Language Learning Anxiety: A Biometric Investigation of Stress and Language Learning using Wearable Devices

A thesis submitted in fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

of

RHODES UNIVERSITY

(HUMANITIES)

by

William Tait MacDonald

December 2021

ABSTRACT

As Sapolsky (2015) notes, stress research has long been characterised by definitional debates that gave rise to a lack of agreement between theorists. An example of this definitional confusion can be seen in the area of Language Learning Anxiety, where there appears to be confusion in the literature over terms such as anxiety and stress. This study investigated the link between stress and language learning using wearable devices measuring heart rate variance, as a means of establishing the feasibility of using this new technology in stress research. The results indicate that the contextualised longitudinal data delivered by wearable devices mitigates against the current dominant paradigm in Language Learning Anxiety, which postulates a straight-line negative correlation between stress and learning. Instead, the inverted U stress relationship proposed by theorists such as Hebb (1955) seem to be a better fit for the data.

The nature of the contextualised data generated in this study allowed for comparisons between participants' stress readings in academic contexts, such as language and non-language classes, and their free time. The findings suggest that certain long-held assumptions about heightened stress in academic contexts may not hold true. While the findings of this study did not reach the levels of statistical significance, they constitute proof of concept that the type of contextualised data delivered by wearable devices may allow for a new type of stress research that incorporates contextualising longitudinal perspectives on participants' stress levels. In this study the inclusion of contextualising data led to fundamentally different conclusions about the relationship between stress and language learning. The same may be true of many areas of stress research. The findings presented in this study have broader paradigm-altering implications not only for educational policy, but also for stress research in general.

Perhaps equally important was that the type of data delivered by wearable devices was qualitatively different from that normally associated with quantitative studies. This presented challenges in data analysis in this study, but also opens intriguing possibilities regarding a means of reconciling the qualitative and quantitative split in research modalities.

The use of wearable devices is not without issues, and some of the issues, ranging from practical considerations to ethical conundrums, are presented for the reader's consideration and to inform future researchers regarding potential pitfalls.

Keywords: Language Learning Anxiety, LLA, Stress, Anxiety, Wearable Devices, Heart Rate Variance, HRV.

Table of Contents

List of Tables

Table of Figures

Acknowledgements

My most profound thanks to all those who provided help and support during this long process. I would like to acknowledge the special contributions of some people without whom I would never have completed this thesis:

To Professor Jacqueline Akhurst and Professor Charles Young, my thanks for your endless support and gentle guidance which led me through this complex and often frustrating topic with both kindness and sensitivity.

To my darling wife, Bronwen, and my beloved children, Maria and Kai, without whose patience, love, and support I could never have completed this long journey.

To my parents-in-law, Barry and Joan Wheeler, who provided support and encouragement. It meant more than words can easily express.

To my mother, Harriet MacDonald, and my brother and sister, John and Sarah, whose comments and suggestions helped to untangle the complex issues surround this topic.

To my research participants, without whom this research would have been impossible.

To the University of Nagasaki, and all my colleagues there, who were understanding of the rigors of writing a doctoral thesis and provided help and support in so many ways.

Finally, to my dearly departed father, Dr. John MacDonald; your guidance and wisdom was most sorely missed during this journey.

Note with respect to funding:

This Ph.D. was completed with financial assistance from:

The University of Nagasaki (長崎県立大学)

Rhodes University

The Wheeler family

The views presented and conclusions reached are those of the author and are not necessarily those of the University of Nagasaki, Rhodes University, or the Wheeler family.

Chapter 1: Introduction

The theory of Language Learning Anxiety (LLA) proposes that a specific stressor exists in the language learning environment that has a substantial effect on language learning. This theory has attracted considerable scholarship, with the search term “language learning anxiety” producing a list of 4,940 academic papers on the topic on *Google Scholar*. As a rough metric of the current importance of this concept within academia, 2,280 of those papers were written in the past five years (from 2017 to 2021), indicating that this theory is currently the subject of a substantial amount of scholarship.

The theory of LLA that is currently most popular was proposed in Horwitz (1986), and Horwitz, et al. (1986); and accompanied by the Foreign Language Classroom Anxiety Scale (FLCAS), which theorises a straight-line negative correlation between anxiety and language learning. According to current LLA theory, as stress increases so learning decreases.

Classical theories of stress and performance such as Hebb (1955) propose a different relationship between stress and performance, namely an inverted U pattern, similar to a bell curve. In the Hebbian (1955) model the relationship between stress and performance is not uniform, with peak learning occurring at moderate stress, and declining as stress increases or decreases. This pattern agrees with current research in neurology, as contained in the works of theorists such as Sapolsky (2015). Resolving this apparent disagreement is of importance and interest to not only the field of stress research, but also to educators and learners.

This difference in the theorised relationship between stress and learning is what motivated this study as it has implications for language teaching. The researcher observed first-hand that LLA theory was being used to justify less challenging material being presented in classes in the belief that this would reduce stress and anxiety, and result in enhanced learning. This prompted the researcher to start investigating the question of learning and stress.

A fuller discussion of the history of LLA theory, and stress theories such as Hebb (1955) can be found in Chapter 2. The historical context is critical in this study as definitions of stress and anxiety have shifted as debates around the topic have evolved. An appreciation for how this debate has unfolded, and the context in which the concept of LLA arose, is important in an appropriate and balanced approach to this topic.

In researching the question of stress and learning, a reliable means of measuring stress was sought, and in addition to the General Health Questionnaire 28 item version (hereafter GHQ-28) and a research diary, this research employed the *Garmin Vivosmart 3*

and 4 wearable devices, which measures stress using heart-rate variance (HRV). A search for literature on similar studies using wearable devices in language learning research was conducted using *Google Scholar*, and none could be found as of November 2021. Therefore, this research may represent one of the first studies using wearable devices for this application and may be of interest to other researchers attempting similar research using this new technology to research questions in stress and education.

Chapter 3 presents a discussion of the measurement of stress and the research supporting the use of wearable devices for measuring stress, while Chapter 4 introduces the methodology for the study. The quality and quantity of data delivered by the wearable devices necessitated a mixed methods approach in this study. Multiple case studies were used to explore the detailed, almost narrative, daily data, with statistical analysis to further explore trends seen in the case studies.

As wearable devices are a relatively new technology they present several unknowns regarding ethical issues, and so in a departure from traditional structure ethical issues are presented separately in Chapter 5 to allow a more complete discussion of the associated issues.

The quantity and quality of the data obtained from the wearable devices necessitated two different forms of analysis. Chapter 6 presents detailed data in a series of case studies, allowing a more fine-grained investigation of the qualitative aspects of the data gathered in this study. Chapter 7 focuses on the quantitative aspects of the data using statistical methods of analysis, more traditionally associated with numerical data.

In Chapter 8 the findings from Chapter 6 and 7 are discussed, with a view towards reintegrating the two different forms of analysis and exploring what conclusions the evidence may suggest. Chapter 9 explores the implications of these suggestions for education, stress research, and future research using wearable devices, with a particular view towards highlighting potential future avenues for research.

Further supporting information can be found in the appendices, with Appendix A containing an annotated copy of the research diary used in this research, Appendix B containing the participant briefing information, Appendix C containing copies of correspondence relating to the handling of data by Garmin, and Appendix D containing information relating to ethical clearances.

Chapter 2: Review of the Literature on Language Learning Anxiety

Language Learning Anxiety (LLA), also referred to a Foreign Language Learning Anxiety, is central to thesis, so this chapter will structure itself around LLA theory, presenting the core concepts, and then critiquing them with reference to other theories and models.

Using this approach, it is hoped that the reader will be able to consider the theoretical basis for this thesis and understand the validity of the challenges being raised to LLA theory.

2.1 The Beginnings of LLA Theory

In 1947 at the University of Chicago in the United States of America (USA), Dunkel's research with 46 undergraduate students who had studied two years of Latin at high school observed a disparity between the performance predicted by tests of language learning ability and students' performance in Latin tests. Dunkel (1947) theorised that there must be other personality variables at play.

To identify these, Dunkel administered a group Rorschach test and found correlations between the test's sub-groups and the students' test results. One finding was that a tendency towards anxious thinking correlated with lower scores in Latin than predicted by language learning ability tests. This seems to mark the genesis of the concept of language learning anxiety (LLA), and over the intervening decades the topic has attracted a great deal of contradictory research.

In 1953 Montague investigated anxiety in language learning with 120 undergraduate students at the University of Iowa in the USA, split into 3 anxious groups and 3 non-anxious group using Taylor's Manifest Anxiety Scale. The students attempted to learn a list of words, and Montague (1953) found that the anxious students did less well on difficult lists, but better on less complex lists, and improved more quickly.

In 1955 Taylor and Chapman used Taylor's Manifest Anxiety Scale to split a group of 34 students at the Northwestern University in the USA into two groups, who scored either extremely high or low on the anxiety scale. Taylor and Chapman (1955) gave them the task of learning eight pairs of nonsense syllables. The high anxiety group did better than the low anxiety group.

Time and further experimentation did not seem to bring clarity to this phenomenon. Chastain (1975) used the Sarason Test Anxiety Scale and Taylor's Manifest Anxiety Scale to identify anxiety levels in 229 university students studying introductory courses in French, German, and Spanish at the University of Virginia, USA. Compared their final grades in the

courses to their anxiety levels, Chastain (1975) found inconsistent correlations with anxiety, which was negative in the French group, but positive in the German and Spanish groups.

Backman's (1976) research with 21 Venezuelan students studying English at the University of Boston in the USA, used an interview method and Gardener's 1974 attitude scales to assess anxiety. Backman (1976) found that anxiety only served as a discriminating factor in extreme cases (the two highest and two lowest scoring students).

When considering Dunkel's (1947) study, the mixed findings seem hardly surprising. The foundational concepts of LLA theory are based on Dunkel's (1947) research using the group Rorschach test. The group Rorschach test is known to be unreliable for this type of application (Dawes, 1991). Indeed Dawes (1991) doubts the validity and reliability of the Rorschach test in general, citing, "consistent research findings — of literally thousands of published studies — that the Rorschach interpretation is unreliable and invalid" (p. 1).

The concept of LLA appears to falter at the very first step, with someone without the proper training using a psychometric device for a purpose other than what was intended. This displays a lack of awareness of the precision and sensitivity of psychometric devices. Measuring something physical is generally easily done in a few seconds with a tape measure or a set of scales. However, psychometrics measures things that are not directly observable and whose existence may be inferred from other observable behaviours. For example, measuring intelligence is extremely difficult, but if someone does well at school, then they might be considered intelligent. However, this very specific definition of the word intelligence only applies in specific contexts, and for specific purposes. The tester needs to bear in mind the limitations of the test, and a whole host of caveats about complicating factors such as socio-economic status and specific learning styles. There are also limitations on who the test can be used for, and under what conditions (Furr, 2017).

Psychometric tests are complex devices that require proper training, an awareness of how they are constructed, and their precise uses and limitations to use properly. The misuse or misinterpretation of psychometric tests can cause serious and long-lasting harm, and they should only be used by those with the proper training (Furr, 2017).

This theme of the misuse or misinterpretation of psychometric tests by those who incompletely understand their specific uses and limitations will become a recurring theme in the history of LLA theory. That later researchers failed to replicate Dunkel's (1947) results should probably have been the end of Dunkel's theory.

2.2 Continuing Pursuit of LLA Theory

Despite Dunkel's (1947) results having failed the test of replicability LLA theory persisted. Scovel (1978) presented a systematic review of the status of research into LLA. In this review the results of thirty papers by various authors are discussed, finding mixed results. Scovel (1978) theorised that the key must lie in the difference between debilitating and facilitating anxiety. The notion of facilitating anxiety is based on work by authors such as Kleinmann (1977), who attempted to establish whether anxiety might help students to do better.

In Kleinmann (1977), Spanish and Arabic speaking English second language (ESL) students at the University of Pittsburgh in the USA were asked to respond to questions such as, "Nervousness while using English helps me" (in Scovel, 1978, p. 133), in order to investigate whether anxiety led students to attempt more difficult grammar that was not normally found in their mother tongue. Kleinmann (1977) suggested that anxiety might be a facilitating factor in learning. However, the evidence was mixed and Kleinmann (1977) concluded that facilitative effects seem to be dependent on individual learner variables.

However other research reviewed by Scovel reached the opposite conclusion, that anxiety impeded learning, and was debilitating. In Gaudry and Fitzgerald's (1971, in Scovel, 1978) research with 7th grade students in twelve Australian schools, higher levels of anxiety were associated with lower levels of performance for all but the best performing students. Scovel's (1978) conclusions at the end of the paper state:

> The bad news is that the deeper we delve into the phenomenon of language learning, the more complex the identification of particular variables becomes. As this paper has suggested, before we begin to measure anxiety, we must become more cognizant of the intricate hierarchy of learner variables that intervene: the intrinsic/extrinsic factors, the affective/cognitive variables, and then the various measures of anxiety and their relationship to these other factors (p. 140).

Scovel's (1978) hypothesis about facilitating versus debilitating anxiety touches on a difficult topic in the field of anxiety research, namely that of definitions.

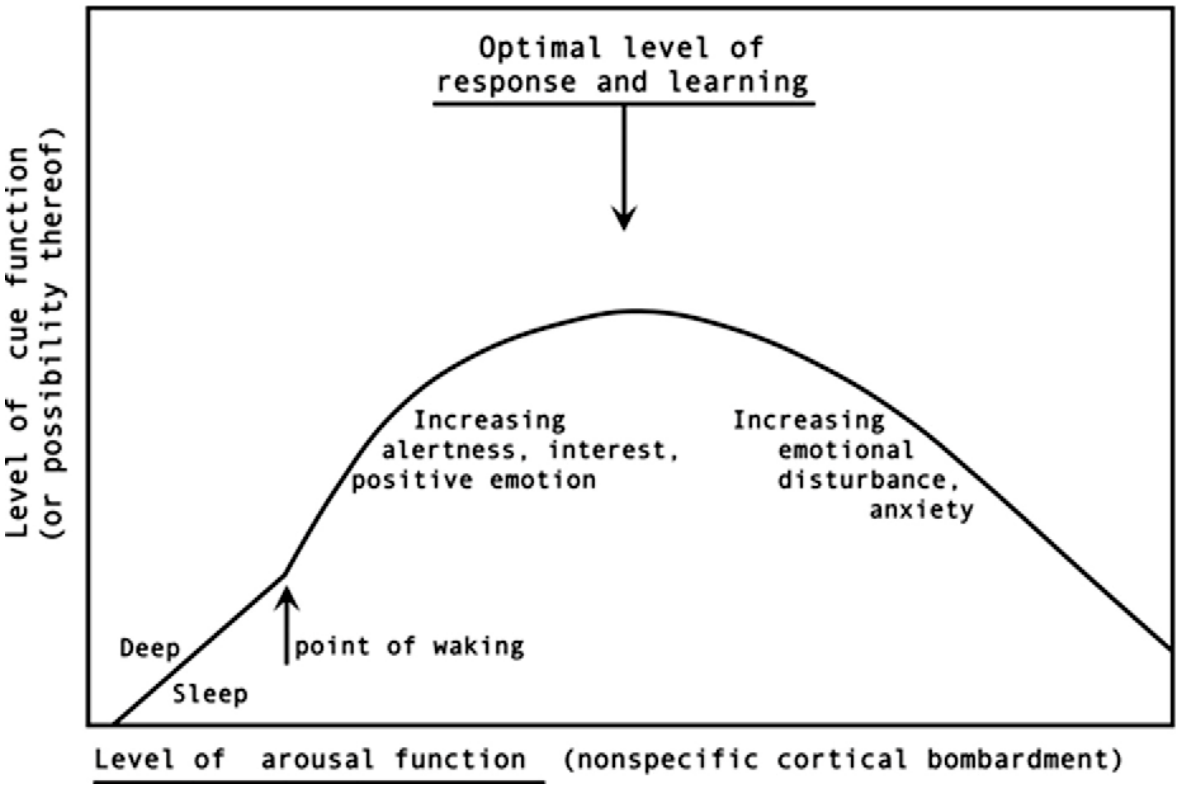
Classical depictions of arousal and resultant anxiety appear quite simple. Hebb's (1955) model is shown in Figure 1. It depicts a neat bell-shaped curve, with arousal rising from deep sleep until waking, then increasing functioning up to a point of optimal arousal, and then declining functioning thereafter. This appears to be a straight-forward homeostatic curve.

Arousal and level of functioning (performance), or the potential for performance, increase until arousal levels reach an optimal point. Thereafter further arousal above the optimum level results in a decline in performance as the individual becomes overstimulated and seeks to avoid or withdraw from the stimulus. In Hebb's model anxiety, a form of avoidance behaviour, results directly from overstimulation, and is always debilitative (Hebb, 1955).

However, despite the apparent simplicity Hebb's model begs serious questions, such as how one can determine the optimal level of arousal for an individual, and whether all arousal ultimately leads to declining levels of functioning once optimal arousal is exceeded.

Figure 1

The Hebbian Model



Note. The original version of this figure was blurred and has been recreated here. Any errors or modifications are unintentional. Adapted from "Drives and the CNS (conceptual nervous system)", (Hebb, 1955).

What is important in the Hebbian model is the word increasing, which indicates that anxiety occurs on a continuum, and that experiencing a degree of anxiety in some contexts is natural and normal, and that this only becomes of clinical concern when the individual experiences unusually high levels of anxiety or exhibits an inability to recover from the anxiety (Dwight, et al., 2005). The point at which anxiety crosses the barrier between normal

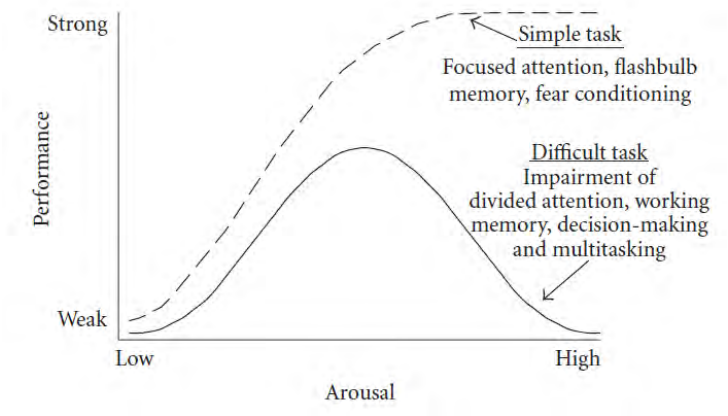
anxiety and clinically significant anxiety is referred to as “caseness” (Public Health England, 2021), the point at which it becomes a clinical case.

There are therefore two overlapping continuums, one for arousal, and then a second for anxiety, with the anxiety continuum beginning as arousal levels peak and performance begins to decline.

Hebb’s (1955) model is remarkably similar to the earlier and more famous Yerkes-Dodson (1908) model and is shown in Figure 2.

Figure 2

The Yerkes-Dodson Model



Note. From “The Temporal Dynamics Model of Emotional Memory Processing: A Synthesis on the Neurobiological Basis of Stress-Induced Amnesia, Flashbulb and Traumatic Memories, and the Yerkes-Dodson Law”, (Diamond, et al., 2007). Copyright 2007 Diamond, et al. Reprinted with permission.

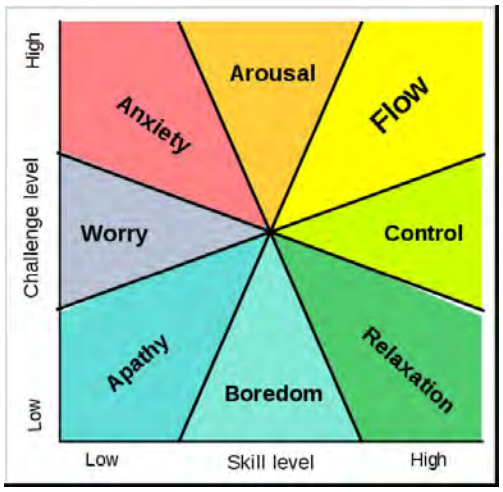
Yerkes-Dodson’s (1908) model does not make explicit reference to anxiety though, and there is sometimes confusion between Hebb’s (1955) less well-known model, and the earlier and more famous Yerkes-Dodson model. Yerkes-Dodson’s model is important in that it adds task complexity, and when combined with Hebb’s model, explains why facilitative anxiety may be more prevalent in simpler tasks where performance continues to strengthen as arousal rises, while facilitative anxiety may be less prevalent in more difficult tasks, where performance reaches a peak and then begins to decline.

Without wishing to in any way denigrate the degree of skill involved in sports, the Yerkes-Dodson model may explain why much of the evidence for, and some of the most ardent proponents of, facilitative anxiety may be found in the area of sports psychology (Polman & Borkoles, 2011).

Continuing this investigation of the interaction between task difficulty and arousal, one should also consider the work of Csíkszentmihály (1997).

Figure 3

Csikszentmihalyi's Flow Theory Model



Note. The following figure is a colourised adaptation of Csíkszentmihályi's Flow Theory Model from "Flow and Education" by M. Csikszentmihalyi, 1997, *NAMTA Journal*, 22, p. 31. This figure is taken from **https://en.wikipedia.org/wiki/File:Challenge_vs_skill.svg** and is a public domain image published without copyright under Wikimedia commons.

Csikszentmihalyi's (1997) model depicts the interaction between the individual's skill level, the difficulty of the challenge, and the emotional response. Csikszentmihalyi was interested in self-motivated activities that minimised traditional rewards, intrinsically motivated activities, and what role challenge level and emotion played in performance.

The research started with a series of 60 unstructured pilot interviews with people who were champions or world-record holders in a range of sports in the USA. From this a questionnaire and more structured interview were administered to 173 individuals in various locations across the USA ranging from experts to beginners, and including a wider range of activities, including 53 chess players, 22 composers, and 28 modern dancers, with the remaining individuals from sports such as rock climbing and basketball. The model presented in Figure 3 was derived from this research (Csikszentmihalyi, 1975).

The top row of the figure depicts a high challenge and seems similar to the centre to right side of the Hebbian (1955) model, with high skill learners achieving peak performance (which Csikszentmihalyi labels as flow), medium skill learners being stressed, but not necessarily to the point of a significant decline in performance, and anxiety in learners with low skill levels where their performance begins to decline significantly.

In Csikszentmihalyi's (1997) model of the three possible results at high challenge levels one is positive (flow), one is neutral (arousal/stress, but without significant impairment), and one is negative (anxiety). Contrasting with this, moderate challenges are evenly split between control (positive) and worry (worry), and low challenge outcomes are largely negative, such as apathy or boredom, or at best neutral in the case of relaxation for high skill learners faced with low difficulty challenges. This may be desirable in an educational context from time to time, but relaxation hardly seems conducive to learning.

In the paragraph above the terms stress and arousal have been conflated. This use of the terms interchangeably is common in the literature (Contrada & Baum, 2011). This is not to say that there are not important differences between the two terms, as shown in the APA's Dictionary of Psychology (Van den Bos, 2007), with the definitions of stress and arousal presented below.

> arousal
>
> n.
>
> 1. a state of physiological activation or cortical responsiveness, associated with sensory stimulation and activation of fibres from the reticular activating system.
>
> 2. a state of excitement or energy expenditure linked to an emotion. Usually, arousal is closely related to a person's appraisal of the significance of an event or to the physical intensity of a stimulus. Arousal can either facilitate or debilitate performance. See also catastrophe theory. (Van den Bos, 2007, para. 1)

> stress
>
> n.
>
> 1. the physiological or psychological response to internal or external stressors. Stress involves changes affecting nearly every system of the body, influencing how people feel and behave. For example, it may be manifested by palpitations, sweating, dry mouth, shortness of breath, fidgeting, accelerated speech, augmentation of negative emotions (if already being experienced), and longer duration of stress fatigue. Severe stress is manifested by the general adaptation syndrome. By causing these mind–body changes, stress contributes directly to psychological and physiological disorder and disease and affects mental and physical health, reducing quality of life. (Van den Bos, 2007, para. 1)

Examining the two definitions offered above, stress could be characterised as a subset of arousal, with the key difference between the two being whether the experience is perceived as positive or negative. The definition of arousal seems more positive, referring to cortical responsiveness, excitement, and both facilitative and debilitative effects.

By contrast the definition of stress contains numerous references to unpleasant physical symptoms and negative emotions and has no reference to any facilitative aspects. In common parlance the word stress generally has negative connotations. Given the list of symptoms it is understandable how confusion might arise between where the line is between stress and anxiety, and if there is indeed a clear line at all.

However, what the definition offered above fails to mention is that stress can be adaptive and facilitative (Contrada & Baum, 2011). For example, consider that without the stressor of bills to pay many people would not get out of bed in the morning. One of the key issues therefore seems to be emotional valence. This issue is far from simple though.

For example, a rollercoaster ride is an objectively physically and psychologically stressful event, associated with high levels physical discomfort, arousal, and fear. The rollercoaster is designed to elicit these reactions. Yet, despite the objectively very high levels of stress experienced during the rollercoaster ride there are many people who report the experience as enjoyable, positive, and wish to repeat the experience.

It follows that one of the key differences between stress and arousal rests on a very tenuous thread, namely the emotional valence attached to an experience. Yet this emotional valence changes over time. In the context of research this raises some thorny issues. When someone on a rollercoaster can report the experience as stressful, yet seconds later after the ride ends reports the experience as arousing, it points to fundamental problems in the definitions being offered.

There are many other more technical differences between the terms stress and arousal. Arousal is used in medicine as a general term referring to “responsiveness to stimuli” (Webster, 2016, para. 1). This term is applied in a wide range of contexts, being used to describe the strength of emotional responses, emotional arousal, to the level of activity in specific portions of the brain, neurological arousal. There is also the most commonly understood meaning, sexual arousal.

While these uses of the term arousal all share common features, the term is very broad and used in different areas in different ways. There seems to be some confusion over the precise definition of arousal, with Pfaff (2006) writing, “Everyone knew that arousal exists, for example. It is intuitively obvious, and it is absolutely necessary to explain

neurobiologic data. But what exactly is arousal?", (p. 5). Pfaff (2006) goes on to offer their own definition, but the debates about the precise meaning of the term is clear and a consistent thread in the literature (Contrada & Baum, 2011).

Stress is likewise a very broad term, and more than a dozen different definitions of the word are offered in Merriam-Webster's Medical Dictionary, including, "one of bodily or mental tension resulting from factors that tend to alter an existent equilibrium" (Webster, 2016, para. 1). This reference to the alteration of equilibrium suggests homeostasis. Hebb's (1955) model also describes a homeostatic curve, which should explain why many researchers have battled with the proposed distinction offered between the terms stress and arousal.

Sapolsky (2015) sums up the debate in the following,

> Seemingly within moments of Selye popularizing the word stress in the world of biomedicine, the definitional debate began. Is stress more about the unpleasantry in the outside world (that is, the stressor) or the resulting changes in the body (that is, the stress response)? Or is it mostly about the neurobiological and psychological space floating between the two? This eventually wearisome debate inevitably constituted the first session of virtually every stress conference for decades; it has finally lost steam, with a sense that the word encompasses all of the above—let a thousand flowers bloom, but just remember to define your particular flower in the Methods section. (p. 1348)

In this thesis the focus will be on the stress response for reasons that will be explored in this chapter.

In some areas of psychology shared research in neurology and other areas of medicine has blurred the lines even more, and it could be argued that the terms arousal and stress have become functionally equivalent. Examples of this can be found in research into examination stress, anxiety, and exhaustion by Strack and Esteves (2014) with 103 undergraduate students in Lisbon, Portugal. The students completed daily self-reports for ten days leading up to the examinations. These showed that individuals who interpreted stress and anxiety as facilitative reported lower levels of emotional exhaustion and performed better than similarly stressed individuals. This effect was present even when experiencing similar levels of stress and anxiety to other individuals. The term stress is being used, but in a manner more similar to the meaning attributed to arousal.

In this thesis the terms stress and arousal will be used interchangeably, with a preference for the term stress, but it is important to be cognisant of the theoretical issues

that underlie the debate. Stress is arguably a negative subset of arousal, with arousal a necessary but insufficient condition for resultant stress.

Scovel's (1978) frustration over the host of factors that influence arousal is the reason for the preference for the focus on resultant stress, and the reasons for this will be explored briefly below.

Even when considering seemingly straight-forward models like the Hebbian model raises questions about where and why peak arousal occurs. Different individuals have different tolerances, different coping strategies, and different perceptions of stress and arousal. This is what Sapolsky (2015) refers to as, "the neurobiological and psychological space floating between" (p. 1348) the stressor and the stress response.

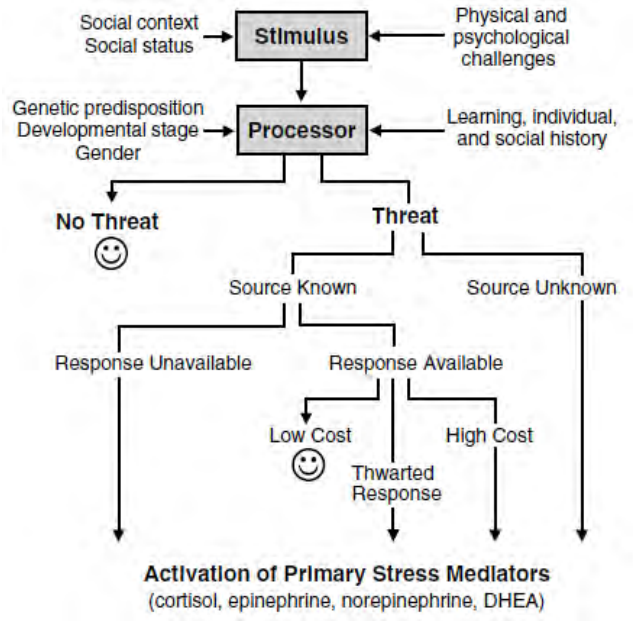
Friedenberg and Silverman (2012) argue that the definitional problems that plague this area of research arise from the interdisciplinary nature of the cognitive revolution in psychology that began in the 1950's. A wide range of fields began to collaborate, each with their own range of approaches and definitions. These included evolutionary approaches, emotional approaches, social approaches, philosophical approaches, and neuroscience approaches.

While the interdisciplinary work done in these fields has contributed immeasurably to our understanding of the mind, it has created fundamental definitional conflicts and debates that have led to positions like those expressed by Sapolsky (2015). Scovel's (1978) work is situated during the cognitive revolution, and bears many of the hallmarks of the ongoing and unresolved definitional debates of the period.

The definitional problem is exacerbated by the incredibly complex nature of the phenomena under examination. Figure 4 from Lupien, et al. (2006) presents McEwan and Stellar's (1993) model of stress, which captures just a small portion of the complexity of stress and arousal. McEwan and Stellar's (1993) model was generated based on a review of "Published original articles from human and animal studies and selected reviews. Literature was surveyed using MEDLINE." (McEwan & Stellar, 1993, p. 2093).

Figure 4

Perception and Mediation of Stress



Note. From "Beyond the stress concept: Allostatic load--a developmental biological and cognitive perspective", (Lupien, et al., 2006, p. 583). Copyright Lupien, et al, 2006. Reprinted with permission.

