating uniform DIF between them. The other seven provinces had a similar shape to one another, but differed in shape from GP and differed in both shape and location from NW. The ICCs indicate that item 23 was biased against NW learners. In contrast to item 23, item 25.2 was biased towards learners from NW, especially for moderate ability where NW learners had a much higher probability of responding correctly than learners from the other provinces with the same ability.

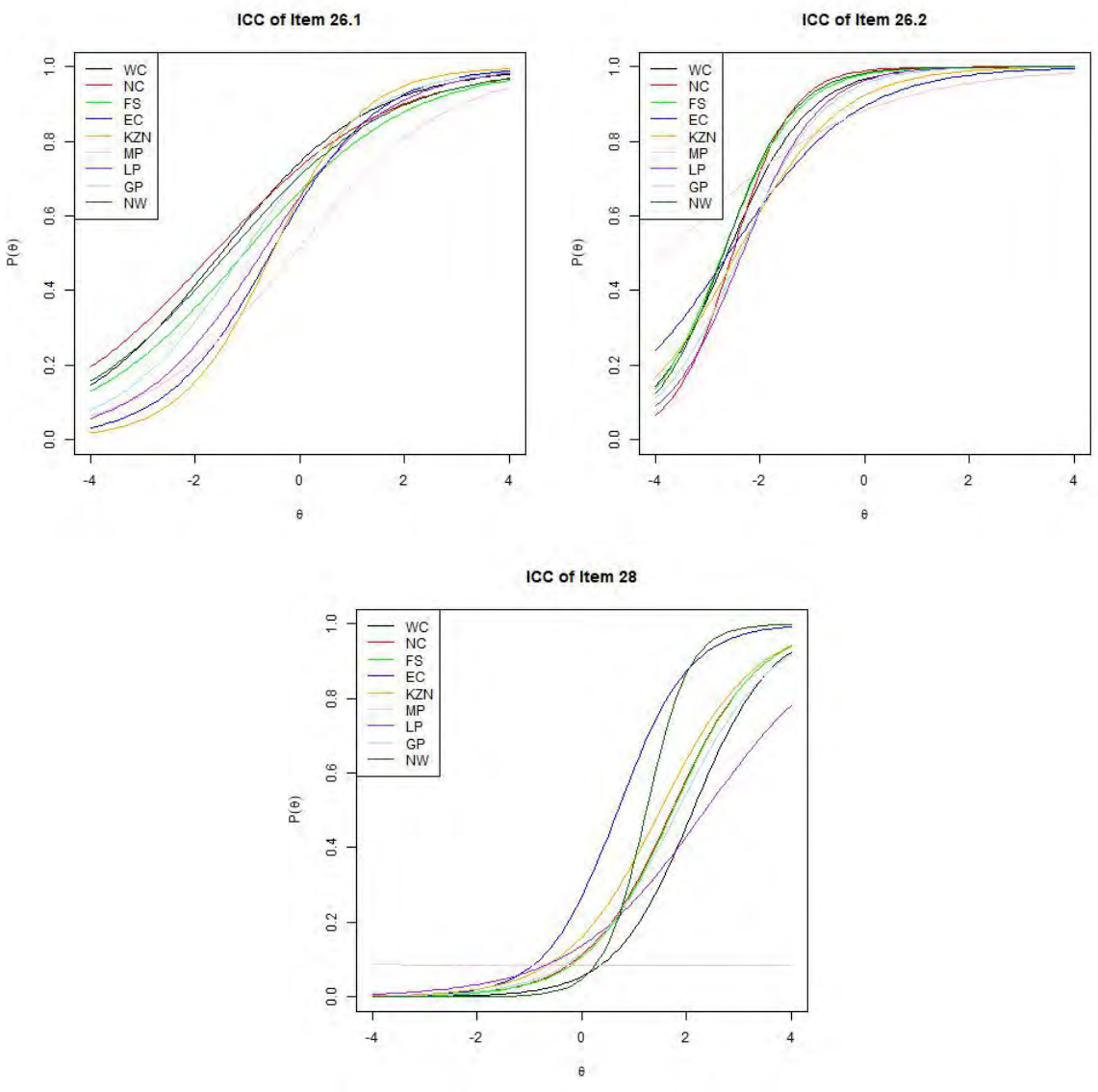


Figure 4.12: ICCs of items 26.1, 26.2, 23, and 28 presenting DIF

For item 26.1, there are large differences in the probabilities of low ability learners responding correctly, such that, at an ability level of negative four, the probability of responding correctly ranges from zero to 20%. Similar outcomes are found for item 26.2, where the probability of a learner with an ability of negative four answering correctly ranges from less than 10% to almost 50%. The ICCs for item 28 show that MP learners had the same probability of answering correctly, irrespective of ability level, as represented by the horizontal ICC in Figure 4.12. Among the remaining eight provinces, the probability of answering correctly was very similar for low abililty learners, and differed substantially for high ability learners from different provinces.

Chapter 5

Discussion

The evaluation of construct validity involves the investigation of item difficulty and discrimination parameter estimates, as well as the comparison of the dimension analysis with the test specifications, and the identification of item bias. This chapter discusses the findings of the study, highlights key outcomes and provides an evaluation of the construct validity of the test.

The parameter estimation results of the unidimensional 2PL model indicated that there were a wide range of item difficulties (-3.093; 2.953) and discrimination values (0.358; 3.102) present in the test.

The difficulty of the items was distributed such that most items were moderately difficult (67.44%) while almost a quarter (23.26%) of the items were either difficult or very difficult and 9.30% were easy or very easy. In the 2014 ANA cycle, the distribution of question difficulty in all the tests was intended to be 20% easy, 60% moderate and 20% difficult; or 40% easy, 40% moderate and 20% difficult, depending on the requirements of the curriculum policy (DBE, 2014b). Based on the results, it is assumed that the grade 6 test was intended to be 20% easy, 60% moderate and 20% difficult. However, some of the moderately difficult items should be replaced by easy items to fit the framework intended by the DBE. The distribution of difficulty indicates that the test is best suited for learners

with an ability level that corresponds to the moderately difficult category. Therefore, the test favours learners within this central band of ability.

All discrimination parameters were greater than zero which satisfies the assumption of monotonicity. This represents a positive slope for all item functions which indicates that for all items, an examinee with a higher ability has a higher chance of providing the correct answer to the item. Since all discrimination estimates were above 0.34, there were no items that were considered poor and needed to be completely eliminated. However, the majority of items were functioning marginally or moderately, indicating that revision was needed. The marginal items (1.6; 1.10; 6 and 24) were all moderately difficult items that provided very little information on the ability levels of the examinees, which can be seen by the extremely flat item information curves for these items in Figure 4.2. Based on the interpretation of discrimination from (Bichi and Talib, 2018), all items that are moderate or good need little to no revision, however, the moderately functioning items would probably need more revision than the good items.

Four items (3; 10.1; 10.2; and 10.3) were functioning satisfactorily and did not need revision. Item 3 involved rounding off and items 10.1-10.3 involved converting between decimals, fractions and percentages. These satisfactory items were very effective in determining the examinees' ability levels at their given difficulty levels. Unfortunately, all of these items were moderately difficult which indicates that the test provides less information on learners whose ability is more towards the upper or lower end of the ability scale. This aligns with the analysis of item difficulty and indicates that the test is most effective in providing information on the ability of examinees who have ability levels ranging from -1 to 1. Since this is in agreement with the intended difficulty distribution, the results of the unidimensional IRT analysis provide evidence to support the claim of construct validity of the test. However, since the test specifications indicate that the test is a multidimensional assessment tool, the multidimensional IRT analysis holds more weight in the evaluation of validity.

Since it was hypothesised that the dimensions would align with the five sub-domains being measured in the test, the fact that the best-fitting MIRT model was the five-dimensional

model was a good indication of a possible alignment between the subdomains and the dimensions. However, once the prominent dimension of each item was extracted, results indicated that for the most part, the dimensions did not align well with the sub-domains of the test. Instead there was a single dimension (m=3) that included a range of general Mathematics skills from all five sub-domains, and each of the other dimensions consisted of fewer items that tested more specific skill groups. The items from each dimension is discussed below in terms of their similarities and differences with regard to sub-domain, skill and cognitive level.

The first dimension was comprised of five items. Three of the items (10.1; 10.2; 10.3) were from the sub-domain of numbers, operations & relationships testing the examinees' knowledge of common fractions, decimal fractions and percentages by requiring them to convert between the different forms. Item 24 was also in the first dimension, and while this item had a different sub-domain (measurement), context (shot put distances) and cognitive level (routine procedure), it also required the skill of converting values into different forms. Item 1.6 seemed to be the odd one out as a patterns, functions and algebra item testing students' ability to conduct complex procedures to determine input and output values.