Even using a stress-based mechanistic approach, Lupien, et al.'s (2006) model shows that one cannot disentangle arousal or stress from a host of variables, such as social context, social status, physical status, psychological status, genetic factors, developmental stage, gender, learning, individual history, social history, threat perception and identification, response generation, the perceived cost of responses, and the perception of whether a response has been successful and control over the situation has been asserted, or whether the individual perceives the response as unsuccessful. Note that many of these variables differ depending on the individual stressor, not the situation as a whole, meaning that any attempt to weight these factors would necessitate breaking the situation down into each component stressor.

Furthermore, Lupien, et al. (2006) show that stress is not a unidirectional process, but rather incorporates a series of feedback loops where each stage may aggravate or ameliorate preceding phases, in a gestalt that modifies the values of each individual component as they interact with each other. The activation of stress mediating substances in the brain changes the individual's psychological state, looping one back to the beginning of the model.

For example, an individual might perceive clowns as threatening, and on seeing a clown the individual's reaction is one to feel stressed, so the brain initiates a chain of chemical reactions that result in adrenaline being released in anticipation of a fight, flight or freeze reaction. The release of adrenaline results in an increased heart rate, which feeds back into their psychological state of feeling threatened and heightens their sense of this being a dangerous situation.

To add another layer of complexity, research such as that by Strack and Esteves (2014) shows that in a longitudinal process such as learning, which occurs over a protracted period, post-hoc rationalisations and value judgements, which may shift over time.

Consider the example of being terrified on a rollercoaster but later viewing the experience as enjoyable. This mean that even if one could tease apart the various variables and score them, then the score might change later as the experience is retroactively re-evaluated.

Scovel (1978) expresses frustration at the host of variables that interact dynamically to produce the complex phenomenon that may be referred to as arousal or stress, and touches on an issue that has long frustrated researchers into the area of stress and arousal and continues to do so. Such a reductionist approach to stress would be like trying to break the Mona Lisa down into its component pieces in search of beauty.

Yet one needs to consider Scovel's (1978) work within the context of the cognitive revolution, where there was this notion from some quarters, such as neuroscience, that if enough variables were catalogued and analysed then anything could be understood (Friedenberg & Silverman, 2012).

In this thesis the approach taken will bypass this sort of approach, focusing on a measurement of inter-beat heartrate variance (the variation between beats of the heart) as a means of measuring resultant stress or arousal experienced by the individual at a particular moment. It is acknowledged that this approach simplifies a tremendously complex and dynamic process, but Scovel's elegant summary of his "bad news" (Scovel, 1978, 140), shows that pursuing the path of trying to count and weigh individual variables is a recipe for frustration and meaningless results that fail to capture any meaningful results.

Scovel's (1978) systematic review summarises many of the challenges that face research into LLA, but sadly seems to have fallen by the wayside, as will be seen in the next section.

2.3. Modern LLA Theory and Research

This section deals with the foundational concepts in modern LLA theory, and the tremendously influential work done by Horwitz, Horwitz, and Cope (1986). There is a great deal to critique in this section, and to prevent a long critique dealing with a diverse range of issues, this section will be divided up and critiqued in smaller chunks, addressing issues as they arise.

2.3.1. Publication of the Foreign Language Classroom Anxiety Scale

In 1986, Horwitz, et al. (1986) published the Foreign Language Classroom Anxiety Scale (FLCAS) in two papers, one by Horwitz and two collaborators (1986), and a second by Horwitz (1986) alone. Please note that in the following Horwitz (1986) refers to the single-authored paper, primarily concerned with the statistical evidence for the FLCAS. Horwitz, et al. (1986) refers to the co-authored paper that is primarily theoretical but does include some numerical data.

Both papers drew from the same research with students at the University of Texas in the USA. The paper by Horwitz, et al. (1986) mentions the involvement of two groups of 15 students who participated in a support group whose experiences were the basis for the FLCAS questions. Horwitz, et al. (1986) also report the results of the questionnaire done by 75 students from four introductory Spanish classes.

Horwitz's (1986) paper mentions "approximately 300 students" (p. 560). Horwitz (1986) specifies 35 students from two beginner Spanish classes and 32 students from two beginner French classes but provides no details as to the identity or origin of the remaining approximately 233 students.

The two papers appear to have been published around roughly the same time, with the paper by Horwitz, Horwitz, and Cope (1986) appearing in the summer edition of The Modern Language Journal. Horwitz's (1986) paper appears in the September edition of TESOL Quarterly.

The FLCAS claimed to, "test an individual's response to the specific stimulus of language learning" (Horwitz, 1986, p. 559), and "assess the degree of anxiety, as evidenced by negative performance expectancies and social comparisons, psychophysiological symptoms, and avoidance behaviours."

There are two items to note here. Firstly, that the FLCAS is, as the name implies, designed to be a specific measure of anxiety in the foreign language learning classroom. Secondly, that the FLCAS perceives of LLA only in terms of negative outcomes and makes

reference to “psychophysiological symptoms” (Horwitz, 1986, p. 559), employing language that suggests a clinical tone and that the anxiety may have crossed the boundary for caseness.

2.3.1.1. Critique of the conceptual basis for the FLCAS. Dealing with the issues in reverse order, one can see that Horwitz, et al. (1986) have adopted an approach where LLA is perceived as only negative. This immediately raises some questions about the FLCAS’s theoretical basis. If one considers the FLCAS in terms of Hebb’s (1955) model of arousal and performance there should be at least the possibility that participants are operating at below peak arousal, and therefore are not anxious and that their performance is not impaired.

Instead, Horwitz, et al. (1986) seem to be operating from the assumption that every individual has already exceeded peak arousal and is experiencing anxiety and declining levels of performance. This seems to be an unwarranted assumption, and seems to ignore the possibility that individuals may not be anxious at all, and might even be bored, or otherwise on the pre-peak side of the stress curve.

There are two references in Horwitz, et al. (1986) to making language learning less stressful as a remedy to language learning anxiety. Horwitz, et al. (1986) state that, “Teachers might create student support systems and closely monitor the classroom climate to identify specific sources of student anxiety.” (p. 131), and conclude the paper with the following:

> Foreign language anxiety can probably be alleviated, at least to an extent, by a supportive teacher who will acknowledge students' feelings of isolation and helplessness and offer concrete suggestons [*sic*] for attaining foreign language confidence. But if we are to improve foreign language teaching at all levels of education, we must recognize, cope with, and eventually overcome, debilitating foreign language anxiety as a factor shaping students' experiences in foreign language learning. (p. 132)

However, there is no engagement with the idea that learners might not be sufficiently stressed to be anxious. Thus, there is reason to suspect that Horwitz, et al. (1986) have conflated stress and anxiety into a single concept.

Horwitz, et al. (1986) offer a definition of anxiety, “Anxiety is the subjective feeling of tension, apprehension, nervousness, and worry associated with an arousal of the autonomic nervous system” (p. 125), which is more similar to stress or arousal than it is to anxiety. Of course, some degree of anxiety can and does occur at subclinical levels.

This assumption of anxiety is further exacerbated by the use of clinical language in making reference to symptoms, a term with medical overtones, and seems to lean towards pathologizing the anxiety that is alleged to exist. The use of psychometric instruments from clinical settings in the validation of the FLCAS, which will be explored in more detail in the following section, lends weight to the assertion that Horwitz, et al. (1986) tends towards a pathologizing approach.

Further, in Hortwitz, et al. (1986) it is explicitly stated that,

> We do know that individual reactions can vary widely. Some students may experience an anxious reaction of such intensity that they postpone required foreign language courses until the last possible moment or change their major to avoid foreign language study. Students who experience moderate anxiety may simply procrastinate in doing homework, avoid speaking in class, or crouch in the last row. Other students seldom, if ever, experience anxiety or tension in a foreign language class. (p. 131)

There seem to be contradictions inherent within Horwitz, et al.'s own writing and the manner in which the FLCAS is constructed. It is explicitly acknowledged that some students may not experience anxiety, but the FLCAS is then constructed in a manner that assumes that anxiety exists.

From the outset there are serious questions as to whether Horwitz, et al.'s model of LLA is even internally logically consistent. This assumption that all students are anxious, despite Horwitz, et al. (1986) acknowledging that this may not be the case, seems a contradiction. It becomes important later in the scoring of the FLCAS, which has no option that allows students to score as not anxious.

Moving on to the second point, namely the notion of situation-specific anxiety relating solely to the classroom environment, this is a deeply problematic position to take. As shown in Lupien, et al. (2006) stress, and possibly resultant anxiety, is a tremendously complex phenomenon that has inputs from a wide range of sources. Most of these stressors are outside of the scope of the classroom. The assumption that all of the inputs into the stress and anxiety dynamic originate solely within the foreign language learning classroom is not supported by the literature on stress and anxiety. Students do not walk into the classroom environment with zero stress or anxiety.

Some of the FLCAS questions refer to situations outside of the classroom, such as, "I often feel like not going to my language class." (Horwitz, et al., 1986, p. 129), and "When I'm on my way to language class, I feel very sure and relaxed." (Horwitz, et al., 1986, p. 130).

There seems to be no engagement with the idea that an individual might not want to go to language class for reasons extraneous to the language class. For example, feeling ill or simply having a bad day. Moreover, there seem to be internal inconsistencies in the FLCAS in that it claims to be solely focused on the language learning environment, but then also asks questions about situations outside of the language learning environment.

Lupien, et al. (2006) present a list of potential moderators of the stress reaction. While it is undeniable certain situations may be stress-provoking, the notion that all resultant stress can be attributed to the situation seems to fly in the face of common sense. Consider two common stress variables, namely sleep and exercise.

Sleep performs a vital role in the regulation of emotional responses (Gruber & Cassoff, 2014), in memory formation (Stickgold & Walker, 2013), and in both physical and mental health (Tamakoshi & Ohno, 2004; Milojevich & Lukowski, 2016). The positive correlation between sufficient sleep in both general academic performance and in language learning specifically has also been established (Wong, et al., 2013; MacDonald, 2015).

Exercise and stress management is also a well-researched area that shows an important role in stress regulation, but also benefits in physical and mental health, and many other areas (Contrada & Baum, 2011).

Salmon (2001) offers a systematic review of the literature on the effects of physical exercise on anxiety and stress sensitivity. Salmon (2001) claims that "… immediately before and after regular exercisers undertake strenuous exercise at a level with which they are familiar … overwhelming evidence confirms mood improvement" (p. 35). Similarly, for those who are normally sedentary mild to moderate exercise produces an improvement in mood, or a less negative mood.

However, there are two important caveats in Salmon (2001). Exercise that is unusually intense may worsen mood, as can competitive activities. This suggests that the effect of exercise on mood is not one of an absolute number of minutes of exercise, but rather an issue of whether it falls within the individual's usual level of activity, suggesting that regularity is a key issue. There is also the nature of the exercise, with non-competitive activities tending to have a more positive impact on mood.

Salmon (2001) also notes that mood can be subjectively reassessed, writing that, "… the measurement of mood is complex… Effects change over time, too, so that initial mood-worsening during exercise can change into mood-improvement 30 or more minutes later" (p. 36). This is a phenomenon that has already been commented on earlier, but bears repeating as a significant issue in many studies of stress and mood.

In the course of the systematic review Salmon (2001) also raises another important issue, namely that when researching into mood one cannot treat any phenomenon as discrete but must rather consider other ancillary factors such as social and environmental factors.

Relevant to this research is a study cited by Salmon (Steptoe, et al., 1997), of 16,483 undergraduate university students from 21 European countries surveyed by questionnaire that showed a negative correlation between depression and exercise. This finding is corroborated by Stephen's (1988) population surveys in the USA and Canada, with a combined sample size of 55,000 people, that showed that level of physical activity was positively correlated with better mental health, and negatively correlated with symptoms of anxiety and depression.

Salmon's (2001) review of the research on exercise and anxiety suggests that while it may be effective for mild non-clinical cases of anxiety it may exacerbate the effects of anxiety in those with a clinical level of anxiety.

Finally, regarding stress resistance and exercise Salmon's (2001) review found some evidence for a protective effect, however Salmon (2001) raises questions about the validity of much of the research conducted under laboratory conditions as this methodology removes many of the psychological and environmental factors that would make the findings meaningful. Salmon (2001) does cite some research using life stress measures which found a negative correlation between life stress and physical activity; however it was unclear whether this was a cause or effect.

While no researcher can include every possible variable in a study, and there is always the difficult process of determining which variables to include and which to exclude, it should be clear from what has been explored above that this is not a critique of Horwitz, et al.'s (1986) research design. Instead, this is a critique of their conceptual and theoretical framework that attempts to cordon off foreign language learning anxiety from well-known moderating variables outside of the classroom.

There seem to be internal contradictions in Horwitz, et al.'s (1986) stated theoretical framework and how the theory is expressed in the test questions in the FLCAS, as well as contradictions between Horwitz, et al.'s (1986) theory and other theories of anxiety, which challenge the concept of context-specific anxiety.

According to Tran (2011, p. 70), a "large number of studies" have used Horwitz, et al.'s (1986) model, and an admittedly unscientific foray onto Google Scholar revealed over 5,040 papers mentioned Horwitz, et al.'s (1986) FLCAS. Horwitz, et al.'s 1986 paper has

been cited 7,235 times, giving a sense of how influential this idea and measure has become, both in language learning and more broadly in academia.

Cassady (2011) makes reference to “math and science anxiety” (p. 6) in addition to reading, computer usage and foreign language learning anxiety. Cassady (2011) also echoes many of Horwitz, et al.’s concepts, such as situation-specific anxiety, proposing that there is no facilitative anxiety, and even makes the statement, “Obviously, this stress is detrimental to student achievement” (p. 177), conflating stress and anxiety as if the terms were interchangeable and indicating how Horwitz, et al.’s work has been interpreted.

The spread of Horwitz, et al.’s (1986) ideas within the field of foreign language learning, and then more broadly into academia in general underlines the importance of a thorough critique of the ideas presented by Horwitz, et al. in their paper.

There is a well-known bias against the publication of null results (Hubbard & Armstrong, 1997). When insufficiently well curated papers, particularly those that show positive results, are published on a subject they are treated as true and accumulate further research building on these faulty ideas. Papers suggesting that the results are not replicable, such as those papers investigating the idea but with a null result, are less likely to be published. Once the faulty idea accumulates sufficient momentum it is difficult to stop, and may inform debates in other areas, and even public policy. There is an academic duty to investigate, and if necessary correct, these misconceptions before they can cause further harm.

Misconceptions about stress and anxiety can cause tremendous harm, as can the misappropriation and misuse of psychometric instruments by those who lack the proper training or awareness of their limitations (Furr, 2017). The issue of the pathologizing language used in Horwitz, et al. (1986) with references to anxiety and “psychophysiological symptoms” (Horwitz, 1986, p. 559) that the FLCAS can “… address conceptually and clinically important aspects of anxiety” (Horwitz, 1986, p. 560) has been mentioned earlier.

This is worrying in that it may cause learners to over-estimate their stress levels and create perceptions of anxiety that become self-reinforcing. This may occur through the feedback processes that can be seen in stress models such as Lupien, et al. (2006), causing direct harm to learners.

Additionally, the use of clinical language and psychometric tests lends a gravitas to the assertions being made and suggests level of scientific rigour that later discussion will show may be unjustified. The precision of the results delivered by psychometric tests is heavily dependent on their use by an appropriately trained individual. Without an

understanding of proper administration, scoring, and interpretation the reliability and validity of the results is questionable, if not outright invalid (Furr, 2017).

For example, there is frequently comorbidity between anxiety and depression. To address this Spielberger (1983) designed the State-Trait Anxiety Inventory (STAI) to distinguish between anxiety and depression in the diagnosis of anxiety. The test is designed for a specific context, and to be administered, scored, and interpreted in a very specific manner. The validity and reliability of the test are contingent on these factors. A researcher who administered just half the test, such as just the trait inventory, in a non-clinical setting, would not necessarily receive valid or reliable results about participants' anxiety. In relation to scoring the FLCAS, Sparks and Ganshow (2007) note that, "The authors of the FLCAS do not include a scoring procedure with the instrument…", (p. 268).

There is also the issue of how psychometric testing affects teacher expectations of learner performance and the effect this has on learners. In 1965 Rosenthal and Jacobson (1968) conducted an experiment at a public elementary school in San Francisco in the USA with 320 elementary school students. The experiment consisted of a control group of 255 students and a randomly chosen experimental group of 65 students. Teachers were informed that the randomly chosen experimental group students had performed highly on a fictitious test of giftedness. The results showed that teachers favoured these students, assessing them more highly, and that this positive attention led to concrete gains in IQ tests in a self-fulfilling prophecy.

The danger of constructs such as LLA is that it may have the inverse effect, with teachers perceiving those who scored highly on LLA as being less capable of learning.

While experiments such as those by Rosenthal and Jacobson (1968) on the interaction between psychometric testing and teacher expectations and the performance of students have been disputed, there is a long history of the misuse of psychometric testing. The potential for harm to be done to learners means that claims relating to psychometric testing deserve special scrutiny (Mitchell & Daniels, 2003; Furr, 2017).

As such the next section will explore the construction of the FLCAS in some detail as it claims to have been validated against psychometric instruments.

2.3.2. Construction and validation of the FLCAS

Horwitz, et al. (1986) generated the list of 33 questions in the FLCAS from five sources. The first source was, "Two groups of anxious foreign language students" (Horwitz, 1986, p. 560), the second was interviews with counsellors at the Learning Skills Centre at the University of Texas in Austin about, "their experiences with anxious language learners" (Horwitz, 1986, p.

560). It is unclear from Horwitz whether these counsellors were qualified in any way. Third, “The author’s experience with anxious students” (Horwitz, 1986, p. 560). Fourth, “Measures of test anxiety…, speech anxiety… , and communication apprehension… were reviewed to identify relevant items.” (Horwitz, 1986, p. 560), and finally, “five items from the French Class Anxiety Scale… were made generic and added to the item pool” (Horwitz, 1986, p. 560).

Once the FLCAS question list was generated, Horwitz (1986) reports that the FLCAS was administered to about 300 students in an unspecified number of separate studies, then only reports results for one group of 108 students, and it is unclear if the results that follow are reflective of all the separate studies, this one group of 108 students, or from different studies, although where numbers of participants are listed they differ, suggesting different studies.

In Horwitz’s (1986) paper entitled, “Preliminary Evidence for the Reliability and Validity of a Foreign Language Anxiety Scales” the results of 13 statistical tests performed to assess the reliability and validity of the FLCAS are reported. In the following section (unless stated otherwise) the results reported relate to Horwitz’s (1986) paper. The results are summarised below:

1. Internal Consistency (via Cronbach’s alpha coefficient) = 0.93
2. Test-rest Reliability (8 weeks) r= 0.83, p= 0.001, n= 78
3. Self-rated foreign language course anxiety r= 0.77, p= 0.001, n= 108
4. Expected Final grade in a foreign language course r= 0.52, p= 0.001, n= 108
5. Final grade in two foreign language courses (Spanish) r= -0.49, p= 0.001, n= 108
6. Final grade in two foreign language courses (French) r= -0.54, p= 0.001, n= 32
7. Final Grade (controlling for Test Anxiety) r= -0.53, p= 0.002, n= 29
8. Trait scale of the State-Trait Anxiety Inventory (Spielberger, 1983) r= 0.29, p= 0.002, n= 108
9. Personal Report of Communication Apprehension (McCroskey, 1970) r= 0.28, p= 0.063, n=44
10. Fear of Negative Evaluation Scale (Watson & Friend, 1969) r= 0.36, p= 0.007, n= 56
11. Test Anxiety Scale (Sarason, 1978) r= 0.53, p= 0.001, n= 60
12. Test Anxiety Scale (Sarason, 1978) and Final Grade (French) r= -0.16, p= 0.391, n= 32
13. Test Anxiety Scale (Sarason, 1978) and Final Grade (Spanish) <<No data>>

The first test, that of internal consistency, is a measure of whether all the test items are measuring the same construct. Cronbach’s coefficient alpha pairs items and then compares the scores. Scores on this test can range from negative scores to 1, with a higher positive score indicating that items scored similarly to each other, and thus are probably

measuring the same thing. The FLCAS score of 0.93 is above 0.9, indicating excellent internal consistency, but is below 0.95, which indicates that items are all scoring functionally identically and so are redundant (George & Mallery, 2019).

Horwitz (1986) does not report what statistical test was used on the remaining tests, but the reporting format is consistent with Pearson's product-moment correlation coefficients, which measure a linear correlation between two sets of data. The r value indicates the strength and direction of the correlation, ranging from -1 to 1, with both extremes representing the data points lining up perfectly on a straight line when graphed (Cohen, 2013). A positive r value indicates a positive correlation, and that as one variable increases so the other variable increases, and the same for decreasing with the variables moving in the same direction. For example, if the r value was positive then as students scored higher on the FLCAS, indicating higher anxiety, so the second variable also increases. A negative r value indicates a negative correlation, and that as one variable increases the other decreases. For example, if the r value was positive then as students scored higher on the FLCAS, indicating higher anxiety, so the second variable decreases.

The r value can also be squared to indicate the fit between the two variables: how many of the data points fall close to the perfect straight-line configuration, often expressed as a percentage. Horwitz (1986) uses this technique, stating that "These results indicate that anxiety specifically related to foreign language class accounts for approximately 25% of the variance in final grades." (p. 561).

Except for items 5, 6, 7, and 12 all of Horwitz's (1986) r values indicate positive correlations ranging from 0.28 (a 7.8% fit) to 0.83 (a 68.89% fit). Items 5, 6, 7 and 12 are negative correlations ranging from -0.16 (a 2.56% fit) to -0.54 (a 29.16% fit).

The p value indicates the probability that the correlation is random chance, and ranges from a theoretical 0, indicating absolute certainty that the correlation is non-random, to 1, indicating a 100% chance that the correlation is random. Normally in the behavioural sciences 0.05 or lower is regarded as the threshold for regarding a correlation as non-random (Andrade, 2019), indicating about a 95% chance that the correlation is non-random. In the physical sciences the threshold is 0.01, indicating a 99% chance that the correlation is non-random (Cohen, 2013). As an aside, the explanation just presented is not entirely correct, but is a sufficiently correct way to explain a tricky concept about probability, and the author is aware of the nuances in probability theory, but they lie beyond the scope of this thesis (Andrade, 2019).

Except for items 9 and 12 all of Horwitz's (1986) p values are below 0.05, satisfying the threshold for statistical significance and a non-random result in the behavioural sciences.

Items 9 and 12 are above the threshold for statistical significance, although item 9, which has a p value of 0.063 is sufficiently close that it should probably be considered a marginally significant correlation, with this value indicating a 93.7% chance that the result is non-random.

The n value indicates the number of individuals in a particular statistical sample. If a group of 44 individuals took one of these tests, then the n value would be n=44. It is possible for the n value to be different within the same group. For example, if a group of 44 individuals did a series of two tests, but on the second test an individual omitted some questions or spoiled some of the answers, then that individual's test might be discarded. This might result in the same group being reported as n=44 for the first test, and n=43 for the second test. Statistical outliers might also be removed from a group to prevent an unusual result from skewing the results for an entire group (Cohen, 2013). Therefore, it is possible that the n value might vary for the same group, however in cases where groups are unnamed or unspecified the n value can be used as a basis for determining if one is looking at the same group or different groups.

For example, result number five specifies that it belongs to the group of 108 Spanish students. For results three, four and eight Horwitz (1986) does not specify which groups of students took these tests, however the n value of 108 is the same as that for the Spanish students, leading to the reasonable conclusion that the results reported are most likely from the group of Spanish students.

The n values ranged from 29 to 108, with a group of French students (n=32) and a group of Spanish students (n=108) identified in four of the thirteen tests conducted, but no group identified in the remaining nine tests.

2.3.2.1. Critique of the reliability and validity of the FLCAS. The first thing to note is the n values. These indicate that Horwitz has reported some results for some groups, but not for other groups, and in some cases, such as the internal consistency, has completely omitted the n value. This may be because, as the research proceeded it was realised that additional tests or data were required, and it was impossible to gather participants from previous groups to participate in additional testing. This is not an uncommon problem in research. Horwitz (1986) refers to the FLCAS having been administered to, "approximately 300 students" (p. 560).

Looking at the n values there seem to have been the following groups:
A. Two Spanish classes n=108 (results 3, 4, 5, and 8)
B. Two French classes n=32 (results 6 and 12)
C. Unnamed Group n=78 (result 2)

D. Unnamed Group n=29 (result 7)
E. Unnamed Group n=44 (result 9)
F. Unnamed Group n=56 (result 10)
G. Unnamed Group n=60 (result 11)

Totalling the n values gives a result of 407, although there may be some overlap between the groups, or some of the lower n values may represent incomplete responses, absent students, or responses that needed to be discarded from larger groups. For example, group B (the two French classes) contained 32 participants, but group D may be the same group, but with three incomplete or spoiled answers, or absent students on test number 7.

Regardless though, the manner in which Horwitz (1986) fails to clearly report which groups completed which items raises questions about whether Horwitz is “cherry-picking” and only reporting favourable results. Horwitz (1986) reports that approximately 300 students completed the FLCAS, but only reports the results for the correlation between the FLCAS and Final Result for two groups (A and B), totalling 140 participants. It is possible that Horwitz (1986) felt that the hypothesis had been sufficiently well tested and moved on to testing other things but given that these tests required completing the FLCAS and that results would have been easily accessible it does raise questions why Horwitz would not report these results.

There are also what appear to be errors in Horwitz’s (1986) statistics that may be typographical errors or may indicate more serious concerns about the accuracy of the results presented. The following section details several irregularities in the results presented by Horwitz (1986) that are not on their own sufficient basis to question the reliability and validity of the statistics presented but contribute to a pattern of seemingly poor data handling.

Result 4 indicates a positive correlation of r=0.52 between Expected Final Grade and the FLCAS. A positive correlation indicates that both variables increase together, so as expected final grade increased, so did the FLCAS score, indicating increasing anxiety. This seems to contradict Horwitz’s (1986) core hypothesis, which is that final grade would decrease as anxiety increased (a negative correlation). Results 5, 6, and 7, which also deal with the FLCAS and Final Grade, all show a negative correlation, which agrees with Horwitz’s (1986) hypothesis that higher FLCAS scores, indicating increasing anxiety, would result in decreasing Final Grades. The easiest explanation for this is to simply assume that a minus sign was omitted as a typographical error, however it does raise questions about the accuracy of the results, and if the results are accurate then the validity of the theory is in question.

Further irregularities are present in the following: result 7 (the p value is reported as 0.002, but is 0.003105), result 9 (the p value is reported as 0.063, but is 0.065643), result 10 (the p value is reported as 0.007, but is 0.006425), and result 12 (the p value is reported as 0.391, but is 0.381715). The correlation significances were calculated using standard parameters for the Pearson's Product Moment Correlation (Stangroom, 2021).

None of these irregularities materially alter the outcomes, result 9's error might easily be considered a simple rounding error. In all fairness this research was done when many statistics were still done by hand, but it contributes evidence to a pattern of irregular handling of statistics and data in this seminal paper on LLA.

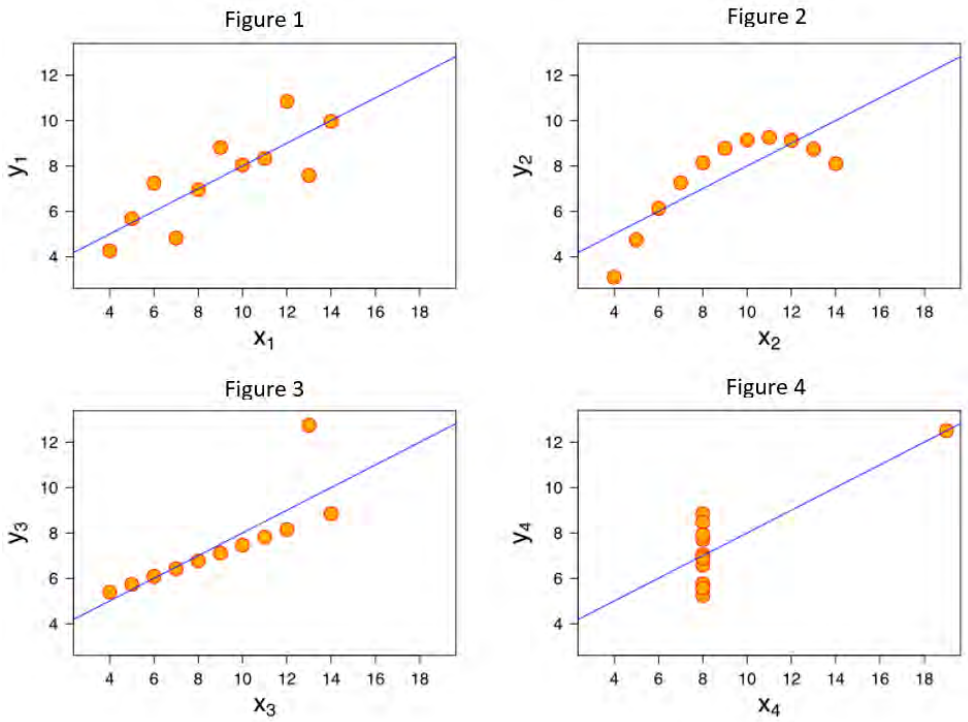
This also raises the issue that of the eleven Pearson product-moment correlation coefficients presented, seven of them have p values of less than 0.01. There is a reason that 0.05 is considered to be sufficient for statistical analysis in the humanities, and this is because humans are not as consistent or predictable as machines or the natural laws of the universe. The threshold of 0.05, or 95% reflects that it is expected that approximately 1 in 20 humans will differ from the norm, producing unusual or unexpected results. Yet in seven of Horwitz's (1986) results the reported p value was $p<0.01$, the threshold normally associated with investigations of physical laws of the universe, not university students. While a novice in statistics might think that presenting a list of results with extremely low p values indicates a great result, in the mind of anyone well-versed in research in the humanities and psychometrics this raises questions as to whether the individual has not perhaps made an elementary error in the handling of the statistics, such as accidentally sorting results from highest to lowest before running the analysis thereby generating this sort of very low p value by accident, particularly given the small sample sizes.

Horwitz's (1986) statement that, "These results indicate that anxiety specifically related to foreign language class accounts for approximately 25% of the variance in final grades" (p. 561) seems to indicate a lack of awareness that correlation does not necessarily indicate causation, and while there may be a correlation that shows a 25% fit for the variance between the two variables that does not mean that anxiety is the cause, merely that there is a correlation between these two things.

Most introductory statistics courses illustrate that a perfectly curved correlation, such as that hypothesised in the Hebbian model of stress (1955), or the Yerkes-Dodson model (1908) would result in a correlation strength of 0 on a linear correlation test. However, if there is an incomplete non-linear correlation, there can be a strong correlation. The most famous example of this is in Anscombe's Quartet (Anscombe, 1973). Figure XX shows Anscombe's Quartet, all of which have strong r values of about $r=0.81$.

Figure 5

Anscombe's Quartet



Note. The original version of this figure was rather small and hard to read, and so has been recreated here, based off of the Wikimedia Commons license version (**https://en.wikipedia.org/wiki/File:Anscombe%27s_quartet_3.svg**) . The original figure can be found in "Graphs in Statistical Analysis", (Anscombe, 1973, p. 19-20).

Both figures two and four in Anscombe's Quartet are relevant to this discussion. Horwitz (1986) provides no graphs of the FLCAS validation data, nor is detailed data available from which graphs could be derived.

This raises the possibility that the data described an incomplete curvilinear relationship (such as in figure two of Anscombe's Quartet). Hebb's (1955) model predicts this sort of curved line relationship, so this should have been considered as a possibility. Framing this within Hebb's (1955) model, if Horwitz's (1986) samples had consisted mostly of participants below peak arousal (to the left side of the peak of the curve) and a minority who were above peak arousal and mildly to moderately clinically anxious, then it may have generated the positive correlation that one sees in Horwitz's statistics according to Anscombe's Quartet.

A second possibility is that statistical outliers were not removed (such as in figure four of Anscombe's Quartet). This produces statistical results that look like sound statistical evidence but prove nothing. Consider the possibility that Horwitz's sample consisted of many individuals experiencing no significant anxiety, and a small minority of individuals who were severely anxious (statistical outliers). If the majority group's results were unaffected by LLA, but the severely anxious (outliers) results were seriously affected then this is the type of situation depicted in figure four of Anscombe's Quartet. Without the single statistical outlier there is no correlation. If the outlier is removed, then the remaining data points best fit a straight vertical line where the variables do not affect each other at all (r=0).

It is therefore possible that the results reported in Horwitz (1986) might be mathematically correct, but not accurately depict the relationship of LLA to performance. Two possibilities have been explored here, the first is that the relationship is curvilinear as predicted by Hebb's (1955) model. The second is that a border condition exists in the LLA phenomenon, with LLA only significantly affecting the results of severely anxious individuals, but the results of those without significant anxiety are unaffected. To understand if either of these possibilities were explored by Horwitz (1986) and Horwitz, et al. (1986) requires a discussion of validity of the FLCAS and the conclusions being drawn.

On examining the five sources from which Horwitz, et al. (1986) generated the FLCAS a word keeps recurring, namely anxious. The first source was anxious language learning students, the second source was interviews with counsellors about their experiences with anxious language learning students, the third source was Horwitz, et al.'s experiences with anxious language learning students, the fourth was Sarason's (1978) measure of text anxiety, and the final source was five of the 33 items in the FLCAS adapted from Gardner, et al.'s (1979) French Class Anxiety Scale (Horwitz, 1986).

With Sarason's (1979) test anxiety measure, and Gardner, et al.'s (1979) French Class Anxiety Scale the focus on anxiety is explicable. However, it seems that Horwitz, et al.'s (1986) remaining sources have focused exclusively on anxious students, which raises questions about whether Horwitz, et al (1986) adequately allowed for the possibility that some students might not be anxious at all in their theoretical framework.

The FLCAS scoring system has no anxiety-neutral option. The FLCAS is scored on a 5-point Likert Scale, with responses ranging from strongly agree to strongly disagree, and total scores ranging from 33 to 165, with Horwitz (1986) proposing a negative linear correlation between FLCAS scores and final results, with this accounting, "for approximately 25% of the variance in final grades." (p. 561). Scoring ranges from one to five, with no provision for negative scoring, even when participants clearly indicated that they strongly

disagreed with statements such as, “I am usually at ease during tests in my language class.” (Horwitz, et al., 1986, p. 129).

To give an example of the problems with the construct validity of the FLCAS in the data presented in Horwitz, et al. (1986) the question, “I am usually at ease during tests in my language class.” (p. 129) was answered with five participants strongly agreeing (SA), 35 participants agreeing (A), 19 neutral (N) responses, 20 disagreeing (D), and 21 strongly disagreeing (SD). Of the 100 responses reported in Horwitz, et al. (1986) the majority (69 out of 100) of participants indicated varying degrees of ease (40 out of 100) or at minimum an anxiety-neutral response (19 out of 100). This is not an isolated example, with another question asking, “It wouldn’t bother me at all to take more foreign language classes”, to which the responses were 15 SA, 47 A, 12 N, 16 D, 11 SD. When 74 out of 101 respondents are not bothered at all by the prospect of taking additional foreign language classes it seems hard to justify the notion that the individuals are anxious about language learning.

As an aside, it is noteworthy that here again there is a mismatch between the data reported in Horwitz, et al. (1986), where there are 99 to 101 respondents, and Horwitz’s (1986) paper on the validation of the FLCAs, where none of the listed group sizes match 100, the closest being 108.

The presumption of omnipresent anxiety and steadily rising debilitative anxiety inherent to the construction of the FLCAS seems unfounded when what initial data is available shows that this is not a fair or reasonable assumption for the majority of respondents, nor is this a defensible approach in terms of existing models of stress and anxiety.

Horwitz, et al. (1986) justify the label of anxiety by demonstrating concurrent validity (a form of criterion validity) with “the Trait scale of the State-Trait Anxiety Inventory (Spielberger, 1983)”, (Horwitz, 1986, p. 560), reporting a positive correlation of r= 0.29, p= 0.002, n= 108. There are numerous problems with this approach, which will be detailed here.

The State-Trait Anxiety Inventory (STAI) has, as the name implies, two scales, one for trait anxiety, which Spielberger (1983) defines as follows:

> The STAI clearly differentiates between the temporary condition of "state anxiety" and the more general and long-standing quality of "trait anxiety." The essential qualities evaluated by the STAI-S Anxiety scale are feelings of apprehension, tension, nervousness, and worry. Scores on the STAI-S Anxiety scale increase in response to physical danger and psychological stress, and decrease as a result of relaxation training. On the STAI-T Anxiety scale, consistent with the trait anxiety

> construct, psychoneurotic and depressed patients generally have high scores. (para. 1)

Horwitz (1986) only validated the FLCAS against the STAI's Trait Anxiety Scale (STAI-T). As the word trait suggests, this scale measures "a relatively stable personality trait" (Barnes, et al., 2002, p. 604), and "People who are high in trait anxiety tend to perceive more situations as threatening or dangerous than people who have lower trait anxiety scores" (Barnes, et al., 2002, p. 604). Bados, et al. (2010) confirm this in their study of the validity of the STAI-T, stating that, "The correlation of the hypothetical anxiety subscale with measures of depression was equivalent to or higher than its correlation with measures of anxiety. These results suggest that the questionnaire does not strictly evaluate anxiety but, rather, negative affect" (p. 560).

It is therefore hardly surprising that Horwitz (1986) found a strong correlation between individuals responding to the more negatively worded items on the FLCAS and those with a high score on the STAI-T. This is what the STAI-T tests for. It does not, as Bados, et al. (2010) point out, necessarily mean that they are anxious, but rather as both Spielberger (1983), and Bados, et al. (2010) point out, the STAI-T is more closely associated with depression and negative emotion.

Furthermore, the STAI-T is not a reflection of context-specific anxiety, but rather "a relatively stable personality trait" (Barnes, et al., 2002, p. 604). If the FLCAS is measuring a stable personality trait then this would also partially explain the seemingly good test-retest reliability scores of $r=0.83$, $p=0.001$, $n=78$. It would be expected that someone would score similarly on retesting as their personality has not substantially changed between one testing and the next.

A longitudinal study by Sparks and Ganschow (2007) seems to confirm this assessment, reporting that, "… anxiety as measured by the FLCAS was negatively correlated with students' native language skills as early as nine years prior to encountering a foreign language course" (p. 279). Something that is stable over a period of nine years, reflecting multiple instructors, and even multiple learning environments, tends to indicate the FLCAS is not measuring a response to a specific stressor, but rather a fundamental personality trait.

Therefore, there are good reasons to suspect that what the FLCAS may be measuring, at least in part, is a personality trait rather than situation-specific anxiety. This has two important implications.

Firstly, if the FLCAS is measuring a personality trait that predisposes individuals towards negative affect then this partially explains the correlations obtained with other tests used in the concurrent validation of the FLCAS. The FLCAS was validated against several other tests, namely the Personal Report of Communication Apprehension (McCroskey, 1970), the Fear of Negative Evaluation Scale (Watson & Friend, 1969), and the Test Anxiety Scale (Sarason, 1978). What all these tests have in common is a tendency towards negative affect, or negative emotion. It is hardly surprising that individuals who scored higher on the STAI-T also scored higher on other tests the involve negative emotional judgements.

Secondly, Horwitz et al. (1986) conclude that, “Foreign language anxiety can probably be alleviated, at least to an extent, by a supportive teacher” (p. 132). If the FLCAS is measuring a stable personality trait, then a supportive teacher is unlikely to change the individual’s personality. This begs fundamental questions about the FLCAS’s content validity.

Further, the FLCAS is supposed to represent foreign language learning anxiety, not just foreign language performance as reflected by final grades. The FLCAS seems based on a uniform learning experience across their whole foreign language learning experience, regardless of instructor or learning environment. The FLCAS’s validation is based on scores obtained from a single language class, and seems to assume that the participants’ experiences in this one class are reflective of their general performance across all foreign language classes and with all instructors. This raises two questions, the first is whether these classes were indeed foreign language classes, and the second is the question of teacher and learning context. Both of these issues are fundamental when the construct in question labels itself as a Foreign Language Classroom Anxiety Scale.

The first question of whether this is a foreign language is questionable when Horwitz, et al.’s (1986) largest group of participants (two Spanish classes n=108) were studying Spanish, in Texas in the United States, where according to the 1990 census out of a total population of 15.6 million people, 3.4 million spoke Spanish at home, approximately 21.79% of the population (U.S. Census Bureau, 1994). When more than one in five people in a particular area speak a language, it seems questionable to consider it a foreign language, and seems to be quite a problematic position to take. Further, assuming that the participants had no knowledge of Spanish whatsoever seems to be unfounded, raising questions about whether all of the students’ exposure to the language can reasonably be considered to be attributable to the classroom environment.

If Horwitz, et al. (1986) were really measuring language learning, and if teacher variables were as important as their conclusions claim, then there should have been some

attempt to distinguish between initial proficiency level and final proficiency level and measuring how much learning actually took place. Instead, what Horwitz (1986) presents is data that does not show learning, but instead could be read as showing that students with higher levels of proficiency are less prone to negative affect with regards to language classes, while learners with lower levels of proficiency are more prone to negative affect with regards to language learning classes. This result would not surprise anyone, it makes sense that people feel better about something when they are good at it, and it certainly lines up with Csikszentmihalyi's (1975) flow model's predictions.

In summation, this critique has raised some serious questions about the validity of the evidence presented by Horwitz (1986) and Horwitz, et al. (1986) for the FLCAS. From the outset Horwitz, et al. have focused on anxious students in their generation of the questions, raising questions about whether there was adequate provision made for the possibility that students might not be anxious. The scoring system seems to bear out these concerns, with no provision for a determination of no anxiety being present. Irregularities in the sample sizes reported for different tests raise questions of inaccurate, irregular, or incomplete reporting of results. The lack of detailed information raises questions about whether the straight-line negative correlation between the FLCAS and final results really reflects a straight-line negative correlation, or whether because of errors in the handling of statistical data, as shown in Anscombe's Quartet (1973), might not be a curvilinear relationship as predicted in other models of stress such as Yerkes-Dodson (1908) and Hebb (1955).

Another critical issue with the data presented by Horwitz (1986) is the evident confusion over the use of the STAI and the decision to use only the Trait anxiety scale. Reading Spielberger's (1983) definitions of the scales should have made it clear that, "depressed patients generally have high scores" (para. 1) on this scale, and since the FLCAS is supposed to measure anxiety, not depression, this seems to suggest that there were certain fundamental misunderstandings about how the test could be used and what each scale measured. Psychometric tests are sensitive instruments, and one cannot simply assume that the name of the test or the scale is sufficient information to use the test in a particular fashion.

There are also serious concerns over the whether the FLCAS is even measuring learning at all, or whether they have simply presented data on emotional affect and language proficiency level, showing that more proficient students tend to have a more positive attitudes than low proficiency students.

Overall Horwitz (1986) and Horwitz, et al. (1986) have failed to present evidence of sufficient quality to, even assuming the lesser standard of on balance of probabilities, demonstrate that the FLCAS measures anxiety, and there are reasonable grounds to question the validity of the test.

In closing, it should be noted that there is no presumption of malice or malfeasance in Horwitz (1986) and Horwitz, et al.'s (1986) work, and there seems to have been a genuine attempt to present statistical evidence of the FLCAS's validity with reference to other tests know to be valid. The use of statistics was not common practice at that time in the field of linguistics, nor are linguists generally trained in psychometrics. In many ways the work done by Horwitz (1986) and Horwitz, et al. (1986) was revolutionary by the standards of the time and the discipline in which it arose, but as with so many pioneering moves made fundamental errors that appear with hindsight to invalidate the results.

2.4. Beyond Horwitz (1986) and Horwitz, et al. (1986)

As the last section shows, the evidence presented for the FLCAS raises reasonable concerns that it is not a measure of anxiety, that there are serious statistical anomalies that invalidate the conclusions reached, and that the results obtained can more easily be explained as identifying individuals with a tendency towards negative affect, specifically possibly depression rather than anxiety.

Despite the above, the concept of foreign language anxiety seems to have become accepted as a fact in the field of language learning, as evidenced by this quotation from Trang (2012) "There is a considerable body of research indicating that foreign language learning anxiety is not merely an abstract concept studied by theorists or by researchers under laboratory on induced-anxiety conditions, but a reality for many students" (p. 69).

The concept of foreign language learning anxiety, and Horwitz's FLCAS has gained considerable traction and attracted a large body of research. This is an inherent danger of iterative knowledge systems where insufficiently well curated research enters a field and then tends to be assumed to be true, and others build on this foundation, seemingly unaware that the foundation is suspect.

To illustrate this point, eight papers are cited in Trang (2012) as evidence of research supporting the reality of foreign language learning anxiety. Casado and Dereshiwsky (2001) use the FLCAS as evidence to support the contention of the existence of foreign language learning anxiety, as did Kostić-Bobanović (2009), Liu (2006) using a version of the FLCAS adapted into Chinese, Liu, and Jackson (2008), Tallon (2009) and Von Wörde (2003).

Of the eight studies cited in Trang (2012) in support of foreign language anxiety, six use the FLCAS as evidence to support the notion. While MacIntyre and Gardner (1995) present a very interesting and nuanced theoretical perspective, there is no actual evidence presented.

If the FLCAS is not a valid and reliable measure of LLA, as the previous section's analysis seems to suggest, then all of this subsequent research using the FLCAS is called into question as well. As noted earlier the FLCAS has been used extensively in LLA research, and is mentioned in over 5,000 papers. The size of this problem is simply staggering, and its impact is indicative of serious issues in the field of LLA research.

Not all research into LLA used the FLCAS. Coryell and Clark (2009) conducted qualitative research based on semi-structured interviews, but their coding system for the interviews is highly problematic, assuming the existence of foreign language learning anxiety as a reality and labelling a wide variety of responses as indicative of anxiety as evidenced in the following quotation: "The participants experienced various levels of anxiety in their learning, which is evidenced by the words they used to describe their anxious experiences; they ranged from feeling 'intense anxiety,' 'frustration,' 'obsession,' and 'stress,' to 'nervous' and 'silly'" (Coryell & Clark, 2009, p. 490).

While labelling "intense anxiety" (Coryell & Clark, 2009, p. 490) is defensible, the rest of the words used in interviews seem unrelated to anxiety, and to label feeling "silly" (Coryell & Clark, 2009, p. 490) as evidence of anxiety seems unwarranted. Conflating the word "stress" (Coryell & Clark, 2009, p. 490) with anxiety reveals a problematic theoretical perspective on the issue of anxiety that seems to have plagued the study of LLA since before Scovel (1978) over forty years ago.

The point being made here is that despite the evidence for the FLCAS being questionable, it has entered the field of linguistics and language teaching and been accepted as a fact, generating more research using it as an instrument. With each generation of subsequent research, the FLCAS becomes more cemented as a valid construct. One sees this clearly in the work of Sparks and Ganschow (2007), where despite finding that the FLCAS's results are stable over a period of nine years across multiple classrooms and with different instructors the results are accepted as valid, despite this evidence clearly strongly suggesting that the FLCAS is not measuring what it claims to measure.

The fundamental question remains whether foreign language learning anxiety even exists, and if it exists what shape it takes and whether it can be modelled.

2.5. Conclusion

The above argument sets LLA research back to Scovel in 1978. The field of LLA research posits the theoretical existence of language learning anxiety, but has no reliable evidence of its existence now that Horwitz's (1986) evidence has been reasonably shown to be questionable. The next chapter will address the complexities of measuring stress and anxiety with a view towards establishing whether the phenomenon of LLA can be verified in any sort of defensible manner.

3. The Measurement of Stress and Anxiety for LLA

The previous chapter reviewed the literature around LLA, exploring the previous research into, and evidence for, the existence of LLA. The conclusion was that despite more than 74 years of research, from Dunkel in 1947 and continuing to this day, no-one has offered any evidence for the existence of LLA that is robust, when considered scientifically. Much of the research has been based on the questionable use of psychometric tests designed to assess clinical anxiety (Dunkel, 1947; Scovel, 1978; Horwitz, et al., 1986: Horwitz, 1986).

In addition, there seems to be a great deal of confusion about what precisely is being measured, namely whether it is clinical anxiety, or sub-clinical anxiety, or something else like stress. The definitional debates, both within and between fields, and differing uses of the terms, make understanding the literature difficult, especially when authors do not or cannot define precisely what they mean by the terms they are using (Scovel, 1978; Horwitz, 1986; Trang, 2012).

The notion of language learning anxiety seems to possess a great deal of face validity, with many researchers proposing that it exists, however it seems rooted in some fundamental misunderstandings about the concept of anxiety, which will be explored in the following section.

3.1. Defining LLA

A major theoretical issue with current LLA theory is the tendency to conflate stress and anxiety into a single concept, as can be seen in Coryell and Clark (2009), but this confusion seems to be persistent. It is also evident in Scovel's (1978) work where the terms tenseness and tension are proposed. This problem is attributable, at least in part, to a great deal of semantic confusion over the precise definitions of these terms, with Sapolsky (2015) concluding that all the terms are valid descriptions of a complex and multi-faceted phenomenon.

While these two phenomena of stress and anxiety are linked in the Hebbian model (1955) there has been a great deal of research since that time that suggests that the problem is considerably more complex. There are many moderating variables, such as beliefs about control (Scott & Weems, 2014), genetic (Smoller, 2016) and epigenetic (Klengel & Binder, 2015) factors, social factors, exercise, coping strategies, personality, gender, dietary factors, pregnancy, physical health, mental health (Contrada & Baum, 2011). An exhaustive list of the factors involved in mediating the stress and anxiety relationship would take more time and space than can reasonably devoted to an analysis that, while fascinating, would ultimately lead one to the same frustrated conclusion as Scovel in 1978.

As noted earlier when discussing Lupien, et al. (2006), attempting to consider all of the contributing and moderating variables involved in the stress-anxiety dynamic would be a daunting task, involving genetic testing, questionnaires about the participants' beliefs about control, and many other factors.

In the context of LLA research this fine-grained approach has not been productive, as shown in Scovel (1978). The problem can be adequately expressed by considering stress and anxiety as gestalt phenomena where the whole is greater than the sum of its parts. Lupien, et al. (2006) suggest the variables involved interact not only with resultant stress, but also with one another producing feedback loops of great complexity, describing this as "when top-down processing meets bottom-up effects" (p. 584). Attempting to stop and break the phenomenon down into its component parts has thus far resulted in mixed findings that fail to capture resultant stress because it is not just a question of adding up the contributing factors. This is the problem that Scovel (1978) found when reviewing the literature on LLA. Scovel (1978) optimistically concluded that "the overwhelming intricacy of these intertwining systems should not deter us from the task of trying to discover natural patterns and continuities" (p. 140). More than forty years later these "natural patterns and continuities" (p. 140) still elude investigators of stress, with Sapolsky (2015) pointing to the "ubiquitous, but nonspecific, role of stress in psychiatric disorders" (p. 1347).

This is not to in any way dismiss the advances in the understanding of stress and anxiety that this sort of variable-focused research has produced, merely to propose that, as yet, it has not produced a model that is practical for research into complex multi-variate phenomena such as LLA.

In Salmon's (2001) discussion of the effects of exercise on stress and anxiety it is suggested that part of the problem may lie in the way data have been collected. Laboratory studies may yield very precise results using sensitive equipment and by limiting the number of variables under consideration. However, Salmon (2001) points out that this very process of limiting the variables and divorcing activities from their normal context substantially changes the results, characterising these results as questionable. Research into areas such as LLA, where the language learning classroom is social environment, cannot safely ignore variables such as social context without risking calling the results into question.

A further complication is that anxiety and stress are normally transitory states in healthy individuals. Some individuals may be predisposed towards more negative emotions, which is what is measured by the STAI trait scale (Spielberger, 1983; Barnes, et al., 2002), but this does not necessarily mean that they are permanently clinically anxious. Further, there is the issue of time and how perceptions of events may change over time. To refer

back to the earlier example of the roller coaster ride, one might display high levels of stress and anxiety before and during the experience, but afterwards recall the event with little anxiety and with a positive emotional valence.

The exception is where individuals experience clinically elevated levels of anxiety, an anxiety disorder, where a key diagnostic feature is the persistence of this phenomenon and the failure to recover from this anxiety over a protracted period of typically 6 months or more. The American Psychiatric Association's (2013) Diagnostic and Statistical Manual of Mental Disorders (5th ed.; DSM-5) clarifies the distinction between short-term fear or anxiety and what is necessary for a clinical diagnosis as follows:

> Anxiety disorders differ from developmentally normative fear or anxiety by being excessive or persisting beyond developmentally appropriate periods. They differ from transient fear or anxiety, often stress-induced, by being persistent (e.g., typically lasting 6 months or more), although the criterion for duration is intended as a general guide with allowance for some degree of flexibility and is sometimes of shorter duration in children (as in separation anxiety disorder and selective mutism). (p. 189)

The DSM-5 makes a distinction between transient anxiety, which is often linked to stress, and clinical anxiety which lasts for six or more months. The DSM-5 differs from the DSM-4 in that some stress-related conditions (post-traumatic stress disorder and acute stress disorder) have been removed from the anxiety disorders chapter and placed in a separate chapter related to stress and trauma. It should be noted that the organisation of the DSM-5 is not random, and that this close placement suggests that these disorders are related, but also a delinking of stress from anxiety when they are at clinical levels (Black & Grant, 2014).

This distinction between clinical and non-clinical anxiety seems critical, and speaks to a number of the core debates in LLA. Evans, et al. (2005) underscore the importance of distinguishing the border between clinical and non-clinical anxiety, writing that, “it is especially problematic to establish the limits between normal behaviour and pathology because when mild, anxiety plays an adaptive role in human development” (p. 161). Further, Evans, et al. (2005) go on to state that, “a person may be characterized by internal dysfunction but not qualify as having a disorder because no resultant harm occurs” (p. 164). When reading Evans, et al. (2005) it is important to bear in mind that the paper is authored by psychiatrists, who have a higher threshold for what they consider harm than in psychology.

This distinction is important when considering the face validity of LLA. Researchers in LLA, such as Horwitz, et al. (1986), report signs of dysfunction, such as, “When I'm in my

Spanish class I just freeze! I can't think of a thing when my teacher calls on me. My mind goes blank." (p. 125), which they characterise as anxiety, but where they seem to be making an error is in assuming that this is indicative of a clinical anxiety disorder that is resulting in harm. This assumption is not necessarily supported if the anxiety is sub-clinical.

Therefore, someone researching the phenomena of stress and anxiety would receive different results depending on when the measurement was taken, as sub-clinical stress is transient. This is a problem that has plagued many studies of stress and anxiety (Salmon, 2001; Lupien, et al., 2006). Laboratory studies removed mediating factors (such as social environment) and introduced additional factors (such as the stress and anxiety from being observed while hooked up to machines), that the results from these studies are of questionable value (Salmon, 2001).

Questionnaire or interview approaches also have had problems with assessing LLA. See the earlier discussion of Scovel's (1978) review. The temporal factor in evaluations of stress and anxiety is of particular importance when assessing LLA if the phenomenon is sub-clinical, and hence transitory. If it is transitory then when the interview or questionnaire was completed becomes important. However, if LLA is characterised by clinical levels of anxiety, one of the defining features of which is its persistence, then the timing of the questionnaire or interview becomes less of an issue (American Psychiatric Association, 2013).

Therefore, the critical question in gathering evidence about LLA is whether the phenomenon is clinical or sub-clinical. If the phenomenon is clinical then measurement requires the use of psychometric instruments specifically designed to detect the presence of anxiety. As noted in Furr (2017) for valid and reliable results this requires that the user be trained in the administration, scoring and interpretation of these tests. The persistent nature of clinical anxiety means that the use of these psychometric instruments isn't very time sensitive as anxiety will persist for a period of six or more months (American Psychiatric Association, 2013).

If the phenomenon is sub-clinical then the DSM-5 suggests that the anxiety is often stress-induced and temporary (American Psychiatric Association, 2013), and does not necessarily cause any harm (Evans, et al., 2005). In this case models such as Hebb (1955) become useful for modelling how stress and anxiety may interact. Sapolsky (2015) recommends that the term stress is very broad and that a specific type of stress should be specified. Scovel's (1978) review of the pre-FLCAS research suggests that approaching the topic of stress from the perspective of trying to count contributing factors to stress is unlikely to yield useful results in the context of LLA research. This conclusion is supported by the work of Lupien, et al. (2006), who show how these contributing factors may interact in

complex ways that change resultant stress levels. Therefore, in line with Sapolsky's (2015) recommendation that the type of stress being investigated be clearly delineated, the logical conclusion in investigating LLA is to measure resultant stress, the final total at the end of the complex equation of contributing variables.

When considering how to measure resultant stress in the context of LLA it is important to bear in mind that, as non-clinical anxiety is transitory (American Psychiatric Association, 2013), any measurement tool used needs to capture what may be a fleeting and changeable phenomenon. Considering Salmon's (2001) comments regarding the questionable results delivered by laboratory testing this approach seems ill-advised. Questionnaires and interviews are likewise potentially problematic given that the duration of the stressor's effect is unknown. Evans, et al. (2005) state that, "A useful rule of thumb for determining the diagnostic threshold is the person's ability to recover from anxiety and to remain anxiety-free when the provoking situation is absent" (p. 162), although the speed of recovery is unknown. Whether a ten-minute interview after a language learning class would or would not capture the phenomenon is unclear. Given Evans, et al.'s (2005) statement that in cases of non-clinical anxiety recovery will occur after the stressor is removed, it seems inadvisable to measure anxiety outside of the context where anxiety occurs. There is also the issue of introducing an additional variable into the stress equation by asking participants in research to stop and fill out a questionnaire or other assessment, which may also affect the stress equation in unknowable ways given the complex manner that intermediate variables interact in the stress process (Lupien, et al., 2006). Ideally therefore the method of measurement should be as close to naturalistic observation as possible, introducing as few contaminating variables as possible, and measuring stress levels in as close to real-time as possible.

There is a final consideration, which is that numbers are meaningless without context (Harford, 2020). For example, if someone registers a stress level of 70 out of 100 in a language learning class, that may create the impression that they are highly stressed in that context, leading to the conclusion that the person is experiencing elevated stress levels in language learning classes. However, if additional data is gathered about their everyday stress levels outside of class and the data shows that this person's stress levels are normally around 90 out of 100, then language learning classes are not places of elevated stress, but rather places of decreased relative stress. This is the sort of problem that Salmon (2001) highlights in their review of the literature around mood and exercise. One of the critical determinants of the effect of mood on exercise was not a question of an absolute quantity of exercise, but rather how the exercise quantity compared to the individual's normal pattern. Therefore, whatever stress measurement instrument is used, it should also be capable of

capturing information about stress levels across a broad range of contexts to allow for meaningful comparisons between contexts and interpretation of the data.

3.2. Heart Rate Variance and Stress Measurement

Recent technological advances have offered an alternative method of investigating stress in a manner in a naturalistic manner that does not interfere excessively with the participants' social environment. Nor does it require participants to stop and analyse their feelings by filling in a questionnaire, thereby possibly contaminating their perceptions through mindfulness of their emotional state (Carmody & Baer, 2008). Instead of attempting to adopt a causally oriented approach to the measurement of stress this new technology allows direct measures of the final resultant stress by monitoring heart rate variability (HRV).

Heart rate (HR) is a measure that should be familiar to most and can easily be assessed by placing a finger on one of the pulse points, most commonly in the wrist or neck, and counting the number of pulses per minute (beats per minute or bpm). This measure of HR is the most common, however there are small variations in the timing between beats. These variations are referred to as heart rate variability or variance (HRV) (Taelman, et al., 2009).

How and why this HRV occurs is of interest in the study of stress, and while the genesis is psychological this triggers a series of physiological changes summarised below by Taelman, et al. (2009):

> Although stress has a psychological origin, it affects several physiological processes in the human body. When a person is exposed to a stressor, the autonomic nervous (ANS) system is triggered: the parasympathetic nervous system is suppressed and the sympathetic nervous system is activated … . This results in the secretion of the hormones epinephrine and norepinephrine into the blood stream which leads to, for example, vasoconstriction of blood vessels, increased blood pressure, increased muscle tension and a change in heart rate (HR) and heart rate variability (HRV). This process is known as the 'fight-or-flight' reaction … . When the stressor is no longer present, a negative feedback system stops cortisol production in the body, and a sympathovagal balance is established through homeostasis between the parasympathetic (vagal) and sympathetic system. (p. 1366)

As with any discussion of this sort the extent of the interactions described are far more profound and far-reaching than the neatly bound summary that Taelman, et al. (2009) present. As discussed in Lupien, et al. (2006) there are multiple feedback loops in this process, not just the negative feedback system to stop the reaction mentioned by Taelman,

et al. (2009). It is here that one enters the realm of what Sapolsky (2015) refers to as, “the neurobiological and psychological space floating between” (p. 1348) the stressor and resultant stress.

Underlying Taelman, et al.’s (2009) neat summary are other reactions that Taelman, et al. (2009) do not discuss, such as the glucocorticoids responsible for converting norepinephrine into epinephrine. Glucocorticoids have different effects on different parts of the brain. For example, in the frontal cortex excessive glucocorticoids have been linked to decreased grey matter volume and decreased synaptic plasticity. This negatively affects processes such as behavioural regulation, motivation, short-term memory, and learning. These changes are also linked to a higher incidence of depression and other conditions such as post-traumatic stress disorder. However, in the amygdala the opposite is true, where glucocorticoids promote synaptic plasticity, and promote the creation of new synapses. While this may seem like a good thing it is important to remember that the amygdala is responsible for processing threat and fear reactions and is associated with the sort of flashbulb memories commonly seen in conditions such as post-traumatic stress disorder. In effect the glucocorticoid excess produced during each stress reaction makes it easier to learn to be fearful and anxious while negatively affecting motivation, behavioural regulation, and memory (Sapolsky, 2015).

This digression into the mediating processes associated with the stress reaction is relevant to LLA. It demonstrates the underlying mechanisms by which stress might affect learning if stress levels were significantly elevated in the language learning environment. It also explains how elevated stress levels might result in the genesis of anxious behaviours in sufficiently stressed individuals. In addition, it clarifies why measuring stress is a valid way to approach investigating the phenomenon of LLA. This addresses the question of validity, and the next section will move on to the question of the reliability of HRV as a measure of resultant stress.