The second dimension identified according to the model was also predominantly items from the cognitive level of knowledge, and were more language focused, requiring examinees to know and apply specific terminology from different sub-domains. The terminology tested in these items were derived from the sub-domains of numbers, operations & relationships; measurement; and space & shape.

Examples of specific terminology tested explicitly in this dimension were "prime number" (1.2) and "boiling point" (1.10). The terminology was more implied in some items: "value of underlined digit" (1.1) tested place value, while "re-arrange ... smallest to the biggest" (6) tested students' problem-solving ability to recognise this as ascending order and arrange the numbers accordingly. Other items were language heavy, both in the wording of the question and in requiring specific terms as answers. Included in this category were items 17.1 and 17.2 which required examinees to label "different kinds of angles" with

correct answers of "obtuse" and "right" angles; and items 19.1 and 19.2 which tested properties of polygons (parallelogram and rectangle) and required examinees to fill a missing word into each sentence.

Based on the results from the unidimensional analysis, the items in this dimension were more towards the difficult end of the scale with -3.019 $\leq \beta \leq$ 0.151 (smaller/negative values indicate more difficult items in the mirt package). This finding is in alignment with those of the qualitative study conducted by Graven and Venkat (2014) with 54 teachers on their experiences administering the 2012 grade three mathematics ANA. Their research revealed teachers' concerns pertaining to the level and speed of reading required for students to understand and respond to the ANA items, and consequent apprehension about the accessibility of the test items. The findings from this study indicate that although the research by Graven and Venkat (2014) included a sample of teachers that was not nationally representative, the reality of their findings may extend, not only to the broader grade 3 population but to other grades (and years).

The fact that a dimension of eight linguistically demanding items was identified by the model indicates that 18.40% of the dichotomous portion of this ANA may have tested an ability that is not purely mathematical. Furthermore, it is possible that this dimension of items gave an advantage to examinees who were more competent in terms of English reading ability and/or more familiar with English as a home language, both of which have been suggested in literature pertaining to assessment more broadly (e.g. Bohlmann and Pretorius (2008), Lahoud (2021) respectively). Although it would be interesting to compare the item functioning of these items for groups of examinees based on learner home language or reading ability, this information was not available to the author, and thus beyond the scope of this study.

Dimension three comprised more than half the items (24/43) and included items from all sub-domains and cognitive levels. Within the sub-domain of numbers, operations and relationships, four items tested basic operations (addition, subtraction, multiplication, and division). One of these items tested knowledge (1.3), while two of these items (9; 13) were routine problems and one (5) required problem-solving. Other routine problems

tested expanded notation (2), multiples of seven (7), and representing a value on a number line (11). Knowledge of factors and rounding off was needed to respond correctly to items 1.4 and 3 respectively.

Patterns, functions and algebra were tested by five items within this dimension. Two routine procedure items involved input and output values (15.1; 15.2), and another routine problem incorporated a simple number sequence. A more complex procedure was required to continue a geometric pattern in item 16, and item 28 involved problem-solving to fill in missing values for a non-routine pattern problem.

All five of the space and shape items in dimension three tested knowledge as follows: recognising symmetry (1.7), locating the position of an object in a 3D space (1.8), and identifying 2D shapes (18.1; 18.2; 18.3). Two items were within the sub-domain of measurement, and both items required routine procedures to convert different forms of measurement (millilitres to litres (23) and kilograms to grams (25.2)).

The final group of items in the third dimension tested data handling ability. Two items tested knowledge of median (1.9) and mode (27) and another required a routine procedure for reading information from a piechart and converting the information to a fraction. The exhaustive range of sub-domains and cognitive levels present in this dimension are indicative that this dimension would likely have tested a general Mathematics ability, while the other dimensions tested specific skills or skill groups.

Dimension four was the smallest dimension, with only two items from the sub-domain of measurement. Both items involved answering problem-solving questions based on a diagram of analogue clocks from different time zones. Item 22.1 involved calculating the time difference and 22.2 involved using this time difference to calculate the time at one place given the other. The specificity of these two items and their inclusion in a separate dimension is indicative that the problem-solving ability required to respond correctly to these items is not tested by any other items. Furthermore, the unidimensional difficulty parameters for items 22.1 (-3.093) and 22.2 (-2.782) indicate that they were very difficult for the student to answer correctly, and only examinees who scored extremely high on the test overall had the ability to respond correctly. This is in agreement with DBE (2014a)

who shed some light on what may have caused so many students to answer incorrectly. A percentage of examinees resorted to counting the time interval from the displayed clocks rather than using a mathematical calculation to subtract the two time intervals. By not considering the information on whether it was morning (a.m.) or afternoon (p.m.), some students concluded that there was a three hour difference rather than nine hours.

The fifth and final dimension included three routine data handling items and one measurement item testing knowledge. Item 25.1 was a simple item requiring students to write down a mass given on a digital scale. The other three items (26.1; 26.2; 26.3) tested students' ability to perform a routine procedure of reading information from a piechart and answering questions that did not require any conversions. All items in this dimension involved using information directly from a diagram and responding to instructions of minimum complexity.

An investigation of item bias was conducted by means of a DIF analysis. The model-data fit statistics (see Table 4.10) indicated that the model with Province as a covariate was the best fitting model. This indicated that the test items functioned differentially for learners from different Provinces. The DBE (2014b) provided the ANA results by province (see Figure 1.1) which showed stark differences in aggregated performance of the learners from different provinces. The identification of province as a potential source of bias indicates that the performance differences could be a result of construct-irrelevant differences, rather than differences in true ability (Gierl, 2004).

The ICCs were plotted for all items for which significant DIF was detected, and three kinds of ICC shapes were noted. Firstly, there were items for which significant DIF was detected, but the provinces followed the same general ICC shape. Since the differences were more due to location differences than shape differences, the type of DIF present was most likely uniform DIF. The second and third kinds of ICC shapes observed Items 1.2, 2, 6, and 25.2 present non-uniform DIF for learners from NW compared to learners from other provinces. This is indicated by the fact that low ability NW learners had a lower probability of answering correctly than learners from other provinces with the same ability, but high ability learners from NW had a higher probability of providing a correct

answer than learners of the same ability from other provinces. In other words, these items discriminated more effectively between ability levels for learners from NW in comparison to other provinces. It is noted that items 1.2, 2, and 6 were items from the sub-domain of Numbers, Operations and Relationships.

Visual inspection of the ICCs for items 1.7, 1.8, 13, 26.1, and 26.2 indicated that there were no provinces for which the ICCs deviated from the general shape. Although significant DIF was detected, there was no single province for which the functioning of the items was in stark contrast to its functioning for the other provinces. Four of these items were moderately difficult and had a moderate quality. Items 1.7 and 1.8 were testing knowledge, while items 13, 26.1 and 26.2 were routine procedure questions. Items 26.1 and 26.2 fell into dimension three and the remaining three were in dimension two.

Items 16, 22.1, 22.2, 23, and 28 were identified as having non-uniform DIF, with high ability learners from some provinces having a much lower probability of correctly answering the items than learners from the majority of provinces. These items were all on the more difficult side of the scale, but the ICCs indicate that they were even more difficult for learners from some provinces. For items 16 and 23, which were both in the third dimension, learners from NW were disadvantaged as the items did not discriminate effectively between NW learners of differing ability levels. Item 28, which was also in dimension three, disadvantaged learners from MP. Items 22.1 and 22.2 – which were very difficult, problem-solving questions that required learners to calculate time across time zones – disadvantaged learners from MP, LP and NW.

No strong relationships were found between the presence of DIF and the items' sub-domains, dimensions or cognitive levels. In addition, there is huge diversity among learners within each province, which makes it difficult to pinpoint any root causes of the DIF associated with province. There is, therefore, not nearly enough information available to make any comments on the possible reasons for DIF between provinces. However, DIF is indicative of item bias, and the presence of item bias means that the inferences and decisions about the true ability of examinees based on this test are incorrect (Chaimongkol, 2005). The fact that 14 items were identified as having significant DIF for learners from

different provinces, provides evidence against the validity of the test (Abedlazeez, 2010, Chen *et al.*, 2021).

Chapter 6

Conclusion

This study demonstrated how multidimensional item response theory and its associated statistical techniques, which have been underutilised in the development of large-scale assessments in South Africa, can be employed to investigate the utility value of a large-scale assessment regarding its validity. The following three paragraphs outline the three key findings of this study.

The evaluation of construct validity in terms of the difficulty and discrimination parameter estimates provided evidence to support the notion of construct validity due to the alignment of the difficulty contributions of the test specifications and the two parameter logistic item response theory model. The test was found to provide the most information on learners within the central band of ability, which indicates that it was better suited for learners within this central range than for those on either end of the ability scale.

The factor structure identified by the multidimensional item response theory model did not align well with the table of specifications. This presented some concerns as it provided evidence against the validity of the test. Most notably, the results suggested that a substantial proportion of the test measured an ability that is not purely mathematical. All items in this dimension required knowledge of specific terminology and/or a high level of literacy to understand and respond correctly. Alarmingly, it is possible that this dimension of items gave an advantage to examinees who were more familiar with English

as a home language and/or more competent in terms of English reading ability, despite the claim that the assessment was suitable for the target grade and used appropriate language.

The differential item functioning analysis revealed that fourteen items were biased, such that construct-irrelevant factors affected the test performance of learners from different provinces. There was insufficient information available to make any conclusions about the reasons for differential item functioning; however, its mere presence in about a third of the test items contributed towards evidence against the validity of the test.

It should be noted that the lack of involvement of the author in the data collection process prevented the inclusion of potentially more informative covariates, such as the home language or reading ability of the learners. In addition, the large proportion of unexplained missing values provides a limitation to the study. It is imperative that this information is accessible to educational measurement specialists when conducting these evaluations for the implementation of large-scale assessments. Despite these limitations, the results clearly show that there are substantial differences between the validity claims and the evidence generated by this study. If this test is a reflection of the broader context of South African national assessments, steps need to be taken to enhance the utility value of future assessments through more effective methods, such as item response theory techniques.

When the test validity is found to be lacking, it is recommended that the necessary stakeholders and content experts are involved in making appropriate adjustments to ameliorate the psychometric properties of the assessment, and confirmatory analyses should be conducted to verify the efficacy of the adjustments. In the context of differential item functioning analyses, it will be of benefit for all social and linguistic covariates to be assessed to establish the test as an unbiased measure for all social and linguistic groups. The covariate information accompanying the test responses should be collected with these goals in mind.

In conclusion, multidimensional item response theory provided an effective and rigorous approach to establishing the validity of a large-scale assessment. To avoid the pitfalls of

the annual national assessments, it is recommended that this method is utilised for the development and evaluation of future national assessment instruments in South Africa.

References

Abedlazeez, N. Exploring dif: comparison of ctt and irt methods. *OIDA International Journal of Sustainable Development*, 1(7):11–46, 2010.

Ackerman, P. L. Predicting individual differences in complex skill acquisition: dynamics of ability determinants. *Journal of applied psychology*, 77(5):598, 1992.

Adams, R. J., Wilson, M., and Wu, M. Multilevel item response models: An approach to errors in variables regression. *Journal of educational and behavioral Statistics*, 22(1):47–76, 1997.

Adedoyin, O. and Mokobi, T. Using irt psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*, 3(4):992–1011, Apr. 2013.
URL https://archive.aessweb.com/index.php/5007/article/view/2471

AERA. Standards for educational and psychological testing. American Educational Research Association, 2014.

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi:10.1109/TAC.1974.1100705.

An, X. and Yung, Y. F. Item response theory: What it is and how you can use the irt procedure to apply it. 2014.

Angoff, W. H. and Ford, S. F. Item-race interaction on a test of scholastic aptitude 1. *Journal of Educational Measurement*, 10(2):95–105, 1973.

Association, A. E. R., Association, A. P., on Measurement in Education, N. C. *et al.* Standards for educational and psychological testing. American Educational Research Association, 1999.

Baker, F. B. The basics of item response theory. ERIC, 2001.

Bansilal, S. What can we learn from the kzn ana results? *SA-eDUC*, 9(2), 2012.

Bayes, T. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, (53):370–418, 1763.

Beanland, C., Schneider, Z., LoBiondo-Wood, G., and Haber, J. Nursing research: Methods, critical appraisal and utilization. 1999.

Bichi, A. A. and Talib, R. Item response theory: An introduction to latent trait models to test and item development. *Int. J. Eval. Res. Educ. (IJERE)*, 7(2):142, June 2018.

Binet, A. and Simon, T. Recherches de pédagogie scientifique. *L'Année psychologique*, 12(1):233–274, 1905.

Bock, D. A brief history of item theory. *Educ. Meas*, 16:21–33, 1997.

Bock, R. D., Gibbons, R., and Muraki, E. Full-information item factor analysis. *Applied psychological measurement*, 12(3):261–280, 1988.

Bohlmann, C. and Pretorius, E. Relationships between mathematics and literacy: Exploring some underlying factors. *Pythagoras*, 2008(1):42–55, 2008.

Box, G. E. P. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976. doi:10.1080/01621459.1976.10480949.

Browne, M. and Cudeck, R. Alternative ways of assessing model fit. *Testing structural equation models*, 154:136, 1993.

Bürkner, P.-C. Analysing standard progressive matrices (spm-ls) with bayesian item response models. 2019.

Burnham, K. and Anderson, D. R. Model selection and multimodel inference, 2nd edn new york. *NY: Springer.[Google Scholar]*, 2002.

Burnham, K. P. and Anderson, D. R. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

Cai, L. A two-tier full-information item factor analysis model with applications. *Psychometrika*, 75(4):581–612, 2010.

Cappelleri, J. C., Jason Lundy, J., and Hays, R. D. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clin. Ther.*, 36(5):648–662, May 2014.

Cardall, C. and Coffman, W. E. A method for comparing the performance of different groups on the items in a test. Educational Testing Service, 1964.

Carlson, J. E. and Davier, M. v. Item response theory. In *Advancing human assessment*, pages 133–178. Springer, Cham, 2017.

Chaimongkol, S. Modeling differential item functioning (DIF) using multilevel logistic regression models: A Bayesian perspective. The Florida State University, 2005.

Chalmers, R. P. mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, 48:1–29, 2012.

Chen, Y., Li, X., Liu, J., and Ying, Z. Item response theory – a statistical framework for educational and psychological measurement. 2021.

Clarke, P., Crawford, C., Steele, F., and Vignoles, A. F. The choice between fixed and random effects models: Some considerations for educational research. *SSRN Electron. J.*, 2010.

Cohen, A. and Kang, T. Irt model selection methods for dichotomous items. *Applied Psychological Measurement*, 31:331–358, 07 2007. doi:10.1177/0146621606292213.

Considine, J., Botti, M., and Thomas, S. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, 12(1):19–24, 2005.

Crocker, L. and Algina, J. Introduction to classical and modern test theory. ERIC, 1986.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951.

DBE. Assessment Policy in the General Education and Training Band: Grade R to 9 and ABET, volume 402. and obtainable from the Government Printer, 1998.

DBE. Report on the annual national assessments of 2011. 2011.

DBE. Report on the annual national assessments of 2012. 2012.

DBE. The annual national assessment of 2013 2013 diagnostic report and 2014 framework for improvement. Technical report, Department of Basic Education, 2013a.

DBE. Report on the annual national assessments of 2013. 2013b.

DBE. The annual national assessment of 2014 diagnostic report intermediate and senior phases mathematics. Technical report, Department of Basic Education, 2014a.

DBE. Report on the annual national assessments of 2014. 2014b.

de Ayala, R. The theory and practice of item response theory. 2009.

DeVellis, R. F. Scale development: Theory and applications (Second Edition). Sage publications, 2003.

Emberston, S. and Reise, S. Item response theory for psychologists. marwah. 2000.

Finch, W. H. and French, B. F. Educational and psychological measurement. Routledge, October 2019.

Fox, J.-P. and Glas, C. A. Bayesian estimation of a multilevel irt model using gibbs sampling. *Psychometrika*, 66(2):271–288, 2001.

Foxcroft, C. D. Ethical issues related to psychological testing in africa: What i have learned (so far). *Online readings in psychology and culture*, 2(2):2307–0919, 2011.

Frey, B. B. The SAGE encyclopedia of educational research, measurement, and evaluation. Sage Publications, 2018.

Gamerman, D., Gonçalves, F. B., and Soares, T. M. Differential item functioning. In *Handbook of item response theory*, pages 67–86. Chapman and Hall/CRC, 2017.

Gibbons, R. D. and Hedeker, D. R. Full-information item bi-factor analysis. *Psychometrika*, 57(3):423–436, 1992.

Gierl, M. Using a multidimensionality-based framework to identify and interpret the construct-related dimensions that elicit group differences. In *Annual Meeting of the American Educational Research Association (AERA)*, pages 12–16. 2004.

Goldstein, H. Multilevel statistical models. John Wiley & Sons, Ltd, Chichester, UK, October 2010.

González-Betanzos, F. and Abad, F. J. The effects of purification and the evaluation of differential item functioning with the likelihood ratio test. *Methodology*, 2012.

Govender, R., Bowen, P., and Edwards, P. Measurement scales for aids-related knowledge and stigma in south africa: An evaluation using item response theory. *Journal of AIDS and HIV Research*, 8(3):12–24, 2016.