A meta-analysis by Thayer, et al. (2012) of “neuroimaging studies on the relationship between heart rate variability and regional cerebral blood flow” (p. 1), particularly the amygdala and ventromedial prefrontal cortex, concluded that, “… this review highlights the importance of HRV as a potential marker of stress …” (p. 754). This conclusion is supported by a more recent meta-analysis by Kim, et al. (2018) of 37 papers with human participants where HRV was used as an objective measure of psychological stress, which concluded that, “… HRV is impacted by stress and supports its use for the objective assessment of psychological health and stress” (p. 235). The meta-analysis by Kim, et al., (2018) suggests that HRV allows a reliable “objective assessment” (p. 235) of resultant stress.

In the past this sort of measurement would have required equipment normally only found in a cardiologist's office, however relatively recent changes in the miniaturisation of technology have moved equipment capable of making this sort of fine measurement of HRV from a purely clinical environment to something that a person might wear without impeding normal daily activities.

3.3. Wearable Stress Measurement Devices

These devices are collectively referred to as *wearables*, or *wearable devices*, and are defined in Ometov, et al. (2021) as:

> … small electronic and mobile devices, or computers with wireless communications capability that are incorporated into gadgets, accessories, or clothes, which can be worn on the human body, or even invasive versions such as micro-chips or smart tattoos. (p. 1)

The first wearable was arguably the smartphone, and while the first smartphone was released in 1992 it was not until 15 years later in 2007 when the first iPhone was released that smartphones entered the mainstream (Tweedie, 2015). Wearable devices technology is very new, but has seen rapid and widespread acceptance, so much so that during the COVID-19 pandemic smartphones were sufficiently ubiquitous in some countries that they represented a viable tool for contract tracing. However, the high cost of the technology and its relative newness did not make it a universally viable solution, particularly in poorer countries and communities (Hernandez-Orallo, et al., 2020).

Ometov, et al. (2021) highlight the "convenient, seamless, portable" (p.2) nature of this technology, and while early devices such as smartphones were not always convenient or seamless, the technology has evolved rapidly. Smart watches and fitness trackers, such as those offered by companies like *Fitbit* and *Garmin*, replace familiar items which people are accustomed to wearing, such as watches. A major difference currently is that wearables require periodic recharging, however researchers are already working on self-charging smart clothing that would remove this requirement (Tian, et al., 2021).

Research by Hirten, et al. (2021) into the use of wearable devices to monitor inflammatory bowel disease for four hundred patients at the Mount Sinai Hospital in New York City, USA found that 42.7% of participants already used or had used a wearable device, and that 93.8% were willing to use a wearable device. Participants indicated that their preferred device was a wrist-worn device, and that they were prepared to use it daily.

A meta-analysis by Pang, et al. (2019) of nine studies using wearables to detect near falls in elderly patients shows that, at least in some areas of medicine, the use of wearables

is gaining traction, and they are seen as a reliable and valid research tool. There also seems to be a general acceptance of the role of wearables in medicine, with a survey of 19 Swiss general practitioners' perceptions of the use of wearables by Volpato, et al. (2021) showing that "Participants perceived wearables as user-friendly devices that could foster patients' empowerment and support them throughout behaviour change processes" (p. 8).

While some research has been done using these wearable devices, they are still a relatively new and rapidly evolving technology, and as such there are no common guidelines available on how they should be used. As has been discussed, the reliability and validity of research can often hinge not just on the goodness of a measure, but also on how and where data was gathered.

The first question is whether these devices deliver valid and reliable results. Pakhomov, et al. (2020) conducted research at the University of Minnesota on pre-examination stress in 18 students who already owned *Fitbit* wearable devices, obtaining seven days of data from these devices for analysis. Pakhomov, et al.'s (2020) research is important for three reasons. Firstly, Pakhomov, et al. (2020) conclude that, "These results are consistent with prior laboratory findings and indicate that consumer wearable fitness trackers could serve as a valuable source of information on exposure to psychosocial stressors encountered in the naturalistic environment." (p.1). This finding by Pakhomov, et al. (2020) is important in that it suggests that consumer wearable fitness trackers are delivering accurate and reliable results comparable to laboratory equipment. Secondly, while Pakhomov, et al. (2020) did also use a questionnaire, they state that this instrument was problematic. Recall bias, where individuals may recall events incompletely, or inaccurately was a factor. The questionnaires were also intrusive, interfering with the individual's natural processing of the stressor by requiring the individual to stop and fill out the questionnaire. Additionally, the questionnaires had to be completed multiple times a day. While Pakhomov et al. (2020) do not explore this final point in any detail it raises a range of possible issues, from placing an unreasonable time burden on research participants to participants getting bored with filling out the same questionnaire again and again. Pakhomov, et al. (2020) state that the wearable devices allowed a more naturalistic form of observation that did not cause the issues they encountered with the questionnaires. Thirdly, while Pakhomov, et al. (2020) do not claim to be establishing a guideline for the appropriate duration of a study into stress using wearables, their seven-day study did produce meaningful results, suggesting that this may be a viable length of observation.

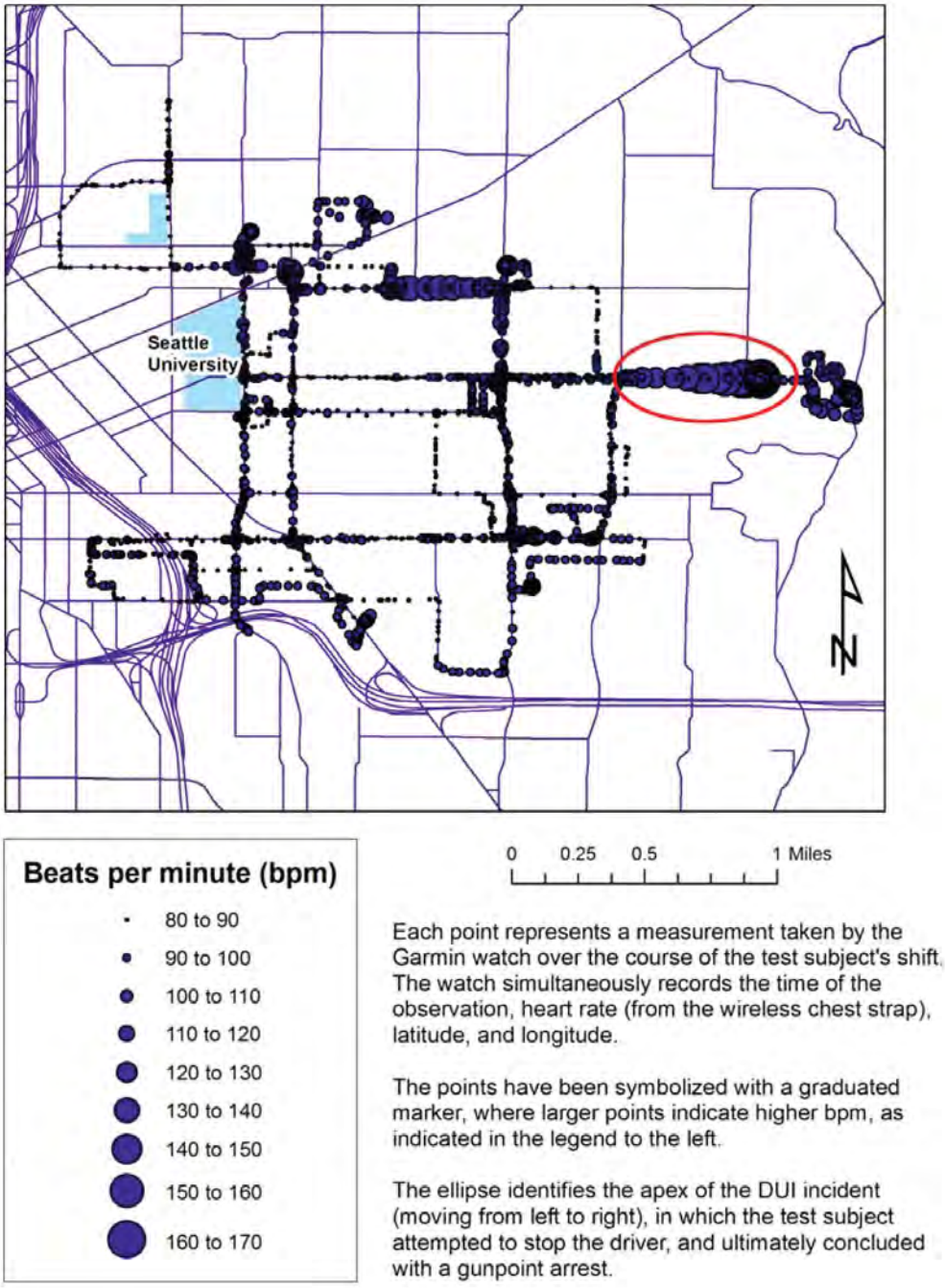
In an earlier study Hickman, et al. (2011) investigated on-the-job stress on police officers in Seattle, USA. The study by Hickman, et al. (2011) was a pilot study using a single

device with a single participant for a single shift to investigate the feasibility of further research using wearables. The device used in this study was a *Garmin* chest strap monitor which collected data on heart rate, and GPS location in real time. Hickman, et al.'s (2011) research is relevant to LLA because it highlights the important of contextualising stress data. In the paper Hickman, et al. (2011) were able to map the police officer's movements and cross-reference them to heart rate and what the police officer was doing at the time. Heart rate alone is not as reliable an indicator of stress as HRV (Taelman, et al., 2009), however the data gathered clearly shows increased heart rate approaching incidents, then the return to normal heart rate after incident.

One of the visualisations of the data from Hickman et al.'s (2011) study is shown below in Figure 6. What is interesting is how the data describes a narrative and shows how the police officer's heart rate increases as they receive the call and approach the driving under the influence incident, culminating in the arrest. Despite being technically quantitative studies there is a qualitative, almost narrative element, present in the type of data that can be gathered using wearables that permits elements of both quantitative and qualitative analysis. Wearable technology offers researchers a richer dataset that can be used to bridge the gap between quantitative and qualitative methods, offering a more holistic approach to research. Hickman, et al.'s (2011) paper illustrates the importance of real-time contextualised stress data in providing an accurate representation of what the individual was experiencing at that moment. It also demonstrates how quickly homeostatic systems return heart rate to normal levels, even after extremely stressful situations, and the importance of not only real-time data, but also contextualised data where measurement before, during and after exposure to a stressor can be considered.

Figure 6

Map of Data Collected via GPS-enabled Heart-Rate Monitor



Note. From "Mapping Police Stress" by Hickman, et al., 2011, in *Police Quarterly*, 14, page 239. Copyright Hickman, et al, 2011. Reprinted with permission.

In the same year Jovanov, et al. (2011) conducted a pilot study into occupational stress in nurses using a chest band heart rate variance monitor and a step monitor, both produced by *Garmin.* Details are sparse, for example the number of nurses involved is not stated and the results are more concerned with the feasibility and technical aspects than with drawing any

conclusions. This study is interesting because it reiterates the transient nature of subclinical stress. Jovanov, et al. (2011) also emphasise that absolute stress values above a certain value or cut-off are not as important as changes in stress values, i.e., whether the stress value is higher than normal or below normal. This is important because traditional stress measures, such as the FLCAS, measure stress once, and assume that stress is consistent in that context, however Jovanov, et al. (2011) and Hickman, et al. (2011) both show that stress levels fluctuate considerably, and that rather than being a stable phenomenon it is highly variable. What is also noteworthy in Jovanov, et al. (2011) is the calibration process where participants in the study sit and relax to establish resting heart rate variance as a baseline for further measurements. This process has been automated in newer devices, with devices establishing resting heart rate during sleep, so it is no longer necessary to do it manually, but the description of this process in Jovanov, et al. (2011) highlights the degree to which stress is an individual phenomenon. Even with HRV, which allows an "objective assessment" (Kim, et al., 2018, p. 235) of resultant stress, there are individual variables that need to be accounted for.

Since these early studies, research has continued using wearables. Akbar et al. (2021) published a study of stress data from 47 physicians at 5 medical centres in the USA. The goal of the research was to measure stress levels during administrative work, specifically electronic health recording, which is where patients' records are updated electronically as opposed to old paper and pen charts. This study is relevant because it addresses what minimum duration may be required for a study of this kind. Participants wore a *Garmin* Vivosmart 3 for a period of seven days, which measured stress through HRV. Of additional interest is that Akbar et al. (2021) cite this device as one that allows researchers to collect, "stress data through unobtrusive means that provide objective and continuous measures" (p. 2).

Looking at the research there is evidence that wearable devices are valid and reliable tools for the measurement of stress through HRV (Kim, et al. 2018) in a naturalistic manner (Pakhomov, et al., 2020). Further, they seem to be acceptable to both research participants (Hirten, et al., 2021: Pang, et al., 2019), and clinicians (Volpato, et al., 2021). There are no guidelines yet on the duration for which they should be worn, however outside of the pilot studies seven days seems to be the minimum period for meaningful results (Pakhomov, et al., 2020; Akbar et al., 2021). When combined with information on what the participant was doing at the time the stress data can be contextualised, delivering a rich data set that has qualitative characteristics (Hickman, et al., 2011). With continuous measurements the data allows for the measurement of changes in stress levels across time and contexts that gives the measurements meaning by allowing comparison to other times and contests (Jovanov,

2011). Therefore. there is evidence to suggest that wearable devices could deliver valid and reliable measurements of sub-clinical stress and anxiety in investigating the LLA phenomenon.

3.4. Conclusion

The previous two chapters have explored the concept of LLA from its earliest mentions and attempts at measurement to the latest techniques available.

The history of LLA has been characterised by definitional uncertainty. This definitional uncertainty can be seen clearly in the review by Scovel (1978), who seems to grasp that there are border conditions in anxiety. Scovel (1978) offers up distinctions such as facilitative and debilitative anxiety, and terms like tension and tenseness. It was only in the DSM-5 (The American Psychiatric Association, 2013) that sub-clinical stress, with periods of transient anxiety, and prolonged clinical anxiety were clearly separated.

The definitional uncertainty in the past research has informed the manner in which previous researchers attempted to measure LLA by utilising clinical measures of anxiety such as in Dunkel (1947), and the work of Horwitz, Horwitz, and Cope (reported in Horwitz, 1986). That there is evidence that suggests that these psychometric devices were poorly understood and were possibly inappropriately administered, scored, and interpreted by researchers not trained in their use has merely muddied the waters even further. In the previous chapter the evidence offered for the FLCAS was examined in detail, revealing numerous questions over whether the evidence is reliable and valid. This detailed examination was necessary because the FLCAS seems to be regarded as capturing the “reality” (Trang, 2012, p. 69) of LLA, and has become the basis for a great deal of research into LLA. The questions raised about the FLCAS, and research based thereon, sets LLA research back to Scovel (1978).

However, after decades of research the definitional uncertainty and confusion seems to be resolving itself as evidenced by the clearer distinction between sub-clinical stress and transient anxiety versus prolonged clinical anxiety offered in the DSM-5 (The American Psychiatric Association, 2013). Sapolsky’s (2015) clear delineation of what can be meant by the word stress, which in this thesis will refer to resultant stress, is also helpful. This offers a way forward for research into LLA as appropriate tools can be selected to measure the correct construct and deliver meaningful results.

For the measurement of clinical anxiety there are a large variety of tools available, and the key considerations seems to be three-fold:

1) That the test administrator be appropriately trained and familiar with the device and its administration, scoring, and interpretation.

2) That the anxiety is prolonged (The American Psychiatric Association, 2013), requiring a high test-retest reliability.

3) That the measure does not place an unreasonable burden on the research participants (Pakhomov, et al., 2020), requiring a measure that is quick and easy for participants to complete and that is not administered excessively frequently.

For the measurement of stress, with Hebb's (1947) model suggesting transient anxiety at higher levels of stress, the requirement list is longer. Given the reasons explored in the last paragraph of the previous section it seems that new technological advances in wearable devices offer a good fit with the requirements suggested by the literature, offering a defensible measure of sub-clinical stress and transient anxiety.

Having identified the critical clinical vs sub-clinical divide in anxiety research, and isolated key elements to be considered in choosing measurement devices, it becomes possible to address the question of research method, and what research questions can reasonably be asked and answered with the tools available.

Chapter 4: Research Methodology

This chapter will introduce the research questions asked in this thesis, the methods used to collect data and the manner in which data was collected, information about the participants and their selection, and the limitations of this study. Ethical issues will be discussed in a separate chapter as there are as yet no clear ethical guidelines for some of the issues associated with wearable devices, and so this section deserves special attention.

4.1. Introduction

Chapter 2 covered the history of language learning anxiety (LLA) theory, sometimes referred to as foreign language learning anxiety. It included an in-depth discussion of why the current evidence for the phenomenon of LLA is insufficient for any reasonable conclusions to be drawn, regarding the existence or non-existence of this phenomenon.

One of the core problems with the literature was the lack of a clear distinction between clinical and sub-clinical anxiety, leading to confusion over what measures of anxiety should be used. The new guidelines in the DSM-5 (American Psychological Association, 2013) clarify the distinction between clinical and sub-clinical anxiety, suggesting that they should be treated as separate, but related, phenomena.

In Chapter 3 this issue was explored more thoroughly, examining the key differences between clinical and sub-clinical anxiety. The most difference in the context of LLA being the long-term nature of clinical anxiety versus the transient nature of sub-clinical anxiety, and that the DSM-5 suggests that sub-clinical anxiety is often linked to stress. The diagnosis of clinical anxiety has a wealth of research supporting it and a wide range of tools available. However, sub-clinical anxiety has not been studied extensively as a result of its transient nature.

New wearable technology, which allows the monitoring of stress through HRV in an unobtrusive manner, was discussed in the context of how it might allow the capture of the LLA phenomenon in a naturalistic manner.

The means of measurement are critical to this thesis as they dictate what questions can or cannot be answered with any reasonable degree of confidence. This naturally leads to the subject of the research questions to be explored, but first a brief discussion of the research paradigm is necessary to situate the questions within an epistemological framework.

4.2. Research Paradigm

Before discussing the research questions, it is necessary to locate this research within a paradigm to provide a framework for understanding the research approach, design, and methods. The research framework chosen speaks to the epistemological orientation of the researcher. To shape this discussion Creswell's (2014) research frameworks will be used to provide reference points. Creswell (2014) discusses four frameworks namely, postpositivist, constructivist, transformative, and pragmatic. Only the postpositivist and pragmatic frameworks will be mentioned here as they align most closely with the approach taken in this research (Creswell, 2014).

Postpositivism, commonly referred to as the scientific method, is normally associated with quantitative research. Postpositivism seeks to uncover causal relationships between variables within a situation. It tends towards reductionism in that it seeks to isolate variables within a complex context, and test for these causal relationships. Underlying these tests is the belief that there is an underlying order to the world that can be observed, measured, and modelled. Postpositivism proposes that knowing the truth is impossible, and that all research is imperfect. Progress towards the unobtainable truth is made by the process of modifying, and if necessary abandoning, theories based on the best evidence available. Predictions (hypotheses) are made, and they are either shown to be false based on the evidence or accepted as provisionally true until better evidence is obtained that proves them false. Often truth is pursued through tests and experiments, in order to isolate variables. Aiming for objectivity is an important cornerstone of this research paradigm, and validity, reliability and minimisation of bias are important goals (Creswell, 2014).

By contrast, the pragmatic paradigm holds that the goal of research is to produce practical results and is more eclectic in its epistemological model than the other frameworks. It is more concerned with investigating a phenomenon in context, and considers that a phenomenon, "arises out of actions, situations, and consequences rather than antecedent conditions (as in Postpositivism)", (Creswell, 2014, p. 17). The pragmatic paradigm is often associated with mixed methods research, drawing on quantitative and/or qualitative data sources, and varying methods of analysis. As context is key, pragmatic research tries to capture data as naturalistically as possibly (Creswell, 2014; DiVincenzo, 2014).

This research contains elements of both the postpositivist and pragmatic approaches. As the earlier analysis of the literature in chapters two and three shows, this research does draw on some elements of the postpositivist paradigm. Earlier research is discussed in terms of postpositivist concepts such as validity and reliability. The underlying assumptions in these earlier chapters are postpositivist notions that hypotheses are false

unless reliable and valid evidence is presented for the hypotheses to be accepted as provisionally true. This research draws on other postpositivist notions, such as the rejection or modification of theory in pursuit of improved theories.

However, there are also important elements of the pragmatic paradigm present in this research as well. The methodology employed in this research was as close to naturalistic observation as possible. An assessor single-blind (Day & Altman, 2000; Kamper, 2018) research design was used, where the participants knew the identity of the researcher, but the researcher was unaware of the identity of the participants and was unable to link research data back to any participant. This was a necessary step given the small size of the department in which the research was conducted and the potential for the research data to contaminate the researcher's perception of students and their abilities (Rosenthal & Jacobson, 1968). The use of wearables to capture contextualising data in a naturalistic fashion aligns more cleanly with the pragmatic paradigm. The level of detail seen in the wearable data begins to approach the type of thick description that typifies qualitative, as opposed to quantitative, research, assuming narrative elements. This is traditionally associated with qualitative studies, but in this study are being accessed through quantitatively oriented data.

This is where this study transcends traditional boundaries and definitions. It is assumed that a mixed-methods approach is mixing or blending different types of data, such as narrative data from interviews and quantitative data from questionnaires. However, in this case the same data is being viewed through different lenses because the sheer density of the data lends itself to different types of analysis.

This is an issue that is difficult to express in terms of traditional language in the research methodology literature. The terms used are loaded with assumptions that seem to perpetuate notions about the division between quantitative and qualitative data types, and the idea that certain types of data lend themselves only to a single type of analysis. While the term used is mixed methods, there is an assumption that one is "'mixing' or combining the two forms of data in a study.", (Cresswell, 2014, p. 242), rather than subjecting the same dataset to different methods of analysis as in this study. Cresswell (2014) repeatedly assumes that mixed methods involves mixing different data types. Thus, this study's approach does not fit the definition of mixed methods advanced in the literature. Much of the methodology around, and critiques of, mixed methods are designed to address the limitations and problems introduced in mixing dissimilar datasets. Yet this is not the case in this research. A prime example of this is the issue of concurrent versus sequential data

collection (Cresswell, 2014), with no option for this being the same dataset being examined using different methods.

Locating this research within a traditional research paradigm is therefore difficult because the language used is loaded with assumptions regarding a divide between quantitative and qualitative research. This is problematic when presented with the quantity and quality of data provided by research instruments such as wearables.

It may be worth reflecting on the fact that new research paradigms emerge from older ones, such as Postpositivism emerging from positivism, and that technological limitations play a role in how research paradigms are constructed. Archetypal postpositivist scientists are shown in their laboratory conducting experiments not because it necessarily represents the best way to gather data, but simply because the equipment involved is not portable. As technology advances it would be surprising if research paradigms did not evolve, but it may take some time for the literature on research paradigms to reflect these changes.

This technological shift has profound implications for research paradigms, particularly the divide between quantitative and qualitative research. Polit and Beck (2010) summarise the divide as follows:

> Generalization, which is an act of reasoning that involves drawing broad inferences from particular observations, is widely acknowledged as a quality standard in quantitative research but is more controversial in qualitative research. The goal of most qualitative studies is not to generalize but rather to provide a rich, contextualized understanding of some aspect of human experience through the intensive study of particular cases. (p. 1451)

The issue of generalisation is core to the concept of research. While a detailed and rich account of a single person's experience may be indispensable in counselling environments. Polit and Beck (2010) point out that it cannot necessarily be held to be valid for other individuals. As such it raises question about what the research has really added to the field. Quantitative research is likewise subject to criticism on the basis that it is reductionist (Creswell, 2014) and minimises or ignores context and the richness of the human experience, and as such misses the essence of human experience.

One of the ideas explored in this research is that technology may be approaching a point where these two goals can be, at least partially, reconciled. New technology allows the gathering of rich and contextualised data on a scale that may be generalisable, delivering quantitative data that takes on qualitative aspects. While there were problems with this research one of the objectives was to serve as evidence of concept for this idea.

This research's epistemological approach fits within a postpositivist paradigm, drawing on notions that reliable and valid evidence needs to be presented to support hypotheses. However, the methodology employs elements of the pragmatic paradigm, utilising a mixed methods approach based on wearable data that provides the basis for both a contextualised narrative approach and a more traditional postpositivist statistical analysis. The issue of context is central to this thesis, and this represents the major departure point from the postpositivist paradigm. It is proposed that this is a natural paradigm shift associated with the increased portability of technology, and the new opportunities for contextualised research that it allows. This is not a departure from the postpositivist approach to epistemology, and at the risk of making predictions, it is a trend that is likely to become the norm in the not-so-distant future as technology becomes more portable.

4.3. Research Questions

The following section will address the research questions to be asked in this thesis. Each question will be presented and then briefly discussed to expand on the fuller implications of the question, as they apply to the methods and measures that should be used.

These questions evolved over the course of the research. The initial line of enquiry was sparked by concerns over the potentially pathologizing use of the word anxiety in LLA theory. There also seemed to be a contradiction between inverted U models of stress and performance (Hebb, 1955), and the straight-line negative correlation in LLA (Horwitz, et al., 1986).

This prompted a deeper examination of stress and anxiety, and an exploration of the definitional uncertainty. This exploration is detailed in Chapter 2. While this definitional uncertainty has been clarified to a certain extent by the DSM-5 (The American Psychiatric Association, 2013) this is not particularly helpful in determining what was intended before the clarification, where terms were used in an inconsistent manner. These definitional debates were not limited to LLA, but are found throughout the literature (Sapolsky, 2015). While there has been some refinement of terms and clarification of definitions, the relationship between stress and anxiety is still an area plagued by scholarship built on lines of discourse where foundational concepts are unclearly defined and operationalised. In pursuing this research, teasing apart the various threads in stress research was a major problem.

The research diary given to participants reflects this struggle. Attempts were made to ask participants to self-rate their own stress, in addition to the resultant stress measures based on HRV from the wearables. This line of enquiry was abandoned when, on reflecting on the lack of clarity around definitions of stress, it became evident that asking participants to rate their own stress would be unlikely to result in meaningful data. The literature also

revealed this confusion even amongst scholars writing on the topic of stress and anxiety. In Chapter 2 the research by Coryell and Clark (2009) is discussed, where in structured interviews participants feeling “silly” (p. 490) is labelled as evidence of anxiety.

The lack of definitional clarity is one of the areas that has been most difficult to navigate in this research. What has added to the complexity of this issue is that over time definitions of stress and anxiety have shifted. These shifts in the meanings of stress and anxiety represent decades of scholarship and are a positive development. However, they also create conceptual distance between the earlier LLA literature and what is understood by the meaning of these terms today; and make it more difficult to guess what the authors intended when using these terms. Definitions are offered, such as, “Anxiety is the subjective feeling of tension, apprehension, nervousness, and worry associated with an arousal of the autonomic nervous system.”, (Horwitz, et al., 1986, p. 125). However, as discussed in Chapter 2, these definitions seem overly broad. Focusing on the last part of the definition, consciousness involves some degree of arousal of the autonomic nervous system. The definition offered could be taken to characterise all conscious behaviour as containing some element of anxiety. This may be what Horwitz, et al. (1986) intended, with the FLCAS containing no potential for a zero score. However, it may be what they were referring to is what would be labelled stress today and that they were battling with the definitional uncertainty that often conflated these two concepts at the time when they were writing.

In all fairness to those early researchers, when this study began the researcher imagined that they had a clear idea of the dividing lines between concepts like stress and anxiety. However, a deeper exploration of the literature revealed that these terms are still less than clear.

In this thesis stress has been defined and operationalised as resultant stress, measured in terms of HRV using wearables. Anxiety is dealt with in two categories, namely sub-clinical anxiety, defined and operationalised as moments of high stress, and clinical anxiety defined and operationalised in terms of a clinical screening device.

The research questions are structured around the most popular model of LLA, namely that proposed by Horwitz, et al. (1986) which has the following features:

a. Situation-specific

b. Debilitative

c. Anxiety (heightened stress)

d. Anxiety (clinical anxiety)

e. FLCAS as a measure

An additional research question has been added investigating the use of wearable devices in stress and anxiety research.

The research questions are detailed below, with a brief mention of the measures to be used. The measures will be introduced in more detail after the questions.

Research Question 1: Is there evidence of a situation-specific stressor that exists only in the language learning environment?

As discussed in Chapter 2 one of the key features of LLA theory is the hypothesised existence of a stressor that is specific to the language learning environment. No evidence could be found in the literature of any attempt to prove this concept. There were no comparisons of stress or anxiety levels outside of the language learning environment, whether in other subjects or in life in general.

This question has profound implications for LLA theory, and for other theories about academic anxiety in other subjects, as the existence of a subject-specific stressor seems to be an unproven foundational assumption in the construction of these theories.

This thesis will investigate this question using resultant stress data obtained from wearable devices to establish whether there is evidence of higher stress levels in language learning classes than in other classes or in everyday life.

Research Question 2: What evidence is there of the effects of stress on language performance or language learning?

As discussed in Chapter 2 there is a mismatch between the straight-line debilitative relationship between stress and language performance and language learning theorised in LLA, and other models of stress and performance which theorise an inverted U relationship. There seems to be agreement in the literature that at high stress levels a debilitative effect is anticipated. Where the models differ is in terms of what happens at moderate or low stress levels. In LLA theory low stress represents peak performance, while in models such as that proposed by Hebb (1955) peak performance exists at moderate stress levels.

This lack of agreement is the subject of the second research question. The question was investigated using HRV data from wearable devices to measure resultant stress. The participants' TOEIC (Test of English for International Communication) results was used as an indicator of language performance, with participants' degree of improvement over the period of a year to indicate language learning.

Research Question 3: Is there evidence that clinical distress has an effect on language performance or language learning?

As discussed in Chapter 3, the DSM-5 (The American Psychological Association, 2013) has suggested that clinical levels of anxiety may not be stress-induced and has introduced a distinction between stress-related and anxiety-related conditions. In LLA theory it is unclear whether the levels of distress are clinical or sub-clinical. Horwitz, et al. (1986) include a section entitled, “clinical experience” (p. 126) and describing symptoms such as, “dread … and … palpitations” (p. 126) that could fall within the scope of a clinical condition, such as panic disorder. LLA theory is not clear on whether the distress described is clinical or sub-clinical. As such the possibility exists that LLA is a clinical phenomenon that may not be stress induced. This distinction between clinical and sub-clinical anxiety is critical in choosing the correct measurement device, as the DSM-5 suggests the distress may not be stress-related. If the distress is clinical, then psychometric tests to assess the presence of clinical distress should be used. However, anxiety disorders are linked by similar symptoms, not by similar causes.

Underlying this research question is one of the more troubling assumptions of LLA theory, namely that the anxious behaviours reported in the language learning classroom are all attributable to a common cause, namely the language learning environment. Diagnosis of a specific type of anxiety is not something that can be done on the basis of a single psychometric test. It is a multi-step process involving expert judgements, often from multiple experts, such as a medical doctor to rule out other medical conditions, and a clinical psychologist or psychiatrist. A specific diagnosis of the type of clinical anxiety present will not be attempted for practical reasons.