Graven, M. and Venkat, H. Primary teachers' experiences relating to the administration processes of high-stakes testing: the case of mathematics annual national assessments. *African Journal of Research in Mathematics, Science and Technology Education*, 18(3):299–310, 2014.

Haladyna, T. M. Developing and validating multiple-choice test items. Lawrence Erlbaum, 1999.

Hambleton, R. and Jodoin, M. Item response theory: Models and features. *Encyclopedia of psychological assessment*, pages 509–514, 2003.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. Fundamentals of item response theory, volume 2. Sage, 1991.

Hambleton, R. K. and Van der Linden, W. J. Advances in item response theory and applications: An introduction. 1982.

Hartig, J. and Höhler, J. Representation of competencies in multidimensional irt models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie/Journal of Psychology*, 216(2):89, 2008.

Hartig, J. and Höhler, J. Multidimensional irt models for the assessment of competencies. *Studies in Educational Evaluation*, 35(2):57–63, 2009. ISSN 0191-491X. doi:https://doi.org/10.1016/j.stueduc.2009.10.002. Assessment of Competencies. URL https://www.sciencedirect.com/science/article/pii/S0191491X09000212

Holland, P. W. and Thayer, D. T. Differential item functioning and the mantel-haenszel procedure. *ETS Research Report Series*, 1986(2):i–24, 1986.

Holland, P. W. and Thayer, D. T. Differential item performance and the Mantel-Haenszel procedure, pages 129–145. Routledge, 1988.

Holzinger, K. J. and Swineford, F. The bi-factor method. *Psychometrika*, 2(1):41–54, 1937.

Immekus, J. C., Snyder, K. E., and Ralston, P. A. Multidimensional item response theory for factor structure assessment in educational psychology research. In *Frontiers in Education*, volume 4, page 45. Frontiers Media SA, 2019.

Jodoin, M. G. and Gierl, M. J. Evaluating type i error and power rates using an effect size measure with the logistic regression procedure for dif detection. *Applied measurement in education*, 14(4):329–349, 2001.

Jun, H. W. Diagnostic measurement from a standardized math achievement test using multidimensional latent trait models. Ph.D. thesis, Georgia Institute of Technology, 2014.

Kanjee, A. Using logistic regression to detect bias when multiple groups are tested. *S. Afr. J. Psychol.*, 37(1):47–61, April 2007.

Kanjee, A. Item response theory: Applications and challenges for large-scale assessments in south africa, 2010. Tshwane University of Technology.

Kanjee, A. Review of the annual national assessments (anas) – 2009 to 2014: Implication of the systemic evaluation studies, 2016.

Kanjee, A., Govender, V., Greer, E., Herholdt, R., Makgamatha, M., and Sikala, E. Key findings and recommendations for the annual national assessments, 2013. Assessment Advisory Committee.

Kanjee, A. and Moloi, Q. South african teachers' use of national assessment data. *South African Journal of Childhood Education*, 4(2):90–113, 2014.

Kanjee, A. and Sayed, Y. Assessment policy in post-apartheid south africa: Challenges for improving education quality and learning. *Assessment in Education: Principles, Policy & Practice*, 20(4):442–469, 2013.

Kline, T. Classical test theory: Assumptions, equations, limitations, and item analyses. *Psychological testing: A practical approach to design and evaluation*, 91, 2005.

Kruglova, N., Dykhovychnyi, O., and Lysenko, D. Application of irt and mirt models to analysis of analytical geometry tests. , 38:36–49, 06 2021.

Lahoud, T. A. 2020 matrics face the pandemic of educational disparity, 2021. Honours Thesis, Rhodes University.

Lopez Rivas, G. E., Stark, S., and Chernyshenko, O. S. The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, 33(4):251–265, 2009.

Lord, F. M. A theory of test scores (psychometric monograph no. 7). *Iowa City, IA: Psychometric Society*, 35, 1952.

Lord, F. M. Applications of item response theory to practical testing problems (First Edition). Routledge, New York, England, November 1980.

Lord, F. M. Applications of item response theory to practical testing problems. Routledge, London, England, November 2012.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores / Frederic M. Lord and Melvin R. Novick ; with contributions by Allan Birnbaum. Addison-Wesley Pub. Co Reading, Mass, 1968, xvii, 568 p. ; pages.

Magno, C. Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, 1(1):1–11, 2009.

Mantel, N. and Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4):719–748, 1959.

Martin-Löf, P. The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data [with discussion]. *Scandinavian Journal of Statistics*, pages 3–18, 1974.

Maydeu-Olivares, A. Goodness-of-fit assessment of item response theory models. *Measurement (Mahwah NJ)*, 11(3):71–101, July 2013.

McKinley, R. L. and Reckase, M. D. An extension of the two-parameter logistic model to the multidimensional latent space. Technical report, American Coll Testing Program Iowa City Ia Resident Programs Dept, 1983.

Mislevy, R. J. Estimation of latent group effects. *Journal of the American Statistical Association*, 80(392):993–997, 1985.

Nering, M. L. and Ostini, R. Handbook of polytomous item response theory models. Taylor & Francis, 2011.

Penfield, R. D. Assessing differential item functioning among multiple groups: A comparison of three mantel-haenszel procedures. *Applied Measurement in Education*, 14(3):235–259, 2001.

Popham, W. J. Assessment literacy for teachers: Faddish or fundamental? *Theory into practice*, 48(1):4–11, 2009.

Rasch, G. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. 1960.

Raykov, T. On the use of confirmatory factor analysis in personality research. *Personality and Individual Differences*, 24(2):291–293, 1998.

Reckase, M. D. The difficulty of test items that measure more than one ability. *Applied psychological measurement*, 9(4):401–412, 1985.

Reckase, M. D. The discriminating power of items that measure more than one dimension. 1986.

Reckase, M. D. Multidimensional item response theory models. In *Multidimensional item response theory*, pages 79–112. Springer, 2009.

Reckase, M. D. and McKinley, R. L. The discriminating power of items that measure more than one dimension. *Applied psychological measurement*, 15(4):361–373, 1991.

Reddy, V., Winnaar, L., Juan, A., and Arends, F. Education policy and curriculum in mathematics and science: South africa. *TIMSS 2019 Encyclopedia*, 2019.

Rose, N., Von Davier, M., and Xu, X. Modeling nonignorable missing data with item response theory (irt). *ETS Research Report Series*, 2010(1):i–53, 2010.

Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464, 1978. doi:10.1214/aos/1176344136.
URL https://doi.org/10.1214/aos/1176344136

Sheng, Y. and Wikle, C. K. Bayesian multidimensional irt models with a hierarchical structure. *Educational and psychological measurement*, 68(3):413–430, 2008.

S.J. Howie, V. S. E. V., C. Long. The role of irt in selected examination systems. Technical report, Umalusi Council for Quality Assurance in General and Further Education and Training, 2009.

Spaull, N. South africa's education crisis: The quality of education in south africa 1994-2011. *Johannesburg: Centre for Development and Enterprise*, 21(1):1–65, 2013.

Spearman, C. 'general intelligence', objectively determined and measured. *The American Journal of Psychology.*, 15(2):201–293, 1904.

Stemler, S. E. and Naples, A. Rasch measurement v. item response theory: Knowing when to cross the line. *Practical Assessment, Research & Evaluation*, 26:11, 2021.

Sulis, I. and Toland, M. D. Introduction to multilevel item response theory analysis: Descriptive and explanatory models. *The Journal of Early Adolescence*, 37(1):85–128, 2017. doi:10.1177/0272431616642328.
URL https://doi.org/10.1177/0272431616642328

Swaminathan, H. and Rogers, H. J. Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, 27(4):361–370, 1990.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., and Featherman, C. Analysis of differential item functioning (dif) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, 27(1):53–75, 2002.

Taylor, C. S. Validity and validation. Oxford University Press, 2013.

Thurstone, L. L. A method of scaling psychological and educational tests. *Journal of educational psychology*, 16(7):433, 1925.

Ukanda, F. I. Effectiveness of mantel-haenszel and logistic regression statistics in detecting differential item functioning under different conditions. Ph.D. thesis, Maseno University, 2019.

Van der Berg, S. What the annual national assessments can tell us about learning deficits over the education system and the school career. *South African Journal of Childhood Education*, 5(2):28–43, 2015.

Van der Linden, W. J. and Hambleton, R. Handbook of item response theory. *Taylor & Francis Group. Citado na pág*, 1(7):8, 1997.

von Davier, M. and Sinharay, S. Analytics in International Large-Scale Assessments: Item Response Theory and Population Models, page 155–171. Taylor amp; Francis Group, 2014.

Warm, T. A. A primer of item response theory. Technical report, COAST GUARD WASHINGTON DC, 1978.

White, C. and Van Dyk, H. Theory and practice of the quintile ranking of schools in south africa: A financial management perspective. *South African Journal of Education*, 39(Supplement 1):s1–19, 2019.

Yan, D., Von Davier, A. A., and Lewis, C. Computerized multistage testing: Theory and applications. CRC Press, 2016.

Zimkowski, M., Muraki, E., Mislevy, R., and Bock, R. Bilog-mg: Multiple-group irt analysis and test maintenance for binary items (computer software and manual). chicago, il: Scientific software international. 1996.

Zumbo, B. and Thomas, D. A measure of dif effect size using logistic regression procedures. *National Board of Medical Examiners, Philadelphia, PA*, 1996.

Zumbo, B. D. A handbook on the theory and methods of differential item functioning (dif). *Ottawa: National Defense Headquarters*, 160, 1999.

Zwick, R. and Ercikan, K. Analysis of differential item functioning in the naep history assessment. *Journal of educational measurement*, 26(1):55–66, 1989.

Appendix A

Tables

Item	Mean	SD	total.r	total.r_if_rm	alpha_if_rm
1.1	0,227	0,419	0,377	0,335	0,904
1.2	0,225	0,417	0,433	0,393	0,903
1.3	0,410	0,492	0,488	0,444	0,902
1.4	0,314	0,464	0,383	0,336	0,904
1.5	0,615	0,487	0,404	0,357	0,904
1.6	0,382	0,486	0,307	0,256	0,905
1.7	0,455	0,498	0,512	0,468	0,902
1.8	0,447	0,497	0,463	0,417	0,903
1.9	0,577	0,494	0,454	0,407	0,903
1.10	0,533	0,499	0,318	0,265	0,905
2	0,556	0,497	0,515	0,471	0,902
3	0,532	0,499	0,585	0,545	0,901
5	0,393	0,488	0,586	0,547	0,901
6	0,535	0,499	0,268	0,215	0,905
7	0,183	0,386	0,471	0,436	0,903
9	0,440	0,496	0,573	0,534	0,901
10.1	0,429	0,495	0,669	0,636	0,9
10.2	0,549	0,498	0,567	0,526	0,901
10.3	0,414	0,493	0,666	0,633	0,9
11	0,557	0,497	0,462	0,416	0,903
13	0,454	0,498	0,484	0,439	0,902
15.1	0,592	0,491	0,501	0,457	0,902
15.2	0,553	0,497	0,542	0,5	0,902
16	0,414	0,493	0,396	0,347	0,904
17.1	0,378	0,485	0,505	0,462	0,902
17.2	0,486	0,5	0,558	0,517	0,901
18.1	0,549	0,498	0,431	0,383	0,903
18.2	0,534	0,499	0,461	0,414	0,903
18.3	0,288	0,453	0,337	0,29	0,904
19.1	0,098	0,297	0,413	0,384	0,903
19.2	0,240	0,427	0,521	0,485	0,902
22.1	0,084	0,278	0,372	0,344	0,904
22.2	0,083	0,276	0,288	0,258	0,904
23	0,197	0,398	0,472	0,436	0,903
24	0,344	0,475	0,211	0,159	0,906
25.1	0,753	0,431	0,333	0,288	0,904
25.2	0,120	0,325	0,387	0,355	0,904
26.1	0,620	0,485	0,399	0,351	0,904
26.2	0,909	0,288	0,31	0,28	0,904
26.3	0,701	0,458	0,418	0,373	0,903
26.5	0,468	0,499	0,558	0,517	0,901
27	0,732	0,443	0,4	0,357	0,903
28	0,135	0,342	0,39	0,356	0,903

Table A.1: Descriptive statistics for the test items

Table A.2: Item fit of the 2PL model

Item	S_X2	df.S_X2	RMSEA.S_X2	p.S_X2	Item Fit
Q1.1	137,999	37	0,017	0,000	Excellent Fit
Q1.2	190,748	37	0,022	0,000	Excellent Fit
Q1.3	134,183	36	0,017	0,000	Excellent Fit
Q1.4	211,383	38	0,023	0,000	Excellent Fit
Q1.5	40,457	37	0,003	0,320	Excellent Fit
Q1.6	207,958	38	0,022	0,000	Excellent Fit
Q1.7	91,357	36	0,013	0,000	Excellent Fit
Q1.8	40,637	37	0,003	0,313	Excellent Fit
Q1.9	35,439	36	0,000	0,495	Excellent Fit
Q1.10	53,617	39	0,006	0,060	Excellent Fit
Q2	56,891	36	0,008	0,015	Excellent Fit
Q3	41,628	33	0,005	0,144	Excellent Fit
Q5	38,928	35	0,004	0,297	Excellent Fit
Q6	115,654	39	0,015	0,000	Excellent Fit
Q7	51,811	36	0,007	0,043	Excellent Fit
Q9	56,276	35	0,008	0,013	Excellent Fit
Q10.1	201,651	29	0,026	0,000	Excellent Fit
Q10.2	134,810	33	0,019	0,000	Excellent Fit
Q10.3	121,032	28	0,019	0,000	Excellent Fit
Q11	28,584	36	0,000	0,806	Excellent Fit
Q13	48,076	37	0,006	0,105	Excellent Fit
Q15.1	58,580	35	0,009	0,007	Excellent Fit
Q15.2	39,514	34	0,004	0,237	Excellent Fit
Q16	34,336	38	0,000	0,640	Excellent Fit
Q17.1	46,580	36	0,006	0,111	Excellent Fit
Q17.2	43,859	35	0,005	0,145	Excellent Fit
Q18.1	35,926	37	0,000	0,519	Excellent Fit
Q18.2	51,167	37	0,007	0,061	Excellent Fit
Q18.3	44,192	38	0,004	0,226	Excellent Fit
Q19.1	32,326	34	0,000	0,550	Excellent Fit
Q19.2	24,897	35	0,000	0,897	Excellent Fit
Q22.1	37,428	35	0,003	0,358	Excellent Fit
Q22.2	44,163	36	0,005	0,165	Excellent Fit
Q23	39,997	36	0,004	0,297	Excellent Fit
Q24	47,317	38	0,005	0,143	Excellent Fit
Q25.1	32,741	37	0,000	0,669	Excellent Fit
Q25.2	51,721	36	0,007	0,043	Excellent Fit
Q26.1	75,844	37	0,011	0,000	Excellent Fit
Q26.2	83,858	31	0,014	0,000	Excellent Fit
Q26.3	138,423	35	0,018	0,000	Excellent Fit
Q26.5	128,248	35	0,017	0,000	Excellent Fit
Q27	70,662	36	0,010	0,000	Excellent Fit
Q28	27,122	36	0,000	0,857	Excellent Fit

Appendix B

Assessment Information

Grade 6 Mathematics

Item No	Content Area	Concepts and skills	Cognitive Level	Type	Max
1.1	Numbers, Operations and Relations	Value/ Place value of digits to at least two decimal places.	K	MCQ	1
1.2	Numbers, Operations and Relations	Represent prime numbers to at least 100.	K	MCQ	1
1.3	Numbers, Operations and Relations	Recognize and use the associative property	K	MCQ	1
1.4	Numbers, Operations and Relations	Factors of 2-digit whole numbers	K	MCQ	1
1.5	Numbers, Operations and Relations	Count backwards or forwards in decimal fractions to at least two decimal places.	R	MCQ	1
1.6	Patterns, Functions and Algebra	Investigate and extend numeric patterns looking for rules of patterns involving a constant	C	MCQ	1
2.	Numbers, Operations and Relations	Represent numbers up to at least 9 digit numbers.	R	OEQ	1
3.	Numbers, Operations and Relations	Round off to the nearest 100 000	K	OEQ	1
4.1	Numbers, Operations and Relations	Addition of whole numbers with at least 6-digit number	R	OEQ	2
4.2	Numbers, Operations and Relations	Subtraction of whole numbers with at least 6-digit number	R	OEQ	2
4.3	Numbers, Operations and Relations	Multiplication of at least whole 4-digit by 2-digit numbers	R	OEQ	3
4.4	Numbers, Operations and Relations	Division of at least whole 4-digit by 2-digit numbers	R	OEQ	3
4.5	Numbers, Operations and Relations	Addition of mixed numbers where the denominators are the same.	R	OEQ	2
4.6	Numbers, Operations and Relations	Calculate a fraction of a number	R	OEQ	2
4.7	Numbers, Operations and Relations	Subtraction of mixed numbers where the denominators are the same	R	OEQ	2
4.8	Numbers, Operations and Relations	Subtraction of decimal fractions of at least two decimal places	R	OEQ	2
5.	Numbers, Operations and Relations	Multiple operations on whole numbers	P	OEQ	1
6.	Numbers, Operations and Relations	Arrange decimals - smallest to biggest	P	OEQ	1
7.	Numbers, Operations and Relations	Find multiples of 7	R	OEQ	1
8.	Numbers, Operations and Relations	Word problem involving division	K	OEQ	3
9.	Numbers, Operations and Relations	Solve number sentences by inspection	R	OEQ	1
1.10	Numbers, Operations and Relations	Matching equivalent fractions	K	OEQ	1
10.2	Numbers, Operations and Relations	Matching equivalent fractions	K	OEQ	1
10.3	Numbers, Operations and Relations	Matching equivalent fractions	K	OEQ	1
11.	Numbers, Operations and Relations	Represent numbers on a number line	R	OEQ	1
12.	Numbers, Operations and Relations	Word problem involving division	R	OEQ	2