It cannot be assumed that, were the clinical anxiety diagnosed, it would be a specific phobia to language learning classrooms. Indeed, if Horwitz’s (1986) use of the fear of negative evaluation scale is accurate, it suggests that a common feature amongst those experiencing LLA is a fear of negative evaluation, which might edge a clinician towards a diagnosis of social anxiety disorder under the guidelines for differential diagnosis presented in the DSM-5 (American Psychological Association, 2013). This example is not intended as an attempt at diagnosis, but rather merely to emphasise the delicacy of the diagnostic process and the difficulty in distinguishing not only between types of anxiety without a formal diagnosis.

In raising the issue of fear of negative evaluation, Horwitz (1986) has broadened the range of possibilities beyond anxiety, as fear of negative evaluation is also a factor in conditions other than anxiety. This brings up the possibility that while LLA was initially

theorised as a type of anxiety it may fall under a different classification. Referring back to the example of social anxiety disorder, the differential diagnosis lists more than fifteen conditions, of which less than half are classified as anxiety disorders. Consequently, the measure used in this study needs to not only encompass the possibility that LLA may be describing a clinical condition which may not be stress-induced, but also that the condition may not be classified as an anxiety disorder.

As in the previous question language proficiency and language performance were assessed in terms of participants' TOEIC results. Clinical distress was assessed using the 28-item version of the General Health Questionnaire (GHQ-28). As participants could not be assessed by a health professional to determine a specific diagnosis, the GHQ-28's sub-scales will be used to provide an indication of the chief complaint, which should not be confused with a diagnosis.

Research Question 4: Is there evidence that the FLCAS accurately predicts clinical distress, stress, language performance, or language learning?

As detailed in Chapter 2, there are serious questions over the validity and reliability of the FLCAS. However, as the current dominant measure of LLA in the literature, the FLCAS will be assessed against the other evidence gathered to see if the FLCAS is an accurate predictor of clinical distress, stress, language performance, or language learning.

Research Question 5: Is there evidence about how wearable devices can or should be used in stress research?

Wearable technology is very new, and the literature indicates a relative paucity of stress research using wearables. To underscore this point, Hickey, et al.'s (2021) meta-analysis of research using wearables to monitor stress and mental health conditions found only 21 qualifying studies at the time of the writing of their meta-analysis, although more studies have been published since then. The small number of studies means that there is very little guidance for researchers considering using these devices in terms of how long they should be used for, the potential problems, and other practical considerations. Therefore, one of the goals of this thesis will be to reflect on the issue encountered doing this research, and will hopefully provide some guidance to future researchers using wearables in their research on problems encountered and other issues around the use of these devices.

4.4. Research Measures

This section will explore the measures to be used in this research, explaining why each measure was selected and how it addresses the research questions.

4.4.1. The General Health Questionnaire 28-item version (GHQ-28)

The GHQ-28 is a self-administered psychometric device used in clinical settings, such as hospitals, and has been translated and validated versions in 38 languages (Sterling, 2011). It is suitable for use with both adults and adolescents (Goldberg, 2013).

It is important to note that the GHQ-28 is only intended for use as a screening device "… to detect those likely to have or to be at risk of developing psychiatric disorders" (Sterling, 2011, p. 259). The use of the word clinical should not be taken to indicate that the GHQ-28 is being misrepresented as being equivalent to a diagnosis by a qualified clinician. The use of the GHQ-28 as an indicator of clinical distress in this research is offered with the caveat that in a clinical setting, such as a hospital, clinical, or psychologist's offices, the use of the GHQ-28 would be followed by a more complete examination. A more complete medical and psychological examination of participants was not possible in this study.

However, the GHQ-28 is regarded as a reliable and valid instrument according to Sterling (2011):

> Numerous studies have investigated reliability and validity of the GHQ-28 in various clinical populations. Test-retest reliability has been reported to be high (0.78 to 0 0.9) (Robinson and Price 1982) and interrater and intrarater reliability have both been shown to be excellent (Cronbach's α 0.9–0.95) (Failde and Ramos 2000). High internal consistency has also been reported (Failde and Ramos 2000). The GHQ-28 correlates well with the Hospital Depression and Anxiety Scale (HADS) (Sakakibara et al. 2009) and other measures of depression (Robinson and Price 1982). (p. 259)

For this study, the Japanese translation of the GHQ-28 was used. This version has been validated for use in Japan, and for the various subscales (Iwata & Saito, 1992; Suda, et al., 2007; Goldberg, 2013). There are four sub-scales to the GHQ-28, namely somatic symptoms, anxiety and insomnia, social dysfunction, and severe depression. There are some important limitations regarding the scoring of the GHQ-28 and the appropriate use of the sub-scales.

Scoring of the GHQ-28 can be conducted in two different ways, either using a Likert Scale (0-1-2-3), or the standard GHQ-28 scoring (0-0-1-1). The manual permits both scoring systems and the reliability and validity of the test is not impaired using either scoring system. The Likert scoring was used for this study as it provided more granularity in the results on the sub-scales. The total score on the GHQ-28 suggests a clinical level of distress if a threshold value of 24 is met on the Likert scoring. This threshold is an indicator of caseness. The term caseness distinguishes between cases of clinical concern and those where the

level of distress is likely to be subclinical. Higher scores above this threshold value do not necessarily indicate greater levels of distress, and once the threshold is met all that can be said is that there is likely a clinical level of distress of some kind. The likelihood is related to the reliability of the test, which is between 0.88 and 0.97. In this study the most conservative position was taken that the GHQ-28 might be in error in approximately 12% of cases (Goldberg, 2013).

The sub-scales are for screening purposes in identifying the chief complaint, with the highest score being the predominant source of distress reported. However, this does not rule out comorbidities, so just because someone scores highest on somatic symptoms, such as pain or other physical symptoms, it does not necessarily mean that they are not also severely depressed. There is no threshold value for each individual sub-scale. For example, if an individual scores 26 on the GHQ-28, which is above the threshold for clinical distress, with scores of 14 on the social dysfunction and 12 on the anxiety and insomnia sub-scales. It does not necessarily indicate the presence of a clinical level of anxiety. This result could be explained in a number of ways, such as that the pain from an injury was interfering with the individual's sleep, or that the individual was experiencing shortness of breath and was anxious about what this might mean. In both cases the anxiety and insomnia would not be disproportionate or unreasonable. The interpretation of the GHQ-28 sub-scales is limited, and while they do indicate the chief complaint, the lack of threshold values on the sub-scales means that the results cannot legitimately be interpreted to indicate that a clinical level of distress exists in a secondary area. Nor is the GHQ-28 a diagnostic instrument. As discussed earlier, diagnosis requires a precision and expert judgements. Whilst individual scores that are highest in the area of anxiety and insomnia suggest that this is the chief complaint. This should not be assumed to be the same as a clinical diagnosis of an anxiety disorder (American Psychiatric Association, 2013; Goldberg, 2013). The reasons for using the GHQ-28 in this study were four-fold.

The first reason was related to Research Question 3, to determine whether the degree of distress experienced a participant exceeded the threshold for caseness, and as such might be considered a clinical case. This was important because the DSM-5 (The American Psychological Association, 2013), suggests that stress may not be a causative factor in clinical anxiety. It is unclear in the LLA literature if LLA refers to a clinical or sub-clinical level of distress, and so it was necessary to include a measure that allowed an assessment of clinical distress.

The second reason for using the GHQ-28 was that it provided multiple sub-scales allowing screening for clinical levels of distress where symptoms other than anxiety were the

primary symptom. This use related to Research Question 3. This was important since Horwitz's (1986) research raised the possibility that LLA might arise from something other than anxiety, such as depression (as suggested by the significant positive correlation with the state-trait inventory's trait scale). Fear of negative evaluation (as suggested by the significant positive correlation with the fear of negative evaluation scale), also raises the possibility that social dysfunction is involved in some way (Horwitz, 1986). Therefore, the GHQ-28, which screens for clinical distress in general rather than only anxiety, was an ideal instrument.

Thirdly, while every effort was made to keep the research methods of this study as close to naturalistic observation as possible, the use of wearable devices was at the time a new and untested research method. The model of stress and anxiety presented in Lupien, et al. (2006) raised the possibility that participants' increased awareness of their own stress levels via the wearable devices might produce a feedback loop that heightened stress and so caused harm. The GHQ-28 was used to monitor the participants for any negative impacts on their mental health. This repeated use of the GHQ-28 is one of the special features of the GHQ-28, which is designed to be sensitive to short-term changes and suitable for repeated applications as a means of monitoring changes in the same individual (Goldberg, 2013).

Fourthly, the GHQ-28 is designed to be self-administered with no time limit, although the manual suggested that it should be possible for participants to complete the test in under ten minutes. This was a consideration in selecting the 28-item version of the GHQ, as there was a concern about posing an unreasonable time burden on participants as the assessment was scheduled to be completed three times over the six-week period of the study. The first administration was in the first week of the study, the second in the third week of the study after the participants began to wear the devices, and the third time was in the final week of the study, a week after the participants had stopped wearing the wearable devices.

In order to ensure correct administration, scoring, and interpretation of the Japanese version of the GHQ-28, the Japanese version of the GHQ manual was purchased and read carefully. There are important differences between the English and Japanese version. For example, in the English version of the GHQ-28 the items are grouped by sub-scales and presented sequentially. However, in the Japanese version items on different sub-scales are not presented sequentially. The Japanese official versions of the test were purchased and include a scoring template that reduces the chance of an error in scoring the different sub-scales. The scoring template was not included in the materials distributed to participants, nor

were participants given instructions on scoring or interpreting the GHQ-28. It was clearly stated that no feedback would be provided to participants on their GHQ-28 scores.

The results of the GHQ-28 are presented in the next chapter with detailed discussion of the ethical considerations involved in this research, rather than in the results section.

4.4.2. The FLCAS

While Chapter 2 raised some concerns about the reliability and validity of the FLCAS, it is nonetheless currently the dominant assessment for LLA. As such it would have been remiss to exclude it from this study. The use of the FLCAS in this study relates to Research Question 6, namely whether any significant relationships can be found between the FLCAS and the other aspects of LLA to be investigated. If the FLCAS shows significant correlations with one or more of these measures of anxiety, whether clinical or sub-clinical, or language learning or performance, then the FLCAS may still be of use in LLA research.

A translated version of the FLCAS was used for this study. As noted earlier in Chapter 2, the original FLCAS did not include a scoring method, so in this study the scoring method recommended by the authors of the translated version was used (Yashima, et al., 2007). The core issue in this study was the issue of concurrent validity, namely whether the translated FLCAS delivered similar scores to the original. Brown, et al. (2001) checked the concurrent validity of the translated Japanese FLCAS against the Horwitz's (1986) original FLCAS with 320 students who were, "Japanese nationals enrolled in the Intensive English Programme at Temple University Japan in Tokyo" (p. 361). Brown, et al. (2001) found a Cronbach's coefficient alpha of between 0.92 to 0.93 with the original FLCAS after adjusting the phrasing during pilot testing. This indicates a high degree of agreement between scores on the original FLCAS and the Japanese version.

This item was assessed only once, with previous research having suggested that scores on the FLCAS were stable over the long-term, and that asking participants to complete the FLCAS multiple times would be unnecessary (Sparks & Ganschow, 2007).

No instructions were provided on the scoring or interpretation of the FLCAS, as there were concerns about the validity of this measure, and it would be unethical to expose participants to feedback of questionable validity. It was made clear that there would be no feedback to participants on their FLCAS scores.

4.4.3. The Test of English for International Communication (TOEIC)

The TOEIC test was used to assess both language learning and language performance specifically to answer Research Question 2, regarding how stress affects language learning

or language performance. However, measuring language learning and performance was also a factor in Research Questions 2, 3, and 4. The measurement of language performance and language learning is central to the assessment of the LLA construct, which hypothesises that anxiety plays some role in the efficiency of language learning and/or performance.

The TOEIC was used for this study because participants in this study completed this test as part of their course requirements once a year, allowing both a measure of current level and a measure of improvement over time. The participants' TOEIC tests were paid for by the university, and as such asking participants to report their TOEIC test scores posed no additional cost or time burden on the participants, beyond the time taken to check their records and write down the test scores achieved. Participants are required by departmental policy to achieve a TOEIC score of 750 or higher to graduate, and a higher TOEIC score is advantageous in job hunting. As such it is presumed that participants had good reason to take the TOEIC test seriously and that their scores represent their best efforts.

The measurement of language competency is hotly debated, especially large-scale, commercialised tests such as the TOIEC, with valid and reasonable concerns over whether these tests actually measure language competency at all (McNamara, et al., 2019). The TOEIC test is controversial in that it is a norms-referenced test, meaning that the test norms participant scores against a bell curve. This means that the score does not necessarily reflect absolute competence against a pre-defined standard, but rather a relative score in a national ranking against all the other test takers (Bessette, 2007; Im & Chang, 2019).

This raises arguments that if in a theoretical pool of test takers every participant's competence improved by a similar amount, then individual scores will remain static. This is a valid argument in theory, however if one looks at the situation in the university where this study was conducted as a microcosm of the national TOEIC pool, then each year new first-year students enter the pool, and their scores are normed against second- and third-year students who are within the same pool. This allows a year-on-year comparison of how much learning is taking place. While the criticisms of the TOEIC as not necessarily indicating absolute competence are fair and valid, it is however a good indicator of the amount of learning taking place, and of competence relative to a pool of the participant's peers.

It should also be noted that while the TOEIC is not a perfect measure of language proficiency, it does provide a defensible third-party standardised measurement instrument that all participants are being assessed against under the same testing conditions and against the same items. This removes issues of inter-rater reliability when comparing scores from students in different classes being instructed by different professors who may have different standards. This was a noteworthy problem in previous studies of LLA, such as

Horwitz (1986), where scores from different classes in different languages and taught by different instructors were assumed to be reliable and comparable. Without some sort of moderating instrument this does not seem to be a defensible assumption, particularly in language assessment where there is often a large subjective element in scoring (McNamara, et al., 2019).

Therefore, while there is no attempt to portray the TOEIC as a perfect measure of language proficiency, it is a measure that provides a degree of standardisation that allows for defensible measurements of language performance and learning. In the context of this thesis the TOEIC test results are the operational measures of language performance and learning and will be discussed as such. For example, where a participant scores zero improvement in their TOEIC result it will be stated that there is no evidence of language learning. This is not meant to indicate that the participant has learned nothing, merely that in the context of the measures used in this study there is no evidence of learning. This is an important distinction that needs to be borne in mind when reading the results section.

4.4.4. Garmin's Vivosmart 3 and 4 Fitness Trackers

In Chapter 3 the use of wearable devices to measure HRV and thereby resultant stress and stress-induced anxiety was discussed. These wearable devices were used to investigate Research Question 1 by measuring sub-clinical resultant stress and anxiety, to gather data using HRV on whether there was a link between sub-clinical stress-induced anxiety and LLA. The wearable devices also provided data regarding Research Question 5 about the suitability of these type of wearable devices for stress research.

Hirten, et al. (2021) suggested that wrist-worn devices were most acceptable to research participants, so for this study only wrist-worn devices were considered.

A pilot study was attempted using the *Xiaomi Mi Band*, with a view to establishing whether sufficiently detailed heart rate information could be extracted to analyse HRV, however, despite assistance from a programmer, it was not possible to create an application programming interface (API) to extract the data, nor did requests to *Xiaomi* to allow access to their API meet with any response. The *Xiaomi Mi Band* was considered, despite the additional difficulties, because at approximately $30 it was the cheapest fitness watch on the market at the time.

At the time when this study was initiated in 2018, *Garmin* was the only major wrist-worn wearable manufacturer offering a standard 24-hour HRV-based stress monitoring function. *Fitbit* only released this function in the *Fitbit Versa 2* in 2019. While other *Garmin* wrist-worn devices, such as the *Tactix Charlie* and the *Vivoactive 3* offered HRV-stress

monitoring the *Vivosmart 3* was selected for this study. The other models were considerably more expensive, with the *Tactix Charlie* retailing for US$749.99 and the *Vivoactive 3* retailing for US$299.99, while the *Vivosmart 3* retailed for US$119.99. The *Vivoactive 3* and *Tactix Charlie* also both included GPS tracking, while the *Vivosmart 3* did not include a GPS tracking function. In the context of this study the lack of GPS tracking was advantageous as it enhanced participant anonymity, as participants' data would not be flagged with their home address or other identifying information. It also removed potential concerns about participants being identified by their data should the *Garmin* website be compromised. No information capable of identifying participants was stored on the server.

Garmin already had an API available for converting data gathered by the fitness watch into Excel-readable spreadsheets. In addition, the website had a function for directly downloading the data manually, and automatically generated graphical representations of data. *Garmin* was also responsive to enquiries regarding technical questions on the formatting and interpretation of the data. It should be noted that *Fitbit* was also very responsive when approached about using their devices in research, however as mentioned before they did not have a device that offered HRV-based stress measurements as a standard function in 2018 when the study started.

Finally, both *Garmin and Fitbit* indicated a willingness to quarantine the data used in the study on a secure server usually used for minors' data - that offered additional layers of protection against the exploitation of the data either by hackers or for commercial purposes, and clarified that data would be deleted permanently from the servers on conclusion of the study.

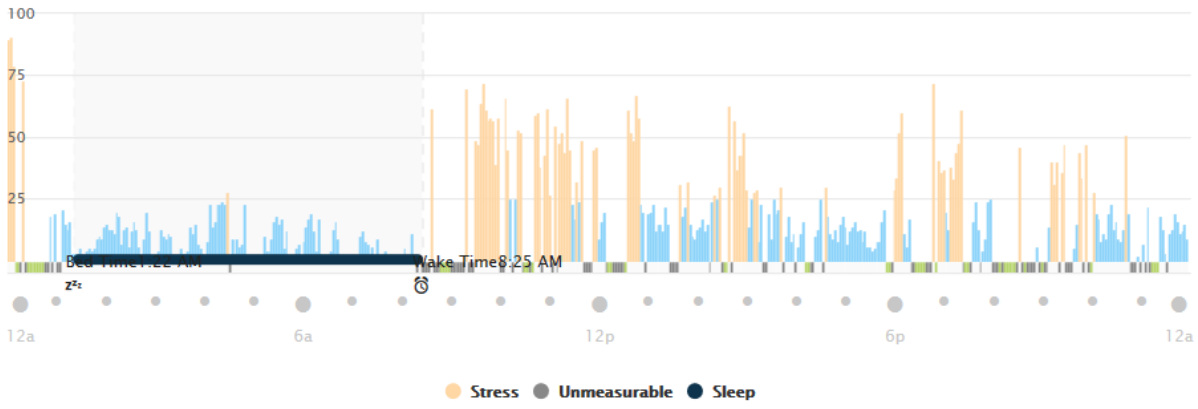
In the interests of not misleading future researchers this should not be read to imply that the data management process was easy or automated. Each day of data for each user had to be manually downloaded. Then each file had to be converted twice, once from *Garmin*'s proprietary .fit file format into a comma-separated value (.csv) file that was readable by Excel, then these files had to be imported into Excel, however some of the data was not compatible with the standard data formats. For example, *Garmin* used a special time format that counted from the founding of their company rather than any of the commonly used dates. As a result of this each day's file, many of which exceeded 10,000 lines of data, had to be manually adjusted for each user to allow analysis. Over the course of the study *Garmin* released several software updates that altered the conversion process. A programmer had to be consulted several times during the study to deal with unexpected problems in the conversion process. The process did become easier towards the end of the research, process as *Garmin*'s API became more sophisticated and capable of delivering

more easily readable data, but any researcher contemplating this type of research would be well advised seek assistance from an expert in computer programming.

At any point in the study during which participants were wearing the smartwatches they could obtain real-time data about their stress levels, step count, heart rate, hours of sleep, and other health data by looking at a free application on their smartphone. An example of the stress data output is included below:

Figure 7

An Example of Stress Data from the Garmin Vivosmart 3/4



Note. This figure was generated from the data collected by a participant's *Garmin Vivosmart 3* and is one of the automatically generated graphs available from the *Garmin Connect* website (https://connect.garmin.com/).

Figure 7 shows the stress data gathered from the *Garmin Vivosmart 3*, although the output is the same for the *Garmin Vivosmart 4*. The *Garmin Vivosmart* takes a stress reading every 3 minutes, for a total of 20 stress readings per hour under ideal conditions, however if there is significant movement (such as rapid movement, sitting or standing) within that 3-minute period the measurement is skipped. This is in line with the best practices for HRV measurement in order to obtain the most accurate measurements (Taelman, 2008). The small grey lines bottom of the graph, below the stress data and above the time index, indicate movement that was sufficient to prevent a stress reading from being taken. The green lines indicate periods of rapid acceleration, such as running, and during these periods stress readings are likewise not taken. The solid black line on the bottom of the graph shows when the participant was sleeping. Above that are the stress readings based on a combination of HRV using the participants' resting HRV during sleep to calibrate the sensors. There HRV-based stress readings appear in the main portion of the graph, on a

scale from 0 to 100. Stress readings from 0 to 25 are in blue, indicating low levels of stress normally associated with rest, while stress levels 26 and above are in orange, indicating levels of stress normally associated with people engaged in activity. Stress levels are grouped descriptively into low stress (26 to 50), medium stress (51 to 75), and high stress (76 to 100), although there is no research supporting any objective basis for these groups.

Concerns have been raised about the possible existence of a stress feedback reaction, resulting in elevated stress levels as a result of participants seeing their stress levels. It is important to note that at the time of the start of this study, stress monitoring was a new function for fitness bands, and had yet to attract any significant research. The existing research at the time indicated that, at least in the short-term such as the usage planned in this study, that wearable technology offered users significant short-term health benefits such as increased exercise, an area known to have a positive effect on stress management, and therefore the literature suggested that the use of these fitness bands was likely to produce positive effects for the wearers, and was therefore ethically defensible (Bravata, et al., 2007; Klasnja & Pratt, 2012; Cadmus-Bertram, et al. 2015; Jauho, et al., 2015; Poirier, et al., 2016). Subsequent research into the use of wearable stress monitoring devices has found weak and preliminary evidence of some minor benefits in stress management, and no evidence of the theorised adverse stress feedback reaction, although this is still a new area of research where no clear consensus has emerged (Smith, et al., 2020; De Witte, et al., 2019).

Initially 22 *Garmin Vivosmart 3*s were purchased, with 20 issued to participants and two held in reserve, to provide immediate replacements in case of loss or damage. Later an additional seven *Garmin Vivosmart 4*s needed to be purchased, to replace lost or damaged smartwatches. The *Garmin Vivosmart 4*s were purchased because *Garmin Vivosmart 3*s were no longer available, and only after checking that there were no changes in the stress measurement system between the two models. At the end of the study only 19 smartwatches remained in functional condition. The primary reason for this was non-return of the smartwatches.

4.4.5. Participant Schedule and Diary

An example copy of the participant schedule and diary is included in Appendix A. It should be noted that the participant schedule and diary included in Appendix A differed slightly between administrations in details such as dates. The dates provided in the copy in Appendix A are from the final administration. The version provided is in English where possible, however some items, such as the Japanese version of the FLCAS have not been translated. The Japanese version of the manual is available on request. Formatting varies

slightly between the versions because of fonts and character spacing. The author has provided notes in red textboxes where the version differs from the original, or where an explanatory note is required.

As the *Garmin Vivosmart 3/4* did not include a GPS function, it was necessary to ask participants to report on their regular schedule and any deviations from their schedule. This was done to link the data from the wearable devices to the participants' activities at a particular time to gather data in relation to Research Question 5, linking resultant stress to a particular activity.

The means for collecting this data was a participant diary. The diary gave detailed instructions for the schedule to be followed, dates on which questionnaires were to be completed (such as the FLCAS and GHQ-28), some preliminary screening questions about biographical data (age, gender, presence of any heart conditions that might affect their HRV readings), and a general schedule so that stress data could later to paired with what participants were doing at that time. A day-by-day diary section was also included so that participants could note absences from class, illness, and other factors that might influence their stress levels on that day.

The participant diary also included a space for the participant to write their login information (username and password) so that the data from the wearable device they had been issued could be downloaded. Participants were instructed not to use their real names or anything that could identify them for the username, and to use a unique password.

An additional detachable copy of the consent form for the study was also included in the back of the diary, as participants were not required to complete the consent form before taking the diary and briefing materials, but exchanged the consent form for the *Garmin Vivosmart 3/4*. The reasons for this will be discussed in the ethics section, but there was no link between the detachable consent form, which necessarily contained identifying information, and the participant's research data.

Activities were coded in data analysis into six generic categories to preserve participant confidentiality and anonymity, as specific combinations of specific subjects might identify a participant. The categories used were as follows:

English Classes. This was the category for any English language learning class, and included classes such as Oral Communication 1, 2, 3, and 4, Debate, (English) Academic Writing, (English) Presentation Skills, Current Affairs 1 and 2, and (English) Writing 1 and 2. All these courses included a large speaking component, even the writing and academic writing classes where participants read and discuss their work in English. In

addition, all these courses were taught by English mother tongue speakers almost exclusively in English, with perhaps the occasional clarification of an instruction in Japanese. The primary source of Japanese in the classroom would have been crosstalk between students in Japanese, and students are encouraged to hold these discussions in English.

Other Language Classes. This was the category for any foreign language learning class other than English and excluded classes where the language of instruction was Japanese. This category included language classes such as Chinese, Korean, Spanish, French, and German. The language of instruction in these courses is primarily the target language, however there is a higher Japanese content as some students had no background in these languages, and while the instructors are mother-tongue speakers of the target language they are also frequently bilingual. As such, these courses were coded in a separate category.

Non-language Classes. This was the category for any classes where language learning was not the primary goal of the course, or where the primary language of instruction was Japanese. These included a large variety of courses, such as Economics, Politics, Media Analysis, Law, Social Studies, and the smaller seminar classes. There were some courses where the distinction was not entirely clear, such as the course on Chinese Politics, where some of the reading materials were in Chinese and the instructor was a bilingual speaker of Chinese and Japanese, and sometimes spoke in Chinese. After consultation with the Professor concerned it was determined that the primary language of instruction was Japanese, and that the goal of the course was not primarily to teach Chinese to the students.

Free Time. This category included any time not in classes, with the exception of sleeping time and time spent at part-time jobs. This was a very broad category. While additional information on what the participants were doing may have been desirable, it also raised difficulties regarding the participants' privacy and possibly imposing an undue burden on participants in noting their activities. It also would have complicated data coding and analysis. As this data was primarily being gathered to contextualise classroom stress levels and provide a general idea of participants' normal stress levels outside of class, additional detail on precise activities was not regarded as being sufficiently important to outweigh the problems with gathering this data.

Part-Time Job. Discussions with students suggested that part-time jobs could be very stressful, and so this category might skew the general free time data if not put into a separate category. It was decided that in this specific case it was necessary to ask participants to report time spent at part-time jobs.

Sleeping. Sleeping times were coded separately, as by definition these are rest periods characterised by low stress/arousal and the inclusion of this data in the averages for any other category would skew the data. Participants were not asked to report in their diaries on sleeping times. One of the advertised features of the *Garmin Vivosmart 3/4* was that it was able to automatically detect sleeping times. In retrospect this was an error, as some of the participants displayed erratic sleep schedules that posed some difficulties for the *Garmin Vivosmart 3/4s* software. These difficulties will be discussed in more detail later.

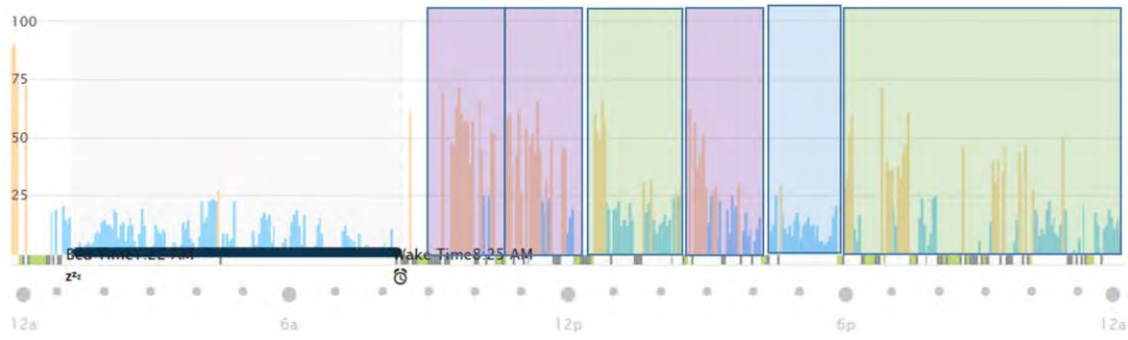
In the results section the various categories (except for sleeping times, which are annotated with a black line at the bottom of the stress graph), have been colour-coded as follows:

Coding Key				
English classes	Other language	Non-language	Free Time	Part-Time Job

These colour codes have then been superimposed onto the stress data from the *Garmin Vivosmart 3/4s* to create a visual link between stress levels and activities:

Figure 8

Garmin Vivosmart 3/4 Stress Data with Activity Codes Superimposed



The schedule for the research contained in the participant diary was as follows. At the beginning of the research the participant was asked to write their age, gender, results of their first TOEIC test, their latest TOEIC test, and to answer five screening questions. The screening questions established whether the participants had any underlying conditions that might affect the results of the study. They were asked about any heart conditions, history of anxiety or depression, current anxious or depressive conditions, chronic pain, and a final open question about any other reason their stress levels might be unusual. Participants were also requested to fill out a table with their normal class schedule and note their self-assessed general stress levels each day on a 10-point scale along with a note in English or Japanese if anything unusual happened that day that might bias their stress readings for that

day. The remaining activities were organised by week, with participants asked to complete tasks at some point during the week when it was convenient for them.

During Week 1 the participants were asked to complete the first GHQ-28 assessment and the FLCAS. The *Garmin Vivosmart 3/4* was not to be worn during this week.

During Week 2 the participants were asked to check the charge on the *Garmin Vivosmart 3/4* and to charge the device if necessary. They were also requested to download the free *Garmin Connect* application to their smartphone and create their user account. Finally, they were requested to wear the *Garmin Vivosmart 3/4* for at least one night to ensure proper calibration. If there were any problems with the *Garmin Vivosmart 3/4* they received, then participants were invited to return the device to the researcher and receive a replacement device. None of the data gathered during Week 2 were included in the research data as this was a period to ensure proper calibration of the device.

During Week 3 participants were asked to wear the *Garmin Vivosmart 3/4* as much as possible. It was clarified in the schedule that the device could be worn while swimming, and as such did not need to be removed for most activities. Participants were reminded to please synchronise the device with their smartphone at least once a week, and to charge the device when it gave a warning that it had less than 24 hours of charge left.

During Week 4 participants were asked to continue wearing the *Garmin Vivosmart 3/4* as much as possible, and to remember to synchronise and charge the device at least once a week. Participants were also requested to complete the GHQ-28 for a second time during this week.

During Week 5 participants were requested to synchronise the *Garmin Vivosmart 3/4* for a final time with the application on their smartphone, and then to stop wearing it.