Item No	Content Area	Concepts and skills	Cognitive Level	Type	Max

13.	Numbers, Operations and Relations	Solve number sentences by inspection	R	OEQ	1
14.	Patterns, Functions and Algebra	Determine input values, output values and rules for patterns and relationships using flow diagram	R	OEQ	2
15.1	Patterns, Functions and Algebra	Determine equivalence of different descriptions of the same relationship or rule presented a by a number sentence. Complete the table.	R	OEQ	1
15.2	Patterns, Functions and Algebra	Determine equivalence of different descriptions of the same relationship or rule presented a by a number sentence. Complete the table.	R	OEQ	1
16.	Patterns, Functions and Algebra	Investigate and extend geometric patterns looking for relationships or rules of patterns	C	OEQ	1
28.0	Patterns, Functions and Algebra	Compare and solve number sequences	P	OEQ	1
1.7	Shape	Draw lines of symmetry in 2-D shapes	K	MCQ	1
1.8	Shape	Link the position of viewer to views of geometric objects	K	MCQ	1
17.1	Shape	Recognise and sort and compare size of angles (-- acute, -- right, -- obtuse, -- reflex)	K	OEQ	1
17.2	Shape	Recognise and sort and compare size of angles (-- acute, -- right, -- obtuse, -- reflex)	K	OEQ	1
18.1	Shape	Recognise and name three 2-D shapes	K	OEQ	1
18.2	Shape	Recognise and name three 2-D shapes	K	OEQ	1
18.3	Shape	Recognise and name three 2-D shapes	K	OEQ	1
19.1	Shape	Similarities and Differences between rectangles, squares and parallelograms	K	OEQ	1
19.2	Shape	Similarities and Differences between rectangles, squares and parallelograms	K	OEQ	1
20.	Shape	Sort and compare 3-D objects in terms of: number of shapes, faces and number of vertices; number of edges	K	OEQ	3
21.	Shape	Word problems involving decimals	R	OEQ	2
1.1	Measurement	Practical measuring of temperature by comparing and ordering	R	MCQ	1
22.1	Measurement	Read time zone maps and calculating time differences based on time zones.	P	OEQ	1
22.2	Measurement	Read time zone maps and calculating time differences based on time zones.	P		1
23.	Measurement	Convert between units.	R	OEQ	1
24.	Measurement	Solve problem in context.	R	OEQ	1
25.1	Measurement	Problems involving mass	K	OEQ	1
25.2	Measurement	Problems involving mass	R	OEQ	1
1.9	Data Handling	Examine ungrouped numerical data to determine the median score in a data set	K	MCQ	1
26.1	Data Handling	Critically read and interpret data represented in pie graphs	R	OEQ	1
26.2	Data Handling	Critically read and interpret data represented in pie graphs.	R	OEQ	1
26.3	Data Handling	Critically read and interpret data represented in pie graphs.	R	OEQ	1
26.4	Data Handling	Critically read and interpret data represented in pie graphs.	R	OEQ	2
26.5	Data Handling	Critically read and interpret data represented in pie graphs.	R	OEQ	1
27.	Data Handling	Examine ungrouped numerical data to determine the most frequently occurring score in the data set (mode).	K	OEQ	1

Appendix C

Test Paper and Memorandum

basic education

Department:
Basic Education
REPUBLIC OF SOUTH AFRICA

MARKS	

ANNUAL NATIONAL ASSESSMENT 2014
GRADE 6 MATHEMATICS
TEST

MARKS: 75

TIME: 90 minutes

PROVINCE _____

DISTRICT _____

SCHOOL NAME _____

EMIS NUMBER (9 digits)

CLASS (e.g. 6A) _____

SURNAME _____

NAME _____

GENDER (✓)

BOY		GIRL	

DATE OF BIRTH

C	C	Y	Y	M	M	D	D

This test consists of 12 pages, excluding the cover page.

Instructions to the learner

1. Read all the instructions carefully.
2. Question 1 consists of 10 multiple-choice questions. You must circle the letter of the correct answer.
3. Answer Questions 2 to 28 in the spaces or frames provided.
4. All working must be shown on the question paper and must not be done on rough paper.
5. The test is out of 75 marks.
6. The test duration is 90 minutes.
7. The teacher will lead you through the practice question before you start the test.
8. The use of a calculator is not allowed.

Practice question

Circle the letter of the correct answer.

8 x 6 = _____

A. 48
B 84
C 72
D 60

You have done it correctly if you circled (A) above.

NB.

- You will answer more questions like the one you have just completed.
- Do your best to answer each question even if you are not sure of the answer.
- Write down the answer that you think is the best and move to the next question.
- When you have answered all the questions on a page, move to the next page.
- Look only at your own work.

The test starts on the next page.

1 Circle the letter of the correct answer.

1.1 What is the value of the underlined digit in 249,15?

A 5
B 0,5
C 50
D 0,05 (1)

1.2 What is the next prime number?
3 , 5 , 7 , _____

A 9
B 11
C 8
D 15 (1)

1.3 Fill in the missing number in 6 + 3 + 5 = ____ + 5.

A 9
B 11
C 8
D 15 (1)

1.4 Which number is not a factor of 96?

A 32
B 16
C 48
D 36 (1)

1.5 What are the missing numbers in the number sequence?
0,9 ; 0,7 ; 0,5 ; _____ ; _____ .

A 0,4 ; 0,3
B 0,03 ; 0,1
C 0,3 ; 0,01
D 0,3 ; 0,1 (1)

1.6 In which number sequence is the rule
(input number + 1) x 2 = output number used?

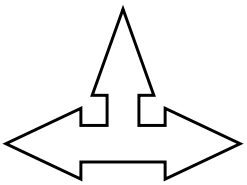
A 3 ; 7 ; 9 ; 11 ; 13

B 4 ; 10 ; 22 ; 46 ; 94

C 6 ; 9 ; 12 ; 15 ; 18

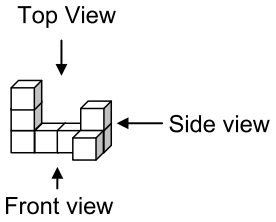
D 5 ; 8 ; 11 ; 14 ; 17 (1)

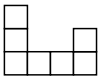
1.7 How many lines of symmetry can be drawn on the shape below?



A 1
B 3
C 5
D 2 (1)

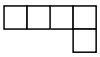
1.8 Which sketch represents the side view of the 3-D object?



A 

B 

C 

D 

(1)

1.9 What is the median of the given set of numbers?

3 5 4 4 5 6 9 8 4

A 8

B 5

C 3

D 4 (1)

1.10 The boiling point of pure water is ...

A 37 °C

B 0 °C

C 98 °C

D 100 °C (1)

[10]

2 Complete:

7 342 651 = (7 000 000) + (3 × 100 000) + (_____) + 2 000 + 651 [1]

3 Round off 59 673 to the nearest 10 000.

_____ [1]

4 Calculate the answers for QUESTIONS 4.1 to 4.8.

4.1 42 152 + 28 945 + 76 361 (2)

4.2 87 546 – 43 968

(2)

4.3 3 107 x 35

(3)

4.4 7 140 ÷ 15

(3)

4.5 $4\frac{3}{8} + 2\frac{1}{8}$

(2)

4.6 $\frac{2}{5}$ of 300 (2)

4.7 $5\frac{3}{5} - 2\frac{1}{5}$ (2)

4.8 59,3 – 25,8 (2)

[18]

5 Complete: (14 ÷ 2) + (51 – 48) = _____ [1]

6 Re-arrange the numbers from the smallest to the biggest.

4,5 , 4,3 , 4,01 , 4,8

_____ [1]

7 Write down the multiples of 7 between 21 and 56.

_____ [1]

8 If there are 8 sweets in a packet, how many packets can be filled with 947 sweets?

[3]

9 Complete: If 336 ÷ 14 = 24, then 24 x 14 = _____ [1]

10

$\frac{1}{4}$	75%	0,5

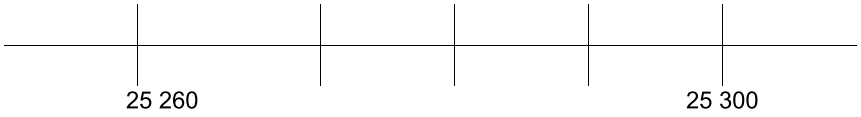
Match each of the three numbers given below with a number in the above frame.