During Week 6 participants were requested to continue not wearing the *Garmin Vivosmart 3/4* and to complete the GHQ-28 for the third time.

The end of Week 6 marked the end of the study and participants were requested to review their participant diary and ensure that they had completed all the tasks, and to write down their username and password for the *Garmin Connect* website so that the researcher could retrieve the data. Both the participant diary and *Garmin Vivosmart 3/4* were then to be deposited into a sealed box outside of the researcher's office door.

4.5. Participant Recruitment

Participants were recruited from the second-year students in the Department of International Studies on the Siebold Campus of the Nagasaki Prefectural University in Nagasaki, Japan.

First-year students were not invited to participate as approximately 20% of students (depending on the year) are from other prefectures, and as such are adjusting to living alone for what may be the first time in their lives. Heightened levels of anxiety and stress would be expected during this period and might contaminate the study's results. Third year and fourth-year students were not invited to participate in the study as job-hunting activities start mid-way through the third-year, and these may likewise introduce additional background stress and anxiety that might make it difficult to distinguish between sources of distress.

Participation in the study was introduced in a brief three-minute summary of the research's goals and methodology at the end of an English lecture. Interested parties were invited to attend a more in-depth 15-minute briefing during lunch later during the week in one of the lecture venues.

At the longer briefing an oral outline of the research procedure, method, and objectives, was presented in English and Japanese, supported by written materials in Japanese detailing the above.

Interested parties attending the presentation were given copies of the participant schedule and diary, and additional briefing materials (included in Appendix B). No register was taken at the briefing session, and it was made clear that there was no commitment to participate in the research at this point. The participant schedule and diary contained the research schedule, stating what had to be done and on what days, and participants could clearly see what information was being requested in this study. The additional briefing materials outlined the functions of the Garmin Vivosmart 3, what data it gathered, the measures taken to safeguard data, and how the data would be handled after the study's completion. The briefing materials also outlined the information to be gathered by the FLCAS and the GHQ-28, however no instructions were included on how to score or interpret the FLCAS or GHQ-28. Additional information on the risks involved in the research was included, such as the possibility of a latex allergy or an adverse stress reaction, and instructions on what steps to take in the eventuality of these issues arising. The written briefing materials and the script for the verbal presentation were checked by a Japanese mother-tongue speaker for accuracy.

On the last page of the research diary was a removable copy of the consent form, for participants to sign only after they had time to read and consider all the material presented and had decided to participate in the study. It was felt that because of the potential for miscommunication, the dual format of both a presentation and a written document was best, and that participants needed to be given time to read and assimilate the documentation. A further complication in this process was that at the time of the study the age of contractual

capacity in Japan was 20 years of age. An unknown number of the potential participants in the study were below that age, but the departmental statistics suggested that approximately 50% of the second-year students were 19. The ethics committee was satisfied with the standard threshold of 18 years of age to consent to participate in the study. However, the researcher included space for parents' signatures if participants felt uncomfortable signing for themselves and allowed two weeks before the start of the study for participants to mail documents to their parents if they lived away from their home prefecture. For reference, the Japanese postal service's standard delivery time for domestic mail is two days, so this was ample time for the participants to send the documentation to their parents for consideration should they so wish. Two participants, both under the age of 20, did include their parents' signatures on the form in addition to their own.

A *Garmin Vivosmart 3/4* could be obtained from the researcher by submitting a signed consent form. Participants handed the signed consent form to the researcher, who checked it was complete and was available to answer any additional questions the participant might have. Then the participant randomly selected a *Garmin Vivosmart 3/4* from a box so that there was no way that the researcher could link the participant with a particular *Garmin Vivosmart 3/4* or the data gathered by that device. This submission of the consent form and receiving a *Garmin Vivosmart 3/4* completed the recruitment phase.

Also included in the participant schedule and diary was a copy of the form for withdrawing from the research. It was made clear that participants could withdraw from the research at any time without stating any reason by simply signing the form and depositing the diary and the *Garmin Vivosmart 3/4* together into the box outside of the researcher's office. The reason for requiring the diary and *Garmin Vivosmart 3/4* together was so that the researcher could then go online and manually delete any data the participant had uploaded to the servers immediately on receipt of the notice of withdrawal. The *Garmin Vivosmart 3/4s* memory would also be reset to factory settings and the participant's diary shredded. Over the course of the study only one participant withdrew for unspecified reasons, and the procedure outline above was followed.

4.6. Participants

All participants were recruited from the second-year student population in the Department of International Studies (国際社会学部) on the Siebold Campus of the Nagasaki Prefectural University (長崎県立大学). The Siebold Campus is a smaller satellite campus of the main Nagasaki Prefectural University, established as a result of a merger between the Siebold University of Nagasaki and the Nagasaki Prefectural University in 2008. The Department of International Studies has approximately 60 students per year.

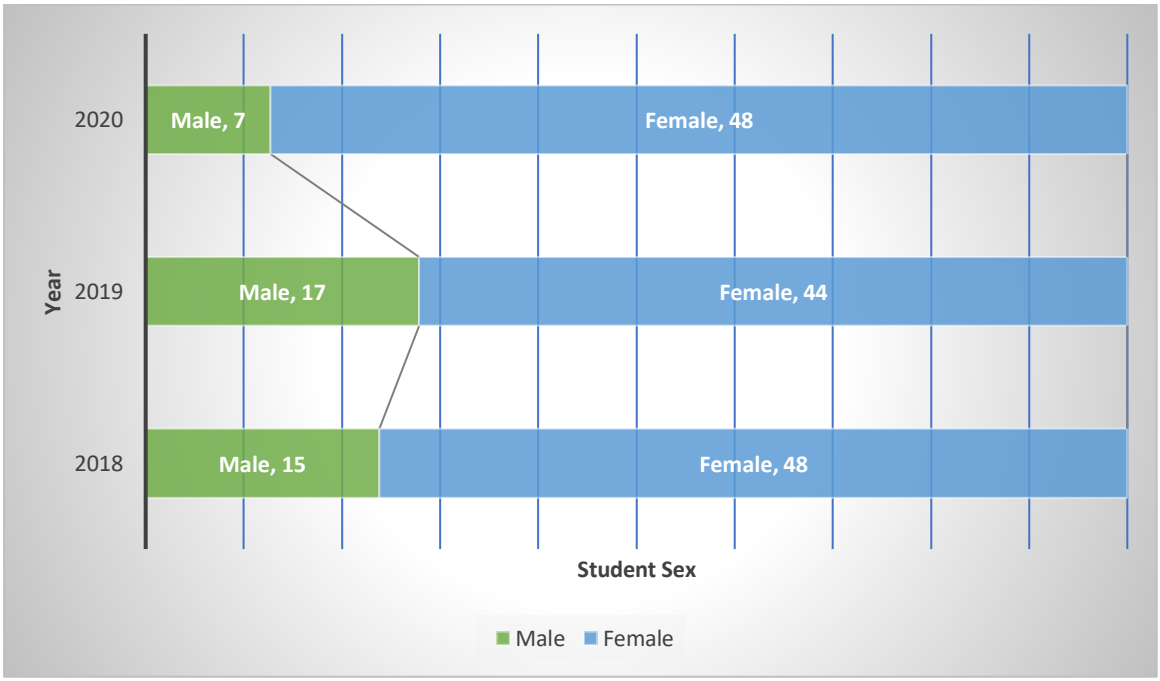
This student population was selected because they are required to take English courses every year as part of their degree requirements, whereas other departments such as Information Technology and Nursing are only required to take it during the first year. The Department of International Studies students are also required to take the TOEIC test every year, with a minimum score of 750 being required for graduation.

All of the participants were aged either 19 (24 of 40, or 60%) or 20 (18 of 40, or 40%), and spoke Japanese as their mother-tongue. The primary variable in participants was sex.

The Department of International Studies students were predominantly female, with the percentage of female students ranging being 83% in 2020, 72% in 2019, and 76% in 2018. The sex demographics of the second-year student population for the years during which this study took place are displayed in Figure 9 below.

Figure 9

Sex Demographics of the 2nd Year Student Population in the Department of International Studies at the Siebold Campus of the Nagasaki Prefectural University



The imbalance in the sex of the population being drawn from does pose some challenges to the generalisability of the findings of this study. This was unavoidable as only this student group met all the criteria for the study.

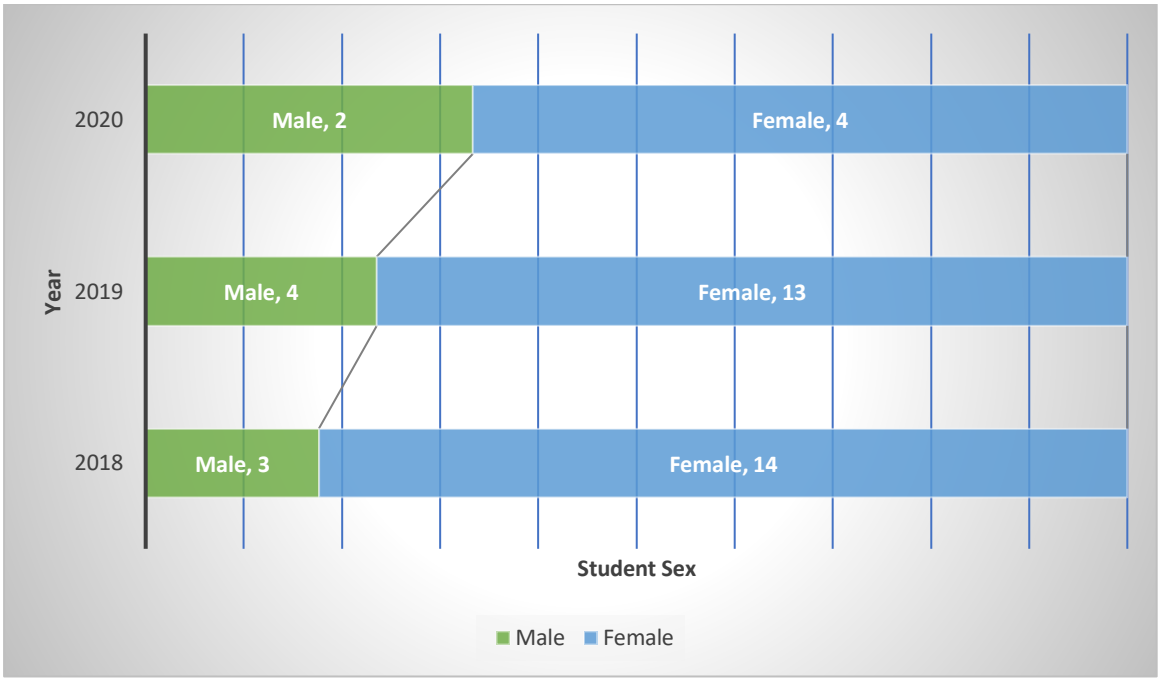
Had there been more interested participants than there were *Garmin Vivosmart 3/4*s, then the excess participants would have been included in a second pool of participants in the

same year, however there were less than 20 interested participants each round of recruitment, and it was deemed that additional rounds of recruitment could have been viewed as pressure from the English teachers to participate, therefore there was only one group of participants from each year.

The sex demographics of the participants will be detailed in Figure 10.

Figure 10

Sex Demographics of the Research Participants



Across the course of the study a total of 40 participants volunteered to participate in the study. The lower number of participants in 2020 was a result of the Covid-19 pandemic, and while the number of cases was still relatively low in Japan, the numbers were rising rapidly, and students were nervous. Almost all of the data returned was incomplete, with 3 of the participants not having completed all the questionnaires in the manual, and only one student having returned partial wearable data from the Vivosmart 3/4. As a result of the unusual conditions the data from the 2020 participants was discarded as potentially tainted, reducing the number of participants to 34. Of the 34 remaining participants 32 returned completed research diaries with all of the questionnaires filled out. One participant failed to complete any of the GHQ-28 questionnaires in the 2018 group, and one participant in the 2019 group failed to complete one of the GHQ-28 questionnaires. The complete data that these individuals did provide were included in the study.

Males were under-represented in the participants, even relative to the smaller number of male students in the department as a whole. The relative increase in male

participants between 2018 and 2019 tracks with changes in the department's general demographics, and the relative increase in male participants in 2020 is simply a product of a lower number of volunteers to participate in the research. Interest in participation in the research was relatively stable between 2018 (14 participants, or 22% of students in the department) and 2019 (13 participants, or 21% of students in the department).

By far the biggest problem was with retrieving data from the *Garmin Connect* website. Participants were requested to choose a unique username and password for their account on the Garmin website. Providing participants with pre-generated usernames and passwords was considered, and this offered certain advantages, such as allowing the researcher to download data on a daily basis, decreasing the administrative burden at the end of the study. However, pre-generated passwords raised concerns about security since the diaries could easily be read by someone else, especially during the briefing sessions when they were simply sitting on tables, and any password that is written down is insecure. Further, there were ethical concerns about accessing student data before the end of the study when students might decide to withdraw at any point before the final submission of their diary and the wearable device. It was decided to prioritise student confidentiality by asking students to write their username and password only at the end of the study before submitting their participant diaries.

Of the 34 participants only 14 participants submitted readable passwords or usernames. In addition, of those 14 participants whose data was accessible, three participants' devices failed to synchronise properly, providing no data, and three participants submitted partial data that was missing two or more days of data over the two weeks when the *Garmin Vivosmart 3/4* was to be worn. In total, including the three participants with partial data, 11 data sets were retrievable for analysis. The data problems are summarised in Figure 11 below.

Figure 11

Garmin Vivosmart 3/4 Data Issues

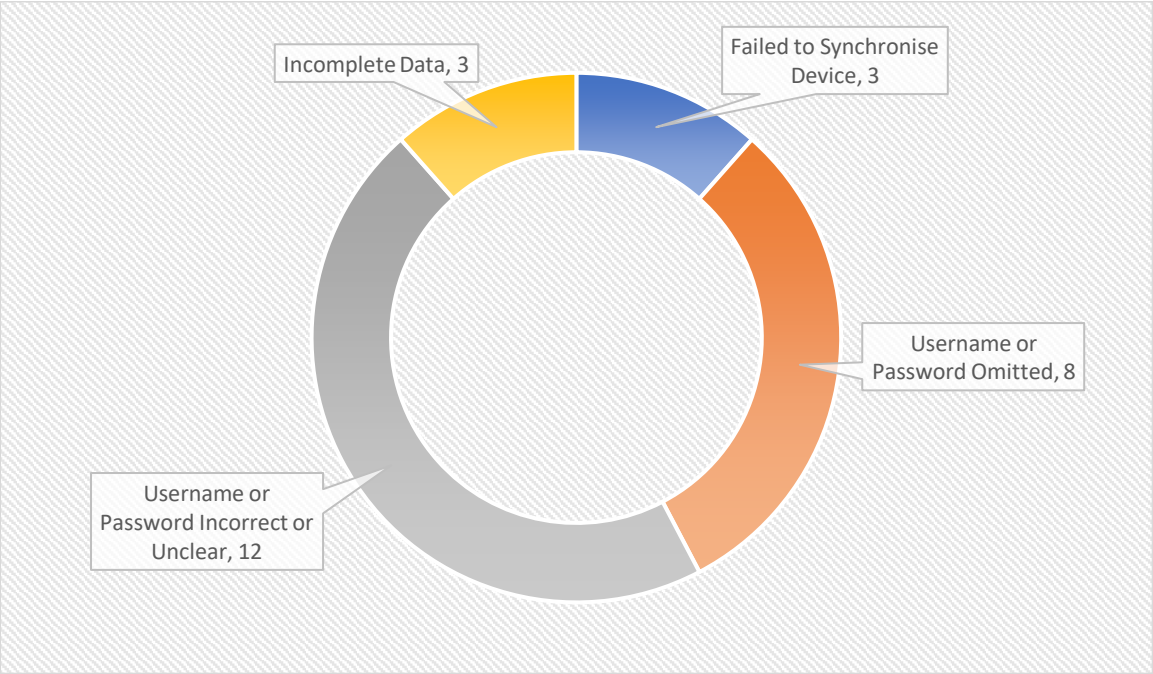


Figure 11 shows that the biggest problems were with incorrect or unclearly written usernames or passwords (12 out of 34), or where usernames or passwords were omitted (8 out of 34). These two factors resulted in the loss the data from the Garmin Vivosmart 3/4 for 59% (20 out of 34) of the research participants, with synchronisation issues accounting for a further 9% (3 out of 34). In total more than two-thirds of the data gathered in this research was inaccessible.

Unfortunately, it was not possible to retrieve any of the missing usernames and passwords without violating the assessor single-blind anonymity of the participants guaranteed in research design. While the participants could be emailed or contacted using the information included on the consent forms, the moment they emailed back a corrected username or password it would link their research data back to an individual. Other reasons existed for not attempting to obtain the missing username and password, which will be explored in the following limitations section.

As Covid-19 prevented the gathering of any further data it was determined that proceeding with the analysis of such data as was available was the best course of action, and the results will be presented in Chapters 6 and 7.

4.7. Limitations

This study had numerous complications that introduced limitations that affected the generalisability of results because of lower than desired sample sizes.

4.7.1. COVID-19

Unfortunately, further research cycles planned for 2020 and 2021 were impossible to conduct because of the impact of COVID-19 and the shift from face-to-face classes to online classes, which introduced many confounding factors. The first of these was the lack of access to research participants. At various stages in 2020 and 2021 when the number of COVID-19 infections spiked the campus entered lockdown, with all clubs, social events, and face-to-face classes being cancelled or moved online. This move online was a major issue given that the construct being study was premised on traditional face-to-face language learning classrooms, and stress data gathered from online learning would be fundamentally incompatible with both previous research and data gathered prior to this point. In addition, many of the classes were supplied in an “on demand” format, meaning that pre-recorded lectures could be accessed at any time, which made any attempt to link stress data to actual learning activity according to a standard schedule an invalid assumption. Finally, given the centrality of the concept of stress to this study and the significant life changes experienced by everyone during the COVID-19 epidemic, it would raise reasonable questions about the validity and reliability of this study if stress data gathered during the COVID-19 epidemic was included in this study.

For these reasons gathering further data was not practical at this time, nor was it clear when normality would be sufficiently restored to allow further data gathering. Given the time limits placed on the completion of this study this necessitated proceeding with such data as had already been obtained.

4.7.2. Participant Population Size and Valid Response Rate

For this study, the fact that the students in the Department of International Studies completed a TOEIC test every year as part of their degree requirements made them an ideal sample group. The pre-existing longitudinal standardised and normed scores on language performance were one of the core measures around which the study was built. This is the first study of those accessed during this time period to use a standardised and defensible measure of language proficiency and language learning.

The small size of this department, with approximately 60 students a year, imposed limitations on sample size per research cycle. The response to calls for volunteers was very positive, with 17 out of 63 students volunteering to participate in 2018 (27%), and 17 out of

61 students volunteering to participate in 2019 (28%). While only 6 out of 54 students volunteered to participate in 2020 (11%), this was probably attributable to the COVID-19 pandemic. The 27% and 28% participation rates were sufficiently close to the researcher's estimates of a 30% participation rate that there seemed to be no reason for concern.

While the problems with usernames and passwords were frustrating, the high quality of the data and the detailed analysis possible on such rich and dense data made the low numbers less of a concern. Medical studies that produce the same sort of data through repeated measures, as in the case of wearable devices repeatedly measuring stress, often have very small sample sizes, but are viewed as valid and reliable (Crowder & Hand, 2017). Bakdash and Marusich (2017) state that repeated measures correlation, "tends to have much greater statistical power" (p. 456).

The initial plan was to conduct repeated rounds of research to gather additional data, however as mentioned above, COVID-19 disrupted these plans. While it is acknowledged that the number of valid responses was far less than hoped for it is nonetheless sufficient for analysis, although generalisability is possibly limited. The possible reasons for the low number of valid responses will be explored below.

4.7.3. Orthography and Cultural Factors

Before attributing the problems with usernames and passwords to cultural factors, it was considered whether some of these cases might be attributable to orthographic differences. English mother-tongue speakers are familiar with the difficulties in distinguishing 0 (zero) from the letter O. Even in typed correspondence the font may make the two difficult to tell apart. In handwritten correspondence, such as the handwritten passwords and usernames in the participant diaries in this study, the difference can be even more difficult to discern. Sometimes it is not just a choice of two similarly shaped symbols. For example, the letter l, the number 1, the exclamation mark (!), and a capital i (I) can all look the same if written in haste or by someone who is unused to writing English. When confronted with a username like OIIVE there are 32 possible combinations, and while Olive seems the most likely to an English native speaker it might equally be 0 L 1 \ / E. Passwords, where numbers and symbols are commonly substituted for similarly shaped letters added to the complexity. This is not to imply that the researcher did not attempt to substitute commonly confused symbols, but that this is one possible explanation that was contemplated.

However irregular or unclear orthography when writing the usernames and passwords would account for, at most, 12 of the inaccessible records, while in a further 8 nothing at all was written. Again, this may be simple forgetfulness and not following the directions in the diary. While this research was of tremendous importance to the researcher

it was a volunteer activity for the participants, who generously gave their time to assist the researcher. It is important to keep a sense of perspective and acknowledge that participants are giving of their valuable time to assist in the research, and that the research period of 6 weeks was long and required a great deal of work from participants, who may have been bored with the research as the novel portion of the research, wearing the *Garmin Vivosmart 3/4*, was two weeks behind them. This also raises the possibility of participants simply having forgotten a password and username that they had not used in two weeks.

The aforementioned possibilities are plausible explanations; however the large proportion of inaccessible participant records, with 23 our of 34 (67.6%) of participants wearable data being inaccessible, suggests that there may be a more pervasive factor. It is important to consider the possibility that the researcher may have not fully considered the role of culture in the study.

Discussions around mental health issues have long been taboo in Japan, and while there has been some relaxation of this taboo in the last decade it is still a sensitive topic for most people in Japan, which makes research into this topic difficult (Borovoy, 2008; Ando, et al., 2013; Lieber, 2017).

It is for this reason that the researcher suspects that at least some the numerous problems with the non-return of smartwatches, incomplete tests in the research manuals, and incorrect passwords and usernames may be attributable, at least partially, to cultural sensitivities in Japan around the issue of mental health. There is still a perception that those who struggle with mental health issues are weak or defective in some manner. This is a contributing factor in individuals failing to seek assistance, which is a part of the reason for Japan's relatively high suicide rate among industrialised countries (Chen, et al., 2009). It may also have interfered with participants' willingness to disclose information that might indicate mental health issues to the researcher (Borovoy, 2008; Ando, et al., 2013).

While this approach of submitting missing, erroneous, incomplete, or unclear passwords and usernames may seem like a round-about way of addressing this issue, it is a very Japanese solution. It allows the participant to "save face" (面目が立つ) and avoid conflict by not directly refusing the researcher's request for the data, and instead maintain the plausible excuse of an error or forgetfulness, while still maintain the personal boundary that they have set in not disclosing the data. This may seem an oddly round-about solution to the issue to those unfamiliar with Japanese culture and language.

When the research briefing material and diary were initially reviewed by a Japanese mother tongue speaker, there was a conversation around the issue of mental health where

the mother tongue speaker suggested that possibly some euphemisms could be used for words such as stress and anxiety. However, these terms were ambiguous, for example the suggested terms for anxiety could also be read as shyness, concerns about something, or exhaustion. There were concerns that the use of more polite oblique references might obscure the research's goals. Even in the interests of politeness and cultural sensitivity this might well cross ethical boundaries in terms of securing informed consent from participants, and so the original word choices were retained, although this may have affected participants' willingness to disclose data on a subject that is culturally sensitive.

As an aside, it is worth stating at this point that the researcher respects the right of the participants to withhold any data that they do not wish to share, and indeed to withdraw their permission for the researcher to use their data at any time. This section should not be read to imply that the researcher had any sense of entitlement to the participants' data.

Returning to the issue at hand, one problem with the hypothesis that the return rate was lower because it touched on mental health issues is that the return rate for participant manuals with the GHQ-28 was much higher. It may be that because the GHQ-28 was unscored participants viewed this data as less threatening, while the *Garmin Vivosmart 3/4s* resultant stress data was displayed on the screens of their smartphones. The difference may lie in the issue of reporting results to participants. This raises ethical questions around the use of wearable devices in general when delivering psychometric data to individuals, whether in research or in daily life, and how that data is understood and interpreted. This leads to the next chapter, in which ethical issues raised by this research will be presented and discussed.

Chapter 5: Ethical Issues

In this chapter the ethical issues surrounding this research will be discussed. This section is traditionally subsumed into the general research methodology section. However, it was felt that the rapid growth in the use of wearable devices in research, and in society in general, is an area that has not received sufficient attention. This chapter will discuss some of the potential ethical issues, and present and discuss some of the data relating to ethical safeguards used in this study.

5.1. Ethical Oversight and Clearance

This study was subject to ethical oversight from two institutions and required ethical clearance from both institutions. The researcher's doctoral studies were under the purview of Rhodes University in South Africa, however the location of the study was at the University of Nagasaki, Siebold Campus, in Japan. As such this study was subject to ethical oversight from both institutions, and clearance had to be obtained from both ethics committees. In some cases, there were differences of opinion regarding ethical conduct the stricter standard was always adopted. Copies of the ethical clearance documents can be found in Appendix D. The Japanese document has been annotated with an English translation. Should there be any enquiries regarding the ethical clearance of this study the relevant tracking numbers are PSY2017/17 for Rhodes University, and application number 368 and judgement number 355 for the University of Nagasaki, Siebold Campus.

5.2. Data Management and Participant Confidentiality

Concerns about privacy are not unique to debates around wearable devices, however wearable devices are of special concern because of the quantity and quality of the data that they gather. The term "lifelogging" (Anaya, et al., 2017) has been used in relation to wearable devices because many are designed to be worn constantly, with only brief gaps in the recording when the device must be charged. Even these gaps may disappear in the near future, with clothing that generates electricity from movement and allows devices to be worn without interruption (Tian, et al., 2021). Devices such as the *Garmin Vivosmart 3* take measurements regularly, monitoring heart rate in real time, and taking HRV measurements every 3 minutes, in addition to information about stride, number of steps, elevation changes to count stairs climbed, and a simply staggering amount of data. A single day of data from participants generated an Excel spreadsheet of over 10,000 rows of data.

Further, this data is qualitatively different from the type of information that could be gathered by simple observation. Much of the data could be considered lifestyle data, such as number of steps taken, stairs climbed, or hour spent sitting, and is the sort of thing that a

person might count themselves, if they want to. However, many of the latest evolutions in the type of sensors wearable devices are stepping into measurements traditionally taken in a doctor's office, such as heart rate, and blood oxygenation levels (Foster & Torous, 2019). This type of data is qualitatively different, and the implications of this will be discussed in more detail later in this chapter. In this section the primary concern is with the privacy implications of these wearable devices, collecting more and more sensitive data that might be characterised as medical data and placing it in the hands of third parties on the internet, who may not exercise due caution with how they store the data or may outright exploit the data for their own gain (Weston, 2015).

A survey of 60 wearable device users in Sydney, Australia by Anaya, et al. (2017) suggests that privacy concerns are the single biggest area of concern amongst wearable device users, with informed consent regarding how the data is being used being the second most pressing concern.

Cognisant of these issues, the research design of this study made the protection of participants' privacy one of the primary concerns. It was for this reason that the assessor single-blind research methodology was employed. If the researcher, holding all the information provided by participants, was unable to identify the participants then it would be extremely unlikely that a third party would be able to deduce the participants' identities. Information was deliberately split between digital (the *Garmin Connect* application and the website where the information was stored), and physical (the *Garmin Vivosmart 3/4* wearable devices, participant diaries, and consent forms).

Digital privacy was protected through the following steps.

First, the researcher approached *Garmin* and obtained assurances that research participants' data would be treated in the same manner as that of minors and other protected individuals. This meant that the data would be excluded from even anonymised aggregate date collect by *Garmin,* and that Garmin would sequester the data on special servers with more restrictive access controls. Unfortunately, this first step was not entirely successful as it was premised on the researcher being able to send a list of usernames to *Garmin* that *Garmin* was able to then sequester. Where participants did not provide a username or provided an unclear, inaccurate, or otherwise invalid username this step was impossible. Participants were informed that this was one of the steps in protecting their data, and the researcher tried their upmost to ensure that it was enacted, but it did require willing cooperation from the participants.

Second, participants were informed that at any point they could delete their data from the *Garmin Connect* application and website, and *Garmin* assured the researcher in writing

that all data would be permanently and irrevocably deleted. It is hoped that participants who chose not to avail themselves of the first protection used this second level of protection on their own. Copies of the relevant email correspondence between *Garmin* and the researcher are included in Appendix C.

Third, participants were instructed to use a unique username on the *Garmin Connect* application and website, so that their information could not be linked to other accounts on social media or other platforms. They were also instructed not to use their personal names, photographs, or other identifying information anywhere on the *Garmin Connect* website. Thus, even if participants chose not to avail themselves of the first and second levels of protection, then at worst all that remained to link their data to them was their age, gender, and preferred language setting. It was deemed that knowing that the participant was a 19-year-old male who spoke Japanese was insufficient information for someone who obtained access to their account to link the information back to an individual.

Physical records, namely the *Garmin Vivosmart 3/4* wearable device, consent form, and participant diary, were disassociated from each other. The participant diary and consent form were received by potential participants at the briefing, but the consent form was then submitted separately to the researcher in exchange for a randomly selected *Garmin Vivosmart 3/4*. The consent forms were then filed separately, and at the end of the research the participant diaries and the *Garmin Vivosmart 3/4* were placed in a box outside the researcher's office.

The consent forms were the only piece of documentation containing the participant's real name and contact information. The disassociation of the consent form from all other research instruments and keeping it as a physical document that could not be digitally accessed ensure that there was no link between the participant's name and any other research instrument. This meant that while the researcher could contact all participants in general with reminders about deadlines and other matters, there was no way that even the researcher could identify any single participant in the study. In addition, the *Garmin Vivosmart 3/4* was reset to factory settings after each round of research, completely wiping all data and settings from the device.

The sole point of contiguity between the physical participant diary and the information on the *Garmin Connect* website was in the username and password contained in the participant diary. Without the username and password there was no way to link the participant diaries and the *Garmin Vivosmart 3/4* data. Similarly, there was no way to know whose password or username had been omitted or was otherwise unreadable. It might have been possibly to narrow the list of possible individuals by a process of elimination using age

and gender had there been just a couple of missing usernames or passwords. However even in this case the researcher would have been unable to conclusively identify missing records.

This may seem like an error or oversight, however from the outset the intention was that the research design would be assessor single-blind, with the participants knowing the researcher's identity, but with the researcher unable to link any information back to participants. The type of highly personal data being gathered in this study argued strongly for this approach. Even with this assessor single-blind approach, the low return rate of data from the *Garmin Vivosmart 3/4s* suggests that there may still have been a reluctance to disclose data to the researcher.

The low return rate of wearable data, particularly when contrasted with the fact that almost all the physical participant diaries were returned, suggests that even more efforts should be taken to ensure that participant confidentiality and anonymity is protected. Confidentiality and anonymity, even from the researcher, seem to be of special concern in research using wearable devices. It is unclear what further possible steps could have been taken to protect participants' personal information, and it may be that this is not a question of what further steps the researcher can take, but rather that a question of participants' concerns and perceptions of the risk of a breach of their privacy.