10.1 $\frac{3}{4}$ = _____ (1)

10.2 50% = _____ (1)

10.3 0,25 = _____ (1)
[3]

11 Write down the number which is half-way between the two given numbers on the number line.



_____ [1]

12 Zonga received R240 for his labour. He received 12 times as much as Peter. How much did Peter get?

[2]

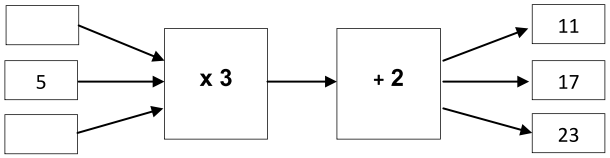
13 Fill in the missing number: 8 x 3 ÷ _____ = 1 [1]

14 Look at the input and output numbers and complete the table.

Input numbers	2	3	4	5	10	
Output numbers	5	8	11	14		44

[2]

15 Complete the flow diagram below.



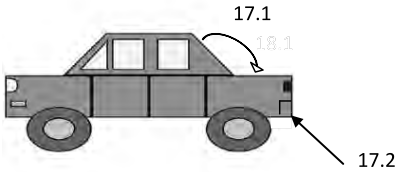
[2]

16 How many matches will there be in the next figure if the diagram pattern is continued?

_____

[1]

17 Name the different kinds of angles that are indicated by the arrows below.

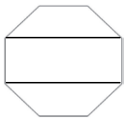


17.1 _____ (1)

17.2 _____ (1)

[2]

18 Name the THREE different 2-D shapes in the diagram.



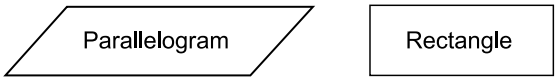
18.1 _____ (1)

18.2 _____ (1)

18.3 _____ (1)

[3]

19 Study the parallelogram and rectangle and complete the sentences below.

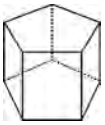


19.1 The _____ sides of a rectangle and parallelogram are equal in length. (1)

19.2 The parallelogram and rectangle each has _____ pairs of parallel sides. (1)

[2]

20 Complete the table.

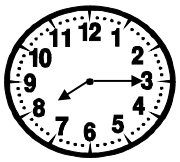
Pentagonal prism	
Number of vertices	
Number of edges	
Number of faces	

[3]

21 Mr Mololo's car uses 9,5 *litres* of petrol to drive to work. He found a shorter route where the car uses only 8,7 *litres* of petrol. How many litres of petrol does he save?

[2]

22 Study the clock faces showing the time in Rome and Tokyo.
Rome and Tokyo are in different time zones.

	
Rome Time in the morning (a.m.)	Tokyo Time in afternoon (p.m.)

22.1 Calculate the time difference between Rome and Tokyo.

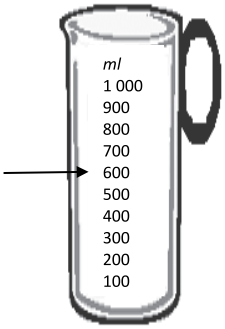
_____ (1)

22.2 If it is 17:00 in Tokyo, what time will it be in Rome?

_____ (1)

[2]

23 Convert the number of millilitres indicated on the jug to litres. _____



[1]

24 Below are the results in a school's final shot-put challenge.

Charles	3,95 *m*
Zola	429 *cm*
Conrad	4,08 *m*
Jabu	387 *cm*

Who threw the shot-put the furthest? _____ [1]

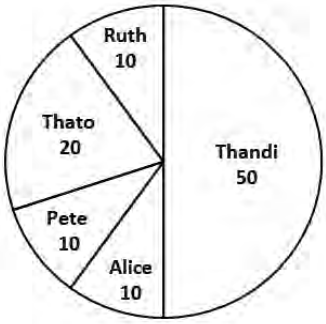
25 Use the kilogram scale below to answer the questions.

25.1 What is the mass indicated on the scale?_____ (1)

25.2 Convert the above mass to grams. _____ (1)

[2]

26 This pie chart shows how 100 marbles were shared amongst a group of children.



26.1 Who has the same number of marbles?

_____ (1)

26.2 Who received 20 marbles? _____ (1)

26.3 How many marbles do Alice and Thandi have together? _____ (1)

26.4 What percentage of the marbles did Pete get? _____ (2)

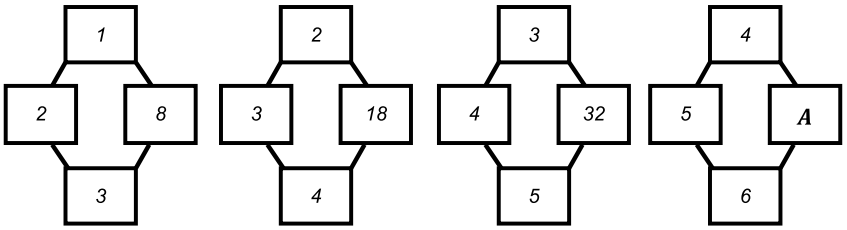
26.5 What fraction of the marbles did Thandi get? _____ (1)

[6]

27 What is the mode of the given set of test marks?

6 7 5 3 7 9 5 8 7 _____ [1]

28 What is the value of A in the fourth figure? A = _____



[1]

TOTAL: 75

basic education
Department:
Basic Education
REPUBLIC OF SOUTH AFRICA

ANNUAL NATIONAL ASSESSMENTS 2014
GRADE 6 MATHEMATICS
MEMORANDUM

MARKS: 75

This memorandum consists of 4 pages.

General marking note:
1. **Give full marks for answers only, unless otherwise stated.**
2. **Accept any alternative correct solution that is not included in the memorandum.**
3. **CA refers to consistency accuracy. See Question 4.3 as an example.**

QUESTION		EXPECTED ANSWER		CLARIFICATION	MARK	TOTAL
1	1.1	D ✓			1	
	1.2	B ✓			1	
	1.3	A ✓			1	
	1.4	D ✓			1	
	1.5	D ✓			1	
	1.6	B ✓			1	
	1.7	A ✓			1	
	1.8	C ✓			1	
	1.9	B ✓			1	
	1.10	D ✓			1	10
2		4 x 10 000 or 40 000 or 4 x 10^4 or forty thousand ✓		Any of the given options : 1 mark		1
3		60 000 ✓		60 000: 1 mark		1
4	4.1	42 152 28 945 +76 361 147 458 ✓ ✓	If the answer is wrong the learner will be credited with one mark if he has added the units, tens and hundreds correctly.	All digits correct 147 458: 2 marks. Digits 458: 1 mark Digits 147: 1 mark	2	
	4.2	87 546 − 43 968 43 578 ✓✓	If the answer is wrong the learner will be credited with one mark if he has subtracted the units, tens and hundreds correctly.	All digits correct 43 578: 2 marks. Digits 578: 1 mark Digits 43: 1 mark	2	

	4.3	3 107 x 35 15 535 ✓ 93 210 ✓ 108 745 ✓ Example of CA 3 107 x 35 12 532 x (incorrect no mark) 93 210 ✓ 105 742 ✓ (added correctly) 3107 x 35 12 532 x (incorrect no mark) 83 210 x (incorrect no mark) 95 742 ✓(added correctly) or 3 107 x 35 = 3 107 x 5 x 7 = 15 535 x 7 = 108 745 or 3 107 x 35 = 3 107 x 7 x 5 = 21 749 x 5 = 108 745	Answer only: 3 marks 3107 x 5 =15 535: 1 mark 3107 x 30 =93 210: 1 mark 15 535 + 93 210 = 108 745: 1 mark	3	
	4.4	476 ✓ 15 ⟌ 7140 − 60 ✓ 114 − 105 ✓ 90 − 90 or 7 140 ÷ 15 = 7140 ÷ 5 ÷ 3 = 1 428 ÷ 3 = 476 or 7 140 ÷ 15 = 7 140 ÷ 3 ÷ 5 = 2 380 ÷ 5 = 476	Answer only: 3 marks 60: 1 mark 105: 1 mark Apply CA	3	
	4.5	$4\frac{3}{8}+2\frac{1}{8}$ or $4\frac{3}{8}+2\frac{1}{8}$ $= 4 + \frac{3}{8} + 2 + \frac{1}{8}$ $= \frac{35}{8} + \frac{17}{8}$ ✓ $= 6 + \frac{4}{8}$ ✓✓ $= \frac{52}{8}$ ✓ $= 6\frac{1}{2}$ $=6\frac{1}{2}$ Do not penalize $6\frac{4}{8}$ or $6\frac{2}{4}$ or $\frac{52}{8}$ ✓✓	Answer only : 2 marks 6: 1 mark $\frac{4}{8}$: 1 mark $6\frac{1}{2}$: 2 marks	2	
	4.6	$\frac{2}{5}$ of 300 or $\frac{2}{5}$ of 300 = 300 ÷ 5 ×2 ✓ = 2 x 60 ✓ (because 300 ÷ 5 = 60) = 120 ✓ = 120 ✓	120: 2 marks Calculation: 1 mark	2	