It may be that this issue is unavoidable, and that it is something that researchers will simply have to account for in research design when contemplating any research using wearable devices. It may be reasonable to assume that approximately two-thirds of research participants will choose not to disclose. Strath and Rowley's (2018) "mini-review" (p. 54) of 11 studies using wearables suggests that a response rate of a third is typical, although Strath and Rowley (2018) attribute this to the long duration, tending towards 6 months or more, of these studies and natural attrition. However, the low response rate seen in this shorter-term study suggests that it may not be related to the duration of the study, but rather that the underlying issue is concerns over privacy.

Strath and Rowley (2018) do cite one study that managed a higher response rate, the "*Fitbit plus cash incentive group*" (p. 61). This raises the difficult ethical issue of whether cash incentives are an ethical option, particularly in light of potential concerns that this reluctance to disclose information may be related to participants' concerns over personal privacy.

5.3. Participant Effort versus Reward

Time is the most basic commodity that every human being possesses, and everyone is born with a finite and unknowable about of time. The gift of someone else's finite time is precious and should be respected. When someone goes to work, they are most often not paid for their output, but rather for their time in the workplace. Economists recognise this in concepts such as the value of leisure time. Hagberg and Lindholm (2010) discuss this in the context of considering how people make decisions on how much time to spend on exercise, leisure, and work, showing that people are aware of these concepts and can even place a cash value on their time. There is an increasing awareness of the value of people's time, especially in such areas as unpaid internships, and the ethical issues around unpaid work in general, and especially in how they perpetuate discrimination against certain groups (Grant-Smith & McDonald, 2018).

Another way of looking at this issue is from the perspective of data. There is also an increasing awareness of the value of an individual's personal data. Every time one uses the internet, even seemingly free services like *YouTube, Facebook,* or typing a search query into *Google*, these services gather data about the individual that they sell to advertisers and other third parties (Reiderer, et al., 2011). Data has value, and research is all about data, whether it is quantitative data like test scores, or qualitative data like interviews, the researcher is still collecting participants' personal data. That data has a value is best demonstrated by the proliferation of digital goods and services purporting to be free yet generate huge amounts of money from the data that their users provide. It has been argued by researchers such as Li, et al. (2019) that users are paying for the use of these so-called "free" services with their valuable personal data. Li, et al. (2019) state that *Facebook* generated a consumer surplus of US$21.4 billion in the USA alone in 2017. The issue of people's data is tremendously ethically sensitive, as the *Facebook-Cambridge Analytics data scandal* demonstrates. In brief, *Facebook* user data was used by *Cambridge Analytics* to build psychological profiles that were used to provide assistance to the 2016 presidential campaigns of Donald Trump and Ted Cruz (Li, et al., 2019).

These seemingly disparate debates have profound implications for research ethics. Participants' time spent on research can be thought of as an economic issue of the cost of the time spent on the research as opposed to working, or as a social issue relating to the fairness and ethics of unpaid work, or a question of the value of data in an increasingly data-driven economy. Regardless of the perspective, the conclusion is the same, which is that the consensus across multiple fields seems to be converging on the conclusion that researchers need to adequately compensate participants for the value of what they are contributing.

Ripley, et al. (2010) conducted a survey of 1,600 researchers and 1,900 institutional review board (IRB) chairpersons in the USA. Table 3 of Ripley, et al.'s (2010) study shows that 69.8% of researchers, and 64.1% of IRB chairpersons consider compensation, above and beyond reimbursement for expenses, to be necessary. Similarly, in Cheff's (2018) survey on research compensation practices in Toronto, Canada 65% of the researchers provided compensation to participants above and beyond recompense for expenses incurred. Cheff (2018) quotes a response from a researcher that sums up the issue most elegantly, "If we value the information the compensation should reflect that. If we don't value the information then why are we collecting it" (p.6). The quotation from Cheff (2018) demonstrates this growing awareness that information (data) has value.

In this study 21 of the 34 participants (61.7%) reported having a part time job. This suggests that the participants were aware of the value of the time that they were volunteering in participating in this study as they were already balancing the demands of university work, leisure time, and their part-time job. This generation of individuals is also probably acutely aware of the value of their personal data on the internet, but may not have considered it in relation to research.

Morse (2005) discusses the issues around this issue of unpaid work in research in their editorial, touching on issues such as intangible rewards, the difficulties of determining adequate compensation, and social issues around how labour is valued. Morse (2005) concludes that this is not so much an issue of money, as an issue of showing respect, and refers to the concept of an *honorarium* in the sense of showing respect and honour to the participants. However, the word *honorarium* has a somewhat fraught history, originally referring to a bribe paid to get appointed to an honorary post (Klein, 1971).

This raises the issue of what amount of *honorarium* might constitute too tempting an offer for participants to refuse, and thus cross the line into a form of duress to participate in research, and whether one should, for example, compensate based on participants' current pay per hour. This seems harmless and logical until one considers that it would in effect perpetuate issues like the gender pay gap. However, the existence of this gender pay gap means that if the honorarium paid was the same for all groups it might be higher than other work paid for female participants, and this crosses the line into an unduly enticing incentive.

Cash payments to research participants are an ethically fraught area where existing social inequalities make determining a fair and ethical payment, or *honorarium*, extremely difficult. Yet clearly some balancing factor is required, or the researcher is taking something that clearly has value, whether it is time away from work, the individual's unpaid labour, or the value of the participants' data, without giving anything in return. This inequality needs to

be addressed by the researcher. It was a question that was pondered in this research. How to give the participants something that all of them would benefit from equally as a means of compensating them for their time and restoring balance and fairness to the relationship?

In this study the wearable devices were the key means of addressing this inequality. On the most basic level there was the novelty and entertainment value of getting to play with a new technological toy. The researcher observed that none of the participants in this research group owned a wearable device like the *Garmin Vivosmart 3/4,* and as such wearing one was a novel experience. At the time when data collection for this research started in 2018 wearable devices were less prevalent, and the research observed none of the students wore wearable devices at that time. In the past year a few students with wearable devices have been observed, but the prevalence of wearable devices still seems to be low. While this may seem a frivolous form of compensation, research suggests a link between novelty and dopamine (Li, et al., 2003). It may be helpful to think of the hours of enjoyment playing with a new device as equivalent to the cost of going to see a new movie. No cash has exchanged hands, but entertainment has a value in terms of the economic theory of the value of leisure time, and all participants were equally compensated.

Another factor that was planned to be form of compensation, was that the literature at that time suggested that these wearable devices conferred health benefits in that they encouraged activity and other health-enhancing behaviours (Bravata, et al., 2007; Klasnja & Pratt, 2012; Cadmus-Bertram, et al. 2015; Jauho, et al., 2015; Poirier, et al., 2016). The gift of improved health is again something that all participants could all benefit from as a means of balancing the scales. People pay large sums for gym memberships in the hopes of improving their health, so the logic went that if these wearable devices helped participants become more aware of adverse behaviours it might, at least partially, compensate them for the value of their time and data.

The *Garmin Vivosmart 3/4s also* contained a new function, the stress measurement system. The system seemed conceptually sound and beneficial to participants. An increased awareness of what situations were resulting in increased stress might potentially help participants to better manage their stress. The wearable device also had an option that could be set so that if stress levels rose above a number set by the participants, then the *Garmin Vivosmart 3/4* would vibrate. The device would then display a message suggesting that the participant engage in a few minutes of guided consciously controlled breathing. Research suggests that this can help with the reduction of stress, and consciously controlled breathing is a common meditation technique (Sasaki & Maruyama, 2014). This seemed like it would be

beneficial to participants, and again form part of the compensation given in exchange for participation.

What was not anticipated was the possibility that seeing their resultant stress levels may have made participants self-conscious about their mental health, which as discussed earlier is something of a taboo in Japanese society. While it is unknown if this was the case the unexpectedly high number of wearable devices that were not returned, and the problems encountered with usernames and passwords strongly suggests some reluctance to share the wearable device data with the researcher. This raises the next issue to be discussed, the use of wearable devices to measure phenomena like stress.

5.4. Wearable Devices and Stress Measurement

As discussed earlier, psychometric tests are sensitive devices that need to be properly administered, scored and interpreted (Furr, 2017). Early wearable devices such as the *Apple Watch* and *Fitbit* were, in the words of Foster and Torous (2019), “a glorified pedometer and electronic watch” (p. 22), but since then have rapidly expanded into measuring heart rate, blood oxygen levels, and other physiological data.

However, in this blur of progress a line has been crossed from delivering purely physiological data to providing feedback on the wearer’s mental state, such as how stressed people are in the case of the *Garmin Vivosmart 3,* which was launched near the end of 2016.

The Garmin Vivosmart 3 manual has this to say about its stress measurement function:

> Your device analyzes your heart rate variability while you are inactive to determine your overall stress. Training, physical activity, sleep, nutrition, and general life stress all impact your stress level. The stress level range is from 0 to 100, where 0 to 25 is a resting state, 26 to 50 is low stress, 51 to 75 is medium stress, and 76 to 100 is a high stress state. Knowing your stress level can help you identify stressful moments throughout your day. For best results, you should wear the device while sleeping. (Garmin, 2016, p. 3)

While the stress measurement system is based on a physiological measure, HRV, the output delivered is labelled with descriptors such as low, medium, and high. As discussed in the literature review, the word stress generally carries quite negative connotations in common usage and is quite poorly defined. The manual makes no attempt to clarify that what is being measured is arousal, and that while high levels of arousal/stress may be detrimental if persistent it is natural and normal to experience at least some moments of high stress/arousal every day. Likewise, constantly low levels of arousal/stress are also not

necessarily a desirable state, as it suggests a lack of stimulation and in the Hebbian model (1955) a sub-optimal level of performance. The topic of stress and arousal, as covered in the literature review, is a tremendously nuanced area of study where consensus is only just emerging after decades of confusion and definitional uncertainty.

This raises ethical questions about whether manufacturers of wearable devices are qualified to deliver feedback about individuals' psychological states, an area previously the domain of psychometric testing. Andersen, et al.'s (2020) recent study with 27 participants who suffered from chronic heart disease and used a *Fitbit Alta HR* to get additional information on their heart rate stated the following,

The practical implications of patient knowledge generation from Fitbit suggest that patients should not be left alone with interpreting activity data as part of self-care (Andersen, et al., 2020, Principal Findings section, para. 5).

While the consequences of inappropriate action were far more serious for the chronically ill heart patients in Andersen, et al.'s (2020) study, the point being made about individuals misinterpreting activity data from wearables and reaching incorrect conclusions is valid. Looking again at the *Garmin* manual the manufacturer has done nothing to clarify the complexity of stress and arousal as a phenomenon. The brief presentation of the issue may lead consumers to think that their goal should be to achieve the lowest possible stress ratings, and to avoid activities that are associated with higher stress levels. While this is unlikely to be a case of life and death the misinterpretation of the data from wearable devices may lead to undesirable behaviours.

This raises the question of whose duty it is to clarify these issues for consumers. These wearable devices deliver information that has previously been only obtainable in a doctor's office or from an expert. In these environments there was an expert on hand to interpret the data and put in its proper theoretical context. While the manufacturers of these devices are careful to clarify that these are not medical devices they are still touted as devices for making changes to one's lifestyle. Yet looking at the example of stress measurement the manuals do not deliver sufficiently comprehensive explanations of the measures for consumers to properly interpret and act on the data being provided. It seems ethically questionable to advise someone to make lifestyle changes based on a data from a device, while not providing sufficient information for the individual to understand how they should interpret or act on that data

This area was unregulated until quite recently. In September 2017, the U.S. Food and Drug Administration (FDA) started the *Digital Health Software Precertification Program* aimed at assessing the software used in products that make medical claims, such as

smartphone applications and wearable devices (U. S. Food and Drug Administration, 2021). Some wearable manufacturers have joined this program, most notably *Apple* and *Fitbit*. Some manufacturers are electing not to declare their wearable devices as medical instruments and are so currently avoiding the additional regulations associated with medical instruments. Foster and Torous (2019) speculate that it is to avoid the additional regulations, which these companies may not be prepared to deal with.

Foster and Torous (2019) criticise the current regulatory framework, stating that the "FDA has expressed little interest in regulating low-risk fitness monitors that are promoted for general 'wellness.' In practice, this means that companies can make hyperbolic claims for the effectiveness of their devices for promoting wellness" (p. 23). This places researchers using these devices in an uncomfortable position ethically speaking. The guidance provided in the user manuals is clearly insufficient and may even border on misleading. This raises a range of ethical issues, such as whether informed consent has been secured when users are unclear what is actually being measured, or whether there may or may not be harm to the user, however slight.

In this research the concern over harm to users was addressed using the GHQ-28. The following section details the GHQ-28 results from participants in this research.

5.5. Monitoring Participants for Potential Harm

In this study the GHQ-28 was administered three times, once in the first week of the study before the Garmin Vivosmart 3/4 was used, once during middle of the study when the Garmin Vivosmart 3/4 was being used, and a final time during the final (6^{th}) week of the study when the device had not been used for at least a week.

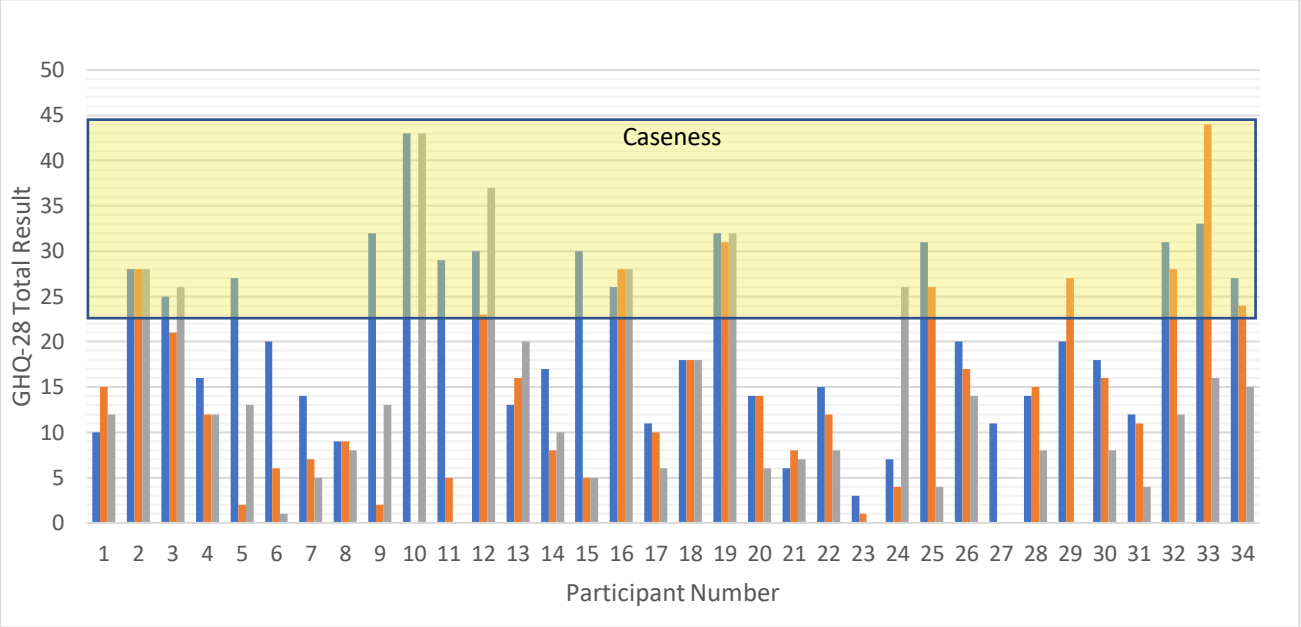
The first GHQ-28 administration was to be used as a baseline for normal pre-study functioning. Ideally participants' scores on the later administrations of the test would be the same or lower than the first administration, plus-minus ten percent because of the test-retest reliability of the device.

All 34 participants completed the first GHQ-28, with an average score of 20.35. Thirty-two of the 34 participants completed the second mid-study GHQ-28, with an average score of 15.4. Thirty-three of the 34 participants completed the final GHQ-28 assessment, with an average score of 13.48.

These averages seem to indicate that, on average, there were no adverse effects from the study, and that the study had a positive impact on participant general health. However, the researcher was concerned about the potential for any harm to even a single individual in the study. As such a more detailed analysis is presented in Figure 12 below.

Figure 12

General Health Questionaire-28 Results (Weeks 1, 4, and 6)



Note[1]. Blue lines indicate the first GHQ-28 administration in week 1, orange lines the second administration in week 4, and the grey lines the final administration in week 6. The yellow block at the top of the graph is aligned to show any scores of 24 or above, as they are above the caseness threshold for the GHQ-28, indicating levels of distress that may indicate a clinical condition, and are therefore of special concern.

Note[2]. On the final administration of the GHQ-28 participants 11, 23, and 29 scored zero. The two participants who did not complete all administrations of the GHQ-28 were participant 27, who omitted the second and third administrations of the GHQ-28, and participant 10, who omitted the second administration of the GHQ-28.

Before starting this discussion, it should be noted that the study was only one of many possible sources of changes in GHQ-28 scores. Therefore, three criteria were required for concern. Firstly, that a participant's GHQ-28 scores were significantly elevated (more than the ±10% expected variance from the GHQ-28's standard test-retest reliability) between the first GHQ-28 administration and the second administration in week four. The second administration in week four would be after the participant had worn the *Garmin Vivosmart 3/4* for one to two weeks, depending on when during week four they completed the GHQ-28. These criteria were intended to help the researcher isolate whether the likely source of distress was the *Garmin Vivosmart 3/4.* This would tend to suggest that the change occurred during the period when the device was being worn and feedback was being received. The

primary risk hypothesised to participants was that seeing their stress scores might aggravate any distress that was present.

The second criterion was that the GHQ-28 score should continue to be elevated above the week one measurement in the final week six measurement. This would tend to indicate that whatever happened to elevate the GHQ-28 score in week four was not transitory distress, but rather a persistent change that implied an ethical obligation on the researcher to investigate and ensure that no permanent harm had inadvertently been done.

If criteria one and two were met, then the participant's GHQ-28 scores would be more closely analysed looking at the scores on the sub-scales. The hypothesised possible source of issues was in the area of increased stress and anxiety as a result of participants seeing their stress scores and becoming more anxious about their stress levels. Therefore, the final criterion for concern was that the increase should be in the anxiety and insomnia scores on the GHQ-28.

It should be noted that the primary purpose of the GHQ-28 is as a screening device for clinical caseness, and as such the test has primarily been validated for assessing cases where the threshold for caseness is crossed. Nonetheless the GHQ-28 manual does indicate that it is sensitive to short-term changes and suitable for repeated applications as a means of monitoring changes in the same individual (Goldberg, 2013).

Participants 1, 13, and 21 met the criteria, and while none of their scores crossed the threshold for caseness, indicating clinical concern, their GHQ-28 scores were analysed in more detail to assess the possible reasons for distress, and will be detailed below.

Table 1

GHQ-28 Scores for Participant 1

	Total	Sub-Scales			
		Anxiety and Insomnia	Severe Depression	Somatic Symptoms	Social Dysfunction
Week 1	10	3	1	1	5
Week 4	15	3	4	4	4
Week 6	12	3	3	1	5

Examining Participant 1's GHQ-28 scores the anxiety and insomnia scores are stable. The increase in week four is because of an increase in somatic symptoms (from 1 to 4), and increased scores on the severe depression sub-scale (from 1 to 4). While the somatic symptoms had subsided by week 6, Participant 1 continued to score higher on the severe depression sub-scale (from 1 in week one to 3 in week 6). The change in scores seems to indicate that Participant 1 suffered some sort of injury (thus the increase in somatic symptoms), and the associated discomfort from the injury causing the individual to score higher on the severe depression subscale.

Table 2

GHQ-28 Scores for Participant 13

	Total	Sub-Scales			
		Anxiety and Insomnia	Severe Depression	Somatic Symptoms	Social Dysfunction
Week 1	13	7	0	5	1
Week 4	16	5	3	3	5
Week 6	20	5	4	6	5

Participant 13's scores on the GHQ-28 do not indicate a rise in the anxiety and insomnia sub-scale, which was the area of primary concern in this study. Rather the main increase seems to be in the area of social dysfunction, which increased from 1 in week 1 to 5 in week 4, and remained elevated in week 6, possibly indicating some sort of interpersonal conflict. There was a secondary increase in scores on the severe depression sub-scale from 0 in week one to 3 in week four, and 4 in week six. This may be explained by the social issues the individual was experiencing, or may be related to the ongoing somatic symptoms, which fluctuated from 5 in week one to 3 in week 4, then rose to 6 in week 6.

Table 3

GHQ-28 Scores for Participant 21

	Total	Sub-Scales			
		Anxiety and Insomnia	Severe Depression	Somatic Symptoms	Social Dysfunction
Week 1	6	5	0	0	1
Week 4	8	5	0	0	3
Week 6	7	5	0	0	2

Participant 21 showed an increase in their GHQ-28 scores during the period during which the Garmin Vivosmart 3/4 was being worn, week 4. Participant 21 showed no increase on the anxiety and insomnia sub-scale. The increase in GHQ-28 scales seems to be attributable to some sort of social problems, with Participant 21's scores increasing from 1 in week one to 3 in week 4 then decreasing slightly to 2 in week six.

The analysis of the GHQ-28 results for the three participants seems to indicate that something other than the wearable device was the most likely source of the increased GHQ-28 scores in this period. There seems to be no evidence of an anxiety-provoking stress feedback loop in this study, and the average scores seem to indicate that over the course of this study the participants on average experienced some improvement in their general health, although without a detailed analysis of each participant's scores it is not possible to assert whether this was attributable to the *Garmin Vivosmart 3/4s* or to other outside factors.

What the more detailed analysis of participants GHQ-28 above demonstrates is that while descriptive statistics such as mean, median and modal values are useful tools they can sometimes create incorrect impressions and be unhelpful in analysing complex phenomena. Therefore, in the following chapter four case studies have been selected for analysis and discussion to see what insights can be obtained from a more detailed engagement with the data. This will be followed by a more traditional statistical approach to see which if what the case studies suggest can be supported by statistical data.

Chapter 6: Case Studies

In this study a great deal of data was gathered from a variety of sources, and while analysing the data it became clear that the data lent itself to two different kinds of analysis. The traditional quantitative approach, focusing on statistics and correlations, does yield some important insights, but using solely a quantitative approach missed some important qualitative distinctions. The wearable devices produce such detailed data that it provides a narrative of the participant's life and activities. Analysis of the wearable data proved challenging because of the level of detail, however in this research the context of that data proved to be critical.

An example of the importance of contextualising data can be seen in the detailed stress data from the *Garmin Vivosmart 3/4*s when it is linked to activity type. This is of critical importance when assessing Research Question 1, whether there is a situation-specific stressor associated with language learning as hypothesised in the LLA literature (Horwitz, et al., 1986, Brown, 2000; Brown, et al., 2001; Kondo & Yang, 2004; MacIntyre, 2007; Tran, 2012; Gregersen, et al., 2014). If this is the case then the stress data should show either higher average levels of stress during language learning classes, or spikes into high stress suggesting moments of anxiety during language learning classes.

In making this assessment regarding whether stress levels are elevated, or normal context is critical, a baseline for comparison is required. In this study two baselines were contemplated, namely stress levels in other non-language classes, and stress levels during participants' free time. This might be accomplished by simply considering average stress levels. As will be shown in this section, the stress data gathered from wearable devices can be misleading if subjected to purely quantitative analysis because of the variability in the stress data.

Averages would also miss temporary spikes in the stress data, and while range might describe these spikes it would not give detail on when they occurred in the class and would limit analysis. Stress readings are taken automatically by the *Garmin Vivosmart 3/4* every three minutes, making it unlikely that periods of elevated stress would be missed unless they were extremely brief. Even then chance suggests that at least some should be captured. At the university where this study took place, class periods are 90 minutes long, allowing for up to 30 readings every class, and up to 480 readings a day. The stress data captured by the wearable devices from the rest of the day is important, in providing contextual information about whether any spikes seen in the participants' stress levels are indicative of a stressor unique to language learning environment, or are normal for that participants' stress patterns outside of the language learning environment.

In this chapter, four case studies have been selected for presentation as a means of exploring the data in a detailed and nuanced way. These data are in the form of numbers, traditionally the hallmark of quantitative studies. It is hoped that this section will demonstrate how the rich and detailed data gathered from wearable devices transcends the traditional categories of quantitative versus qualitative analysis, to allow an approach that may bridge the gap between these two positions.

6.1. Case Study Selection

The four case studies for analysis were selected on the basis of GHQ-28 scores and average stress levels, each case study representing a high or low point on these two scales. The reasons for using the stress scores have been discussed above in the introduction to this chapter and relate to Research Questions 1, 2, and 4.

The GHQ-28, discussed in a previous chapter, was used to assess whether clinical levels of distress may be implicated in LLA. As shown in Figure 12 of the previous chapter only five of the 34 participants (2, 10, 12, 16, and 19) in this study scored above the caseness threshold on all administrations of the GHQ-28. Participant 10 has been included in this number, despite having completed only the first and last administrations. Participant 10 scored well above the caseness threshold of 23, scoring 43 on both the first and last administrations of the GHQ-28, so it is a fair assumption that their mid-study GHQ-28 would likewise have been above the caseness threshold. Again, it should be restated that the GHQ-28 is a screening device used in clinical settings, such as hospitals, to identify individuals who may be experiencing clinical levels of distress. The GHQ-28 should not be confused with a clinical diagnosis and scoring above the caseness threshold merely indicates that clinical levels of distress are likely. Repeated administrations were used to establish that this distress was persistent and not transitory.

There was also an element of necessity in including participants 10 and 12. Only five participants who scored above the caseness threshold on all administrations of the GHQ-28. Of these five participants only two, participants 10 and 12, submitted complete stress data from their wearable devices. This made them the only candidates who satisfied the criteria. This is admittedly not ideal, but the stress data from participants 2, 16, or 19 was not available for analysis.

The participants selected for the case study aspect of the findings, and their scores on the relevant scales, are presented in summary in Table 4 below. The case studies were selected to represent exemplars of high and low scores on the stress and clinical distress continuums. Case Study 1 is an exemplar of “Above Caseness, Below Average Stress”. Case Study 2 is an exemplar of “Above Caseness, Above Average Stress”. Case Study 3A

and 3B are exemplars of “Below Caseness, Below Average Stress”. Case Study 4 is an exemplar of “Below Caseness, Above Average Stress”.

Table 4

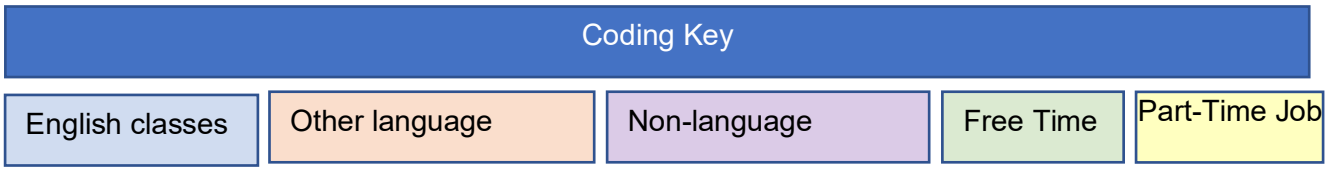
Case Study Participants and Scores

General Health Questionnaire (GHQ-28) (Caseness Threshold 23)	Stress: Below Average (Average 45.09)	Stress: Above Average
Above Caseness	Case Study 1 Participant 12 Stress: 39.56 GHQ-28: 30/23/37	Case Study 2 Participant 10 Stress: 52.44 GHQ-28: 43/*/43
Below Caseness	Case Study 3A Participant 31 Stress: 26.13 GHQ-28: 12/11/4 **Case Study 3B** Participant 28 Stress: 28.64 GHQ-28: 14/15/8	Case Study 4 Participant 30 Stress: 73.88 GHQ-28: 18/16/8

Note: Case study 3 includes two participants because both participant 31 and participant 28 submitted incomplete stress data, so it was decided to include a second similar participant’s data to allow sufficient data for a valid comparison with the other case studies.

6.2. Case Study Structure and Operational Definitions

Each case study will start by presenting data about the participant. Sex, age, GHQ-28 total and sub-scale scores, FLCAS score, TOEIC scores, and average stress values will be summarised. Next, daily stress data will be presented. As in Figure 8 of Chapter 4 the participants’ activities will be coded according to the following key:



Stress levels will be categorised as resting stress (0 to 25), low stress (26 to 50), medium stress (51 to 75), and high stress (76 to 100). After daily stress, the participant's sleep and activity data (in the form of number of steps per day) will be presented as these may help to explain some of the variations seen in daily stress data.

Before starting the case studies a brief word on operational definitions. In conducting this study, measures were chosen for each construct. The rationale for the measures selected, and the limitations of the measures chosen, are discussed at length in Chapters 3 and 4. In previous chapters a great deal of attention has been devoted to the debates surrounding the words "stress" and "anxiety", and the tremendous impact the lack of clarity has had on the scholarship surrounding these terms. For the sake of clarity, the manner in which key terms have been operationally defined in this research will be briefly re-stated. The definitions of these terms are operationalised based on the literature, but in many cases the literature is unclear. It is acknowledged that in some cases the manner in which these terms have been operationalised may be subject to debate. In order to facilitate a clear discussion, the key measures will be briefly restated.

In what follows the construct stress refers to resultant stress drawing on Hebb's (1955) model, which can be seen in Figure 1 in Chapter 2. Stress was measured by the *Garmin Vivosmart 3/4* using HRV, which was then normed using resting HRV by the *Garmin* software and converted into a scale from 1 to 100. Stress readings exceeding the mid-point (51 or higher) may indicate the beginnings of distress, with high (76 or higher) stress levels possibly indicating anxiety. The terms distress and anxiety should not be confused with clinical distress and clinical anxiety.

The GHQ-28 was used to screen participants for "caseness", cases in which distress might be considered to be indicative of a disorder was of clinical concern. This was of importance because the DSM-5 suggests that anxiety disorders, "… differ from transient fear or anxiety, often stress-induced, by being persistent" (American Psychiatric Association, 2013 p. 189). This suggests that in cases of clinical anxiety Hebb's (1955) model may not apply. Participants who scored above caseness on all completed administrations of the GHQ-28, thereby displaying persistence of symptoms for the six weeks of the study, were considered to be possibly experiencing clinical levels of distress. Note that without a diagnosis by a qualified clinician it was impossible to establish whether the distress was indicative of an anxiety disorder. The sub-scales of the GHQ-28 were used to give some indication of the main sources of distress however these are not equivalent to a clinical diagnosis. It should be noted that, as discussed in Chapter 2, while the LLA literature uses the term anxiety it is unclear whether the LLA phenomenon is anxiety as defined today. The

choice of terminology arises from a period in which there was a great deal of confusion over definitions. This was part of the rationale for using the GHQ-28. The GHQ-28 allowed a broader indication of participants' general health, to investigate the main symptoms reported by participants experiencing distress, and whether anxiety was the chief complaint associated with language performance and language learning.