	4.7	$5\frac{3}{5} - 2\frac{1}{5}$ or $5\frac{3}{5} - 2\frac{1}{5}$ $= 3\frac{2}{5}$ ✓✓ $= 5 + \frac{3}{5} - 2 - \frac{1}{5}$ $= 5 - 2 + \frac{3}{5} - \frac{1}{5}$ $= 3\frac{2}{5}$ or $\frac{17}{5}$ ✓✓	Answer : 2 marks 3: 1 mark $\frac{2}{5}$: 1 mark	2	
	4.8	59,3 or 59,3 – 25,8 = 33,5 ✓✓ – 25,8 33,5 ✓✓	33,5 : 2 marks 33 : 1 mark 0,5 : 1 mark	2	18
5		(14 ÷ 2) + (51 – 48) =10 ✓	10 : 1 mark		1
6		4,01 , 4,3 , 4,5 , 4,8✓	1 mark : correct order / sequence		1
7		28, 35, 42, 49 ✓	28, 35, 42, 49 : 1 mark		1
8		Number of packets = 947 ÷ 8 ✓ = 118 r 3 ✓ 11 8 r 3 8⟌947 −8 14 − 8 67 − 64 ∴ Number of packets needed = 118 ✓	118: 3 marks 947 ÷ 8: 1 mark 118 r 3: 1 mark		3
9		336✓	336 : 1 mark		1
10	10.1	75% ✓	75%: 1 mark	1	
	10.2	0,5✓	0,5 : 1 mark	1	
	10.3	$\frac{1}{4}$ ✓	$\frac{1}{4}$: 1 mark	1	3
11		25 280 ✓	25 280: 1 mark		1
12		Peter's amount = R240 ÷ 12 ✓ = R20 ✓	R20: 2 marks R240 ÷ 12: 1 mark		2
13		24 ✓	24: 1 mark		1
14		Input : 15 ✓ Output : 29 ✓	15: 1 mark 29: 1 mark		2
15	15.1	Input : 3 ✓	3: 1 mark	1	
	15.2	Input : 7 ✓	7: 1 mark	1	2
16		13 matches✓	13 : 1 mark		1
17	17.1	Obtuse ✓	1 mark	1	
	17.2	Right angle or reflex angle ✓	1 mark	1	2

18	18.1	Octagon ; Trapezium ; Rectangle or Hexagon ✓✓✓	Any three answers 1 mark each.	1	
	18.2			1	
	18.3			1	3
19	19.1	opposite ✓	1 mark	1	
	19.2	two ✓	1 mark	1	2
20		Number of vertices : 10 ✓ Number of edges : 15 ✓ Number of faces : 7 ✓	1 mark each.		3
21		No. of litres saved = 9,5 - 8,7 ✓ = 0,8 ✓	0,8: 2 marks 9,5 – 8,7 : 1 mark		2
22	22.1	9 hours ✓	9 hours: 1 mark	1	
	22.2	8.00 a.m. ✓or 08:00 or 8 o'clock	1 mark	1	2
23		600 ml = 0,6 l ✓	0,6 l : 1 mark		1
24		Zola ✓	Zola : 1 mark		1
25	25.1	56,8 kg ✓	56,8 kg : 1 mark	1	
	25.2	56 800 g ✓	56 800 g: 1 mark	1	2
26	26.1	Pete, Alice and Ruth ✓	Must write all three names : 1 mark	1	
	26.2	Thato ✓	1 mark	1	
	26.3	Number of marbles = 10 + 50 = 60 ✓	60 : 1 mark	1	
	26.4	Pete's % = $\frac{10}{100}$ ✓x 100 = 10 ✓	10 : 2 marks $\frac{10}{100}$: 1 mark	2	
	26.5	Fraction = $\frac{50}{100}$ or $\frac{5}{10}$ or $\frac{1}{2}$ ✓	$\frac{1}{2}$: 1 mark.	1	6
27		Mode = 7 ✓	7: 1 mark.		1
28		A = 50✓	Fig 1: 1 x 2 + 2 x 3 = 8 Fig 2: 2 x 3 + 3 x 4 = 18 Fig 3: 3 x 4 + 4 x 5 = 32 Fig 4: 4 x 5 + 5 x 6 = 50		1
				TOTAL	**75**

Appendix D

Differential Item Functioning Results

Table D.1 provides the results of the test for DIF items. At 5% significance, items 1.2, 1.7, 1.8, 2, 6, 13, 16, 22.1, 22.2, 23, 25.2, 26.1, 26.2, and 28 were found to function differentially for learners from different provinces.

Table D.1: Likelihood-ratio test of differential item functioning

Item	Converged	AIC	SABIC	HQ	BIC	X2	df	p-value	Adjusted p-value
Q1.1	TRUE	13.982	71.894	51.564	122.738	18.018	16	0.323	0.448
Q1.2	TRUE	-21.596	36.316	15.986	87.160	53.596	16	0.000	0.000*
Q1.3	TRUE	26.532	84.444	64.114	135.288	5.468	16	0.993	1.000
Q1.4	TRUE	20.204	78.116	57.786	128.960	11.796	16	0.758	0.988
Q1.5	TRUE	38.830	96.742	76.412	147.586	-6.830	16	1.000	1.000
Q1.6	TRUE	11.228	69.140	48.810	119.984	20.772	16	0.187	0.288
Q1.7	TRUE	-8.744	49.168	28.838	100.012	40.744	16	0.001	0.001*
Q1.8	TRUE	-35.580	22.332	2.002	73.176	67.580	16	0.000	0.000*
Q1.9	TRUE	12.538	70.450	50.120	121.294	19.462	16	0.245	0.364
Q1.10	TRUE	15.879	73.791	53.461	124.635	16.121	16	0.445	0.597
Q2	TRUE	-8.766	49.146	28.816	99.990	40.766	16	0.001	0.001*
Q6	TRUE	-8.156	49.756	29.427	100.600	40.156	16	0.001	0.001*
Q7	TRUE	33.444	91.356	71.027	142.200	-1.444	16	1.000	1.000
Q11	TRUE	55.589	113.501	93.171	164.345	-23.589	16	1.000	1.000
Q13	TRUE	-3.604	54.308	33.978	105.152	35.604	16	0.003	0.006*
Q15.1	TRUE	38.011	95.923	75.593	146.767	-6.011	16	1.000	1.000
Q16	TRUE	0.192	58.104	37.774	108.948	31.808	16	0.011	0.018*
Q17.1	TRUE	5.270	63.182	42.852	114.026	26.730	16	0.045	0.071
Q18.1	TRUE	77.049	134.961	114.631	185.805	-45.049	16	1.000	1.000
Q18.2	TRUE	38.828	96.740	76.410	147.584	-6.828	16	1.000	1.000
Q18.3	TRUE	59.509	117.421	97.091	168.265	-27.509	16	1.000	1.000
Q22.1	TRUE	-84.377	-26.465	-46.795	24.379	116.377	16	0.000	0.000*
Q22.2	TRUE	-36.691	21.221	0.891	72.065	68.691	16	0.000	0.000*
Q23	TRUE	-14.890	43.022	22.692	93.866	46.890	16	0.000	0.000*
Q24	TRUE	43.759	101.671	81.341	152.515	-11.759	16	1.000	1.000
Q25.1	TRUE	13.927	71.839	51.509	122.683	18.073	16	0.320	0.448
Q25.2	TRUE	-23.848	34.064	13.734	84.908	55.848	16	0.000	0.000*
Q26.1	TRUE	-15.463	42.449	22.119	93.293	47.463	16	0.000	0.000*
Q26.2	TRUE	-10.659	47.253	26.923	98.097	42.659	16	0.000	0.001*
Q26.3	TRUE	23.053	80.965	60.635	131.809	8.947	16	0.916	1.000
Q27	TRUE	4.207	62.119	41.789	112.963	27.793	16	0.033	0.055
Q28	TRUE	-69.095	-11.183	-31.513	39.661	101.095	16	0.000	0.000*