Language performance was operationalised as TOEIC scores, with language learning measured as the difference between the participant's TOEIC scores from the previous year and their latest TOEIC score. The reasons for using the TOEIC are discussed in more detail in Chapter 4, and statements about language learning should be read with the discussion in Chapter 4 in mind.

6.3 Case Study 1: Participant 12 (Above Caseness, Below Average Stress)

Participant 12, a nineteen-year-old female, was the focus of Case Study 1. Their GHQ-28 scores are summarised in Table 5 below.

Table 5

GHQ-28 Scores for Participant 12

	Total	Sub-Scales			
		Anxiety and Insomnia	Severe Depression	Somatic Symptoms	Social Dysfunction
Week 1	30	9	2	11	8
Week 4	23	9	2	11	1
Week 6	37	9	2	14	12

Participant 12 scored 112 on the FLCAS scale. The average participant's FLCAS score was 117.7 out of a possible total of 198.

Participant 12 scored 470 on the TOEIC test on entering the university, against an average participant score of 487.06. One year later (the year of the study) Participant 12 scored 535 on the TOEIC test, against an average participant score of 641.18. In one year their TOEIC score increased by 65 points, which will be used as a reflection of their language learning during that period. The average participant's TOEIC score increased by 154.12 in the same period.

Participant 12's average stress measurements are detailed below in Table 6.

Table 6

Participant 12 Average Stress Measurements by Activity

	Participant 12's Average Stress	Number of Measurements	Average Participant's Stress
English Classes	32.83	437	40.68
Other Language Classes	N/A	N/A	40.83
Non-Language Classes	37.94	276	46.14
Free Time	42.34	1221	44.63
Part-time Job	N/A	N/A	47.91
Overall Average Stress	39.56		45.09

Table 5 shows that Participant 12 is probably experiencing persistent clinical levels of distress, scoring above the caseness threshold for the entire 6-week period of the study. The chief complaint seems to be somatic symptoms, with Participant 12 consistently scoring highest on this sub-scale. In the context of this study it is also noteworthy that Participant 12's anxiety and insomnia scores are their second highest area of concern in the first and second administrations, and remain constant throughout the study.

Despite the clinical levels of distress, with anxiety and insomnia implicated as a secondary area of concern, Participant 12's FLCAS scores are only a little below average (112 against an average of 117.7).

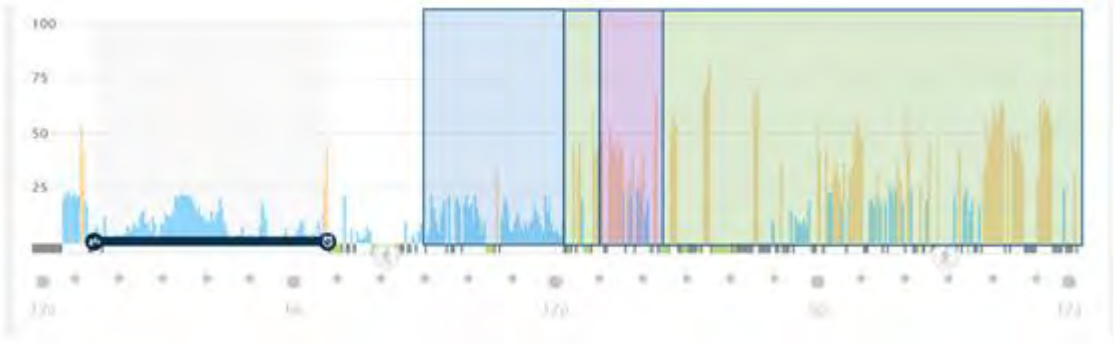
Participant 12's language learning rate seems to be well below average though, improving by just 65 points against an average participant improvement of 154.12, despite starting at 470, only a little below the average participants' TOEIC score of 487.06. This tends to indicate that something is impeding Participant 12's language learning, which the FLCAS does not appear to be predicting.

6.3.1. Case Study 1: Participant 12 Stress

In this section daily stress data will be presented. The first day's stress graph will be presented and then interpreted to provide some guidance on reading what is being shown in the graph. After this, each day's data will be presented without comment and discussed in summary at the end of the section.

Figure 13

Stress Measurement: Participant 12, Day 1 (Monday)



Reading the stress data from left to right shows that Participant 12 went to sleep quite late at approximately 1:30am and woke just before 7am. The grey lines at the bottom of the stress data (for example, the four lines just after waking at 7am) show periods of movement where stress readings could not be calculated. These possibly indicate the participant moving around their house or apartment, getting ready for the day for about half an hour from approximately 7am to 7:30am. This is followed by a period of approximately half an hour of relaxation from about 7:30am to 8pm, followed by leaving home just before 8am and walking to university for approximately 20 to 25 minutes. On arriving at university around 8:20am, they relax for about half an hour before class. The fitness tracker automatically tags activities such as walking, although it is not perfect and in the case of some participants (not participant 12) the software incorrectly tagged walking time as time on an elliptical machine or treadmill.

Looking at the stress data for the early morning the readings range is all blue, indicating values under 25 (readings range from 3 to 23), indicating low stress, in some cases lower than while the individual was asleep.

This is noteworthy because in the classic arousal-stress models, sleep is typically characterised as zero or near-zero stress. What this data shows is that arousal levels during sleep may be higher than arousal levels while awake. This is hardly surprising because during certain types of sleep brain activity levels are similar to, or even higher, than those while awake. This is particularly true of REM sleep, where sleep studies have shown brain wave activity similar to that of individuals who are awake, but often showing higher levels of activity and oxygen consumption (Markov & Goldman, 2006; Reite, et al., 2008).

Participant 12 then attended two English language classes from 9am to 12:10pm. Each class was approximately 90 minutes long, and what is interesting to note is the degree of variation in stress readings within each class, with arousal levels varying seemingly randomly as items in the lecture caught participant 12's attention. There is movement in the

first class (indicated by the grey lines below the stress readings, such as the three periods of movement between 9:30am and 10am) that seem to indicate that the class included some sort of activity, possibly moving into groups for a discussion, which seems associated with higher levels of arousal, but apart from a single spike in stress levels before the start of the second English class at about 10:40am, the stress readings remain firmly in the blue (0 to 25) range, indicating resting levels of stress.

After the English classes is lunch time from 12:10pm to 1pm, which has been flagged as free time. One cannot necessarily assume that participant 12 actually spent lunch time eating lunch, indeed the amount of movement and heightened stress levels, with the majority of stress readings falling into the 26 to 50 range, seems to suggest that participant 12 did not spend this time quietly eating lunch at a table in the cafeteria, and while some portion of the sixty minutes may have been spent this way, they may equally have been up in the computer room printing out an assignment for the next class.

The next 90 minutes, from 1pm to 2:30pm, were a non-language class, and participant 12's stress levels continue to mostly fall in the 26 to 50 range, with a few spikes into the 51 to 75 range, and few dips down into the 0 to 25 range. The degree of variation within this single class is interesting, with the lowest reading at 19 and the highest at 72, indicating the variability of stress within a single period.

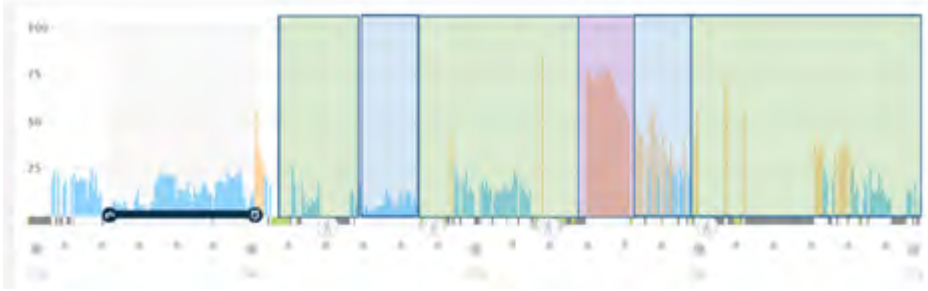
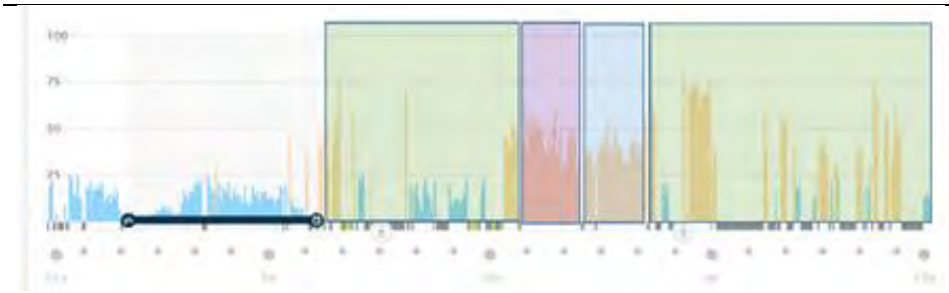
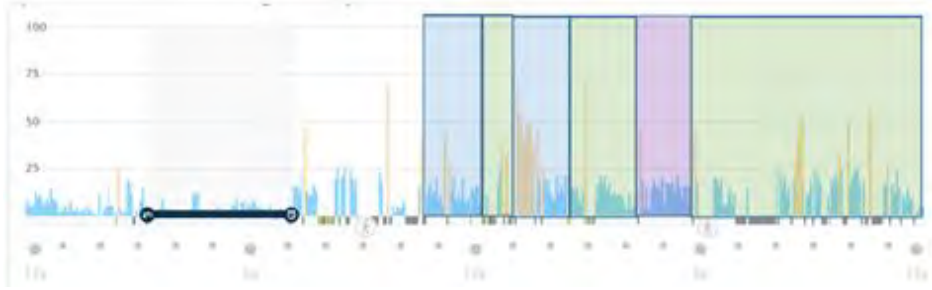
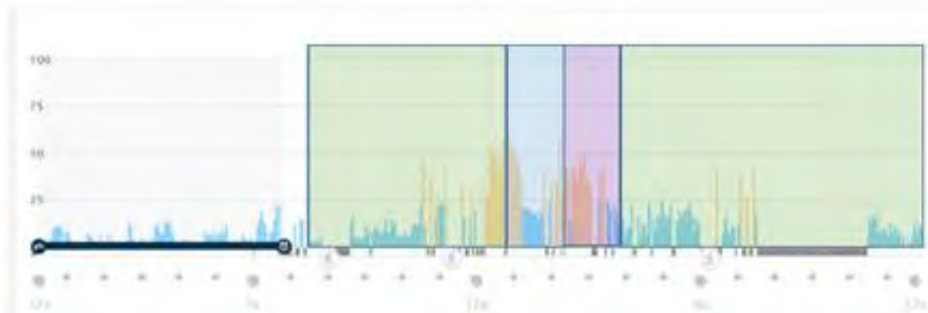
Again, this is interesting in relation to critiquing historical stress studies, which often relied on a smaller number of assessments. These thus may have failed to capture such variations, and earlier studies perhaps therefore theorised arousal levels as being more consistent over a period than this data shows.

After class Participant 12 has more than 9 hours of free time, from 2:30pm to midnight, during which they seem quite active, and quite stressed, with most of the stress readings between 30 and 65. While there are some readings that dip down into the resting stress (0 to 25) range, the majority of the readings are in the low stress (26 to 50) and medium stress (51 to 75) ranges, with a single reading in the high stress (76 to 100) range at about 3:30pm. The readings indicate a further period of exercise at around 9pm, walking for approximately 25 to 30 minutes, and interestingly this is followed by a period of comparatively low stress.

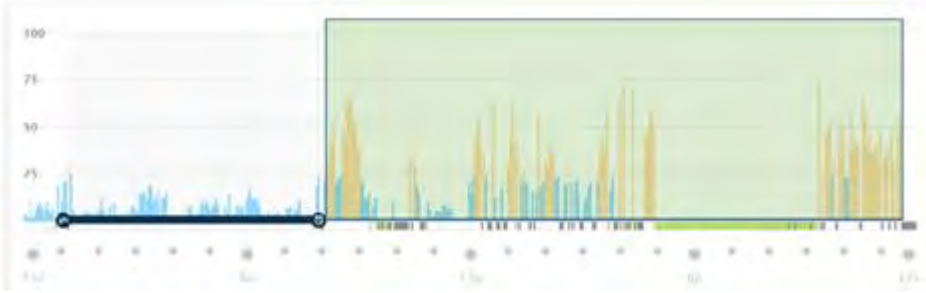
Having analysed one day in detail as an example of how to read the stress data and the various notations, the remaining data for the next twelve days when the *Garmin Vivosmart 3/4* was worn will be presented. The table showing these data will then be followed by a discussion of these to highlight certain features.

Table 7

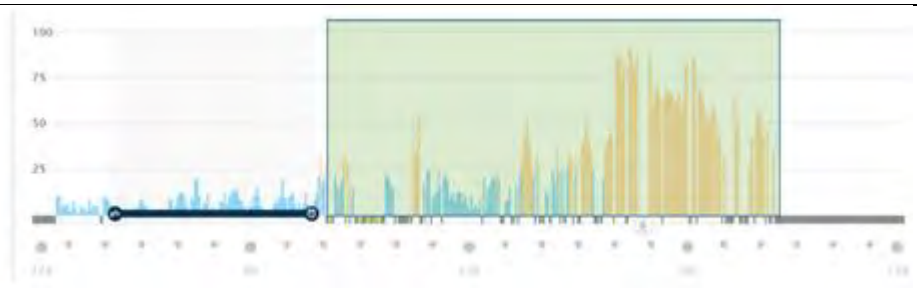
Stress Measurements: Participant 12, Days 2 to 13

Day 2 (Tuesday)

Day 3 (Wednesday)

Day 4 (Thursday)

Day 5 (Friday)


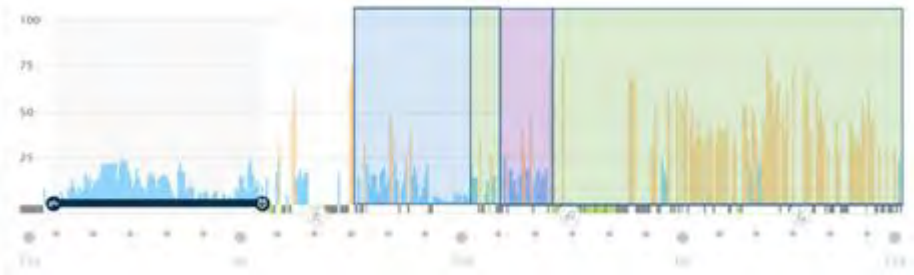
Day 6 (Saturday)



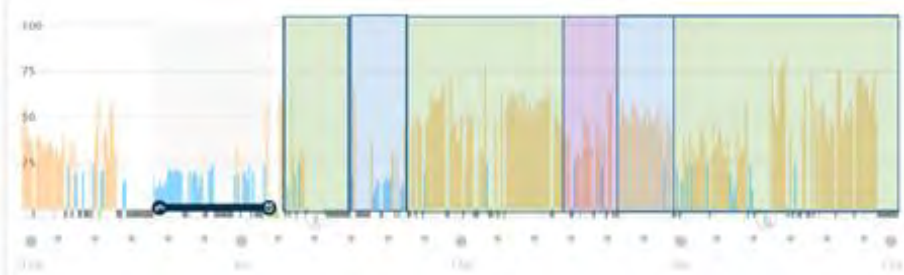
Day 7 (Sunday)



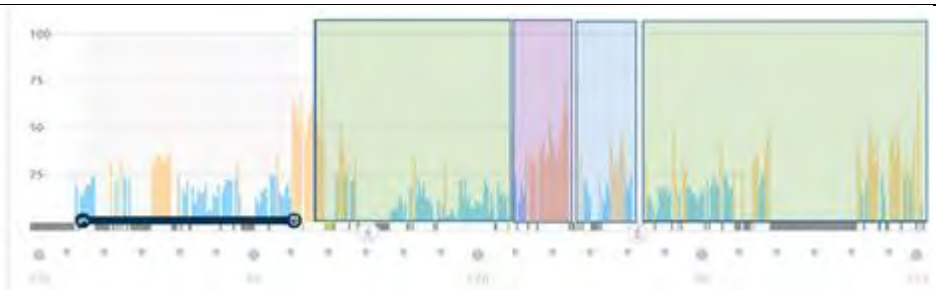
Day 8 (Monday)



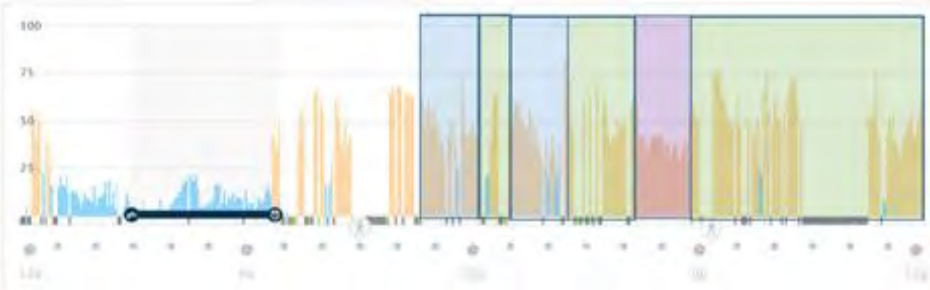
Day 9 (Tuesday)



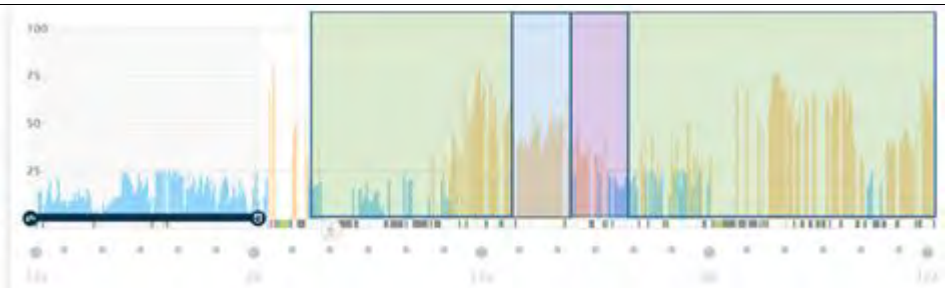
Day 10 (Wednesday)



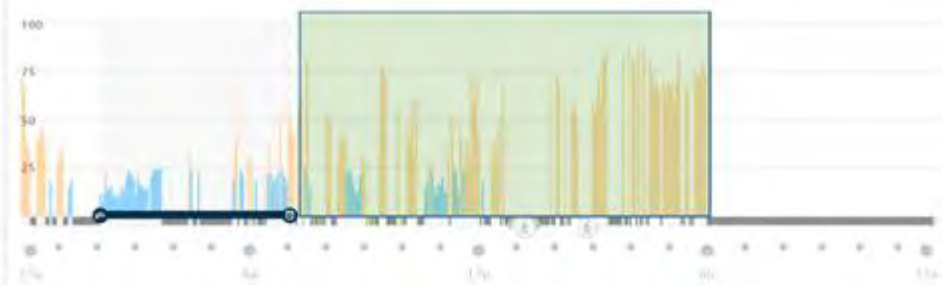
Day 11 (Thursday)



Day 12 (Friday)



Day 13 (Saturday)



Note. Data in Table 7 will be referred to by day. For example, “on Day 8”, rather than “in Table 7, Day 8”.

Participant 12's stress levels in English language learning classes (blue) show repeated spikes, possible indicating moments of stress such as anxiety over being called on in class by the teacher to answer a question. However, these may equally merely indicate the natural waxing and waning of attention during class. It is here that the data from other classes and the participant's free time becomes valuable. Looking at this data the same rapid spikes can be seen both in other classes (see the non-language class on Day 8), and in their free time (on most days, but most notably on Day 10). This sort of pattern of stress seems to be a normal feature of Participant 12's reactions to stress. There are some days where Participant 12's stress levels are unusually high in English classes, such as Day 11 in the first English class of the day Participant 12's stress levels touch the 75 mark, bordering on high stress levels.

Were the English class data the only data being contemplated then this data, coupled with the lower levels of language learning over time, might be viewed as proof for the existence of LLA. However, one of the defining features of LLA is that it is theorised to be context specific. Therefore, the data from the other classes and free time also needs to be considered. Participant 12 shows very similar patterns of stress in other classes. Some days their stress levels are uniformly elevated across all classes (such as Day 11). It seems that the higher stress on this day is not attributable specifically to English classes but may rather be a result of that day being more stressful in general. Again, this displays the importance of contextualising data rather than drawing conclusions from just one context.

On some days Participant 12's stress levels are higher in English class (blue) and lower in non-language (purple) class, such as on Day 12. However, on Day 10 the opposite is true, with higher stress levels in non-language class, and lower in English class. And on Day 9 their stress levels are very low in the first English class of the day, erratic but generally higher in the non-language class, and then highest out of all the classes in the final English class of the day. What is most interesting about Participant 12's data is this lack of consistent stress in any single environment. Stress levels vary rapidly and by a large degree within almost all environments, possibly indicating something more pervasive, such as a dysfunction in emotional regulation.

This case study suggests that with a person experiencing clinical levels of distress, with anxiety implicated as a factor (but not the chief complaint), there is evidence of below average language learning and language performance. However, there is no evidence of situation-specific arousal or stress that explains this impairment. Rather the stress data tends to indicate a global dysfunction in stress regulation across all contexts. Evidence of the global nature of the phenomenon can be seen in Day 9 and Day 11, which show the most

consistency. That consistency is unfortunately achieved by fairly constantly elevated stress levels with comparatively fewer periods of returning to resting stress levels.

This has bearing on Research Question 1, namely whether LLA exists only in the language learning environment, as theorised in LLA. At least in the case of Participant 12 the evidence suggests that language learning impairment may be linked to clinical distress where anxiety is a component, but not the chief complaint. This may manifest in the language learning classroom but is not specific to this environment. Further, in Participant 12's case the chief complaint is somatic, with anxiety as a possible secondary factor. This raises the possibility that language learning teachers may be seeing the secondary anxious symptoms in the classroom and are unaware of the chief complaint. This might lead to the erroneous conclusions that the anxious symptoms are specific to that environment and arise from anxiety as the underlying cause.

This point was raised earlier in Chapter 2, when discussing the development of the FLCAS. There seemed to be no engagement in the LLA literature with the idea that anxiety seen in the language learning environment might have its genesis outside of the classroom, in a clinical condition unrelated to the classroom or teacher. This observation has bearing on Research Question 1, whether LLA exists as theorised in the current LLA literature.

On both Days 9 and 11, the days where Participant 12 showed consistently elevated stress levels across the day, the participant also received fewer hours of sleep than was their norm, so in the next section Participant 12's sleep data will be examined more closely.

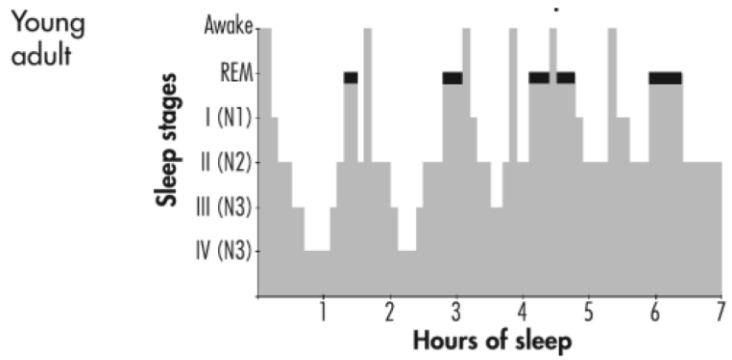
6.3.2. Case Study 1: Participant 12 Sleep Patterns

The stress graphs provided in Table 7 provide some idea of the participant's total hours of sleep, however the *Garmin Vivosmart 3/4* collects additional data about sleep stages which will be presented in this section.

This additional data is important, because not all sleep is equal, and different sleep stages serves different functions. The anatomy of sleep can be divided up in different ways, and the *Garmin Vivosmart 3/4* records three different types of sleep (light, deep, and REM), as well as periods of wakefulness. Disturbed sleep tends to skew the normal distribution of sleep towards more light sleep, and less deep and REM sleep.

Figure 14

Sleep Hypnogram of a Typical Young Adult



Note. From page 28 of "Clinical manual for evaluation and treatment of sleep disorders" by Reite, et al. in 2008, printed by American Psychiatric Publishing. Reprinted with permission.

Figure 14 shows the typical sleep pattern for a healthy young adult, similar to the participants in this study. Sleep stages I and II can be considered analogous to light sleep, with III and IV as deep sleep. REM sleep occurs primarily at the end of each sleep cycle of approximately 90 minutes, and in uninterrupted sleep the duration of REM sleep increases with each sleep cycle. The amount of REM sleep obtained is therefore heavily dependent on the number of completed sleep cycles. Looking at the sleep hypnogram above one can see that the amount of REM sleep obtained at the end of the fourth sleep cycle is nearly double the total amount obtained in the first two sleep cycles. As a rule of thumb REM sleep should comprise approximately 20% to 25% of total sleep, with a typical young adult experiencing approximately 84 to 105 minutes of REM sleep a night (Reite, et al., 2008).

REM sleep is important for certain types of memory formation, particularly procedural memory, which is required for complex processes such as learning and apply grammatical rules (MacDonald, 2015). This is relevant to the discussion of LLA and language learning.

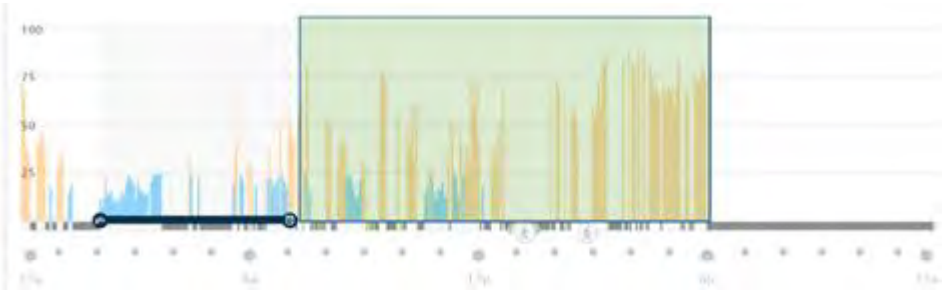
In this case, perhaps more importantly, REM sleep is often associated with emotional self-regulation processes and recovery from negative experiences (Reite, et al., 2008). As a result, this pattern of disturbed sleep is common in cases of anxiety and is listed in the DSM-V as one of the diagnostic criteria for generalized anxiety disorder, "Difficulty sleeping (due to trouble falling asleep or staying asleep, restlessness at night, or unsatisfying sleep)" (American Psychiatric Association, 2013, p. 222).

For these reasons, sleep was considered as a variable of special interest in these case studies.

Looking at the sleeping times in Table 7 shows that Participant 12's sleep onset times are erratic, varying from midnight at the earliest to 3h40 on Day 9, yet waking within about an hour of 7am.

Figure 15

Stress Measurement: Participant 12, Day 13 (Sunday)



The additional data on sleep types should be considered in concert with the stress measurements already presented in Figure 13 and Table 7. The data from Day 13 has been repeated in Figure 15 above for the reader's convenience to briefly discuss some important features of the sleep data contained in the stress measurements presented previously, before proceeding on to the additional sleep data.

Three features will be highlighted for consideration. The first is the black line that indicates the period of sleep, marking the beginning of sleep with a Zzz icon, and the end of sleep with a clock icon. This line is an indicator of total sleep quantity, but does include brief periods of wakefulness, so should be treated as an approximate value only. This is also important in considering the regularity of Participant 12's sleep patterns. On Day 13 participant 12 slept from approximately 02h00 to 07h00.

The second feature to note in Figure 15 is the grey lines at the bottom of the graph, indicating movement. Looking at Participant 12's sleep on Day 13 there are a great many of these lines, indicating a great deal of movement, possibly tossing and turning in their sleep. These would have significantly negatively impacted Participant 12's sleep quality, which is an important consideration when discussing sleep. It is also important to note that this degree of movement interferes with stress measurements. It seems likely that they are indicative of periods of heightened stress during sleep. However, this is speculation and indicates a possible limitation on how wearable devices can be used in stress research, which has a bearing on Research Question 5.

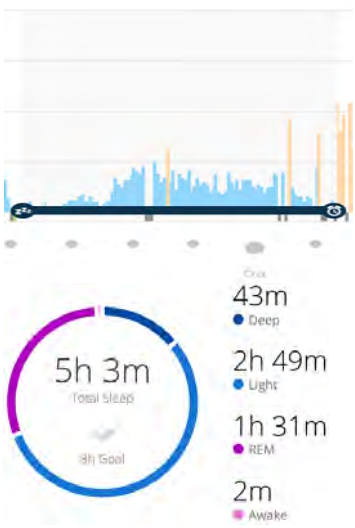
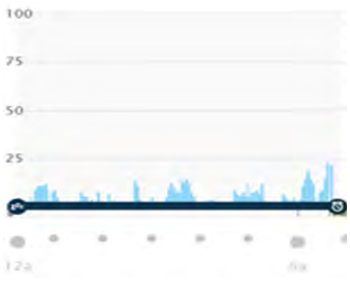
Thirdly, consider Participant 12's stress measurements on Day 13. In theory sleep stress measurements should be in the blue resting stress category (0 to 25), however a noticeable quantity of Participant 12's sleep falls outside this category, and is marked in

orange, falling into the low (26 to 50) and medium (51 to 75) stress categories. Unfortunately stress measurements are unavailable for a significant portion of Participant 12's sleep on Day 13 because of movement. This has a bearing on the quality of Participant 12's sleep, which is poor.

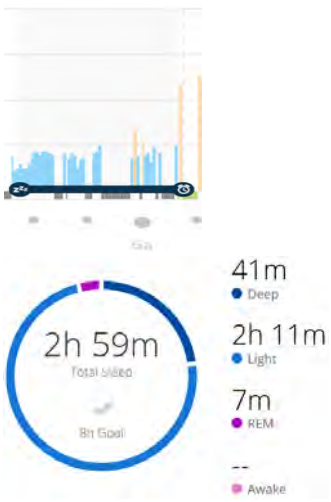
The following additional data presented in Table 8 is intended to be read in concert with the sleep data contained in the stress measurements in Table 7. For the reader's convenience the sleep portion of the stress data has been extracted from the total stress data and is presented above the sleep stage data.

Table 8

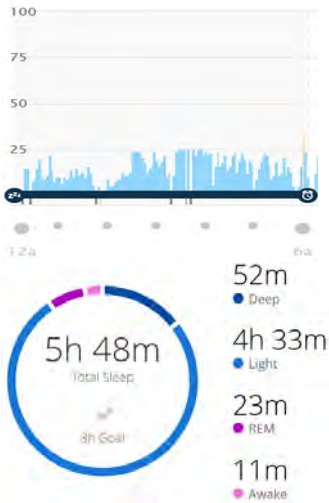
Sleep Data: Participant 12, Days 1 to 13

Day 1 (Monday)	Day 2 (Tuesday)	Day 3 (Wednesday)
		
Day 4 (Thursday)	**Day 5 (Friday)**	**Day 6 (Saturday)**
		

Day 7 (Sunday)	Day 8 (Monday)	Day 9 (Tuesday)



Day 10 (Wednesday)	Day 11 (Thursday)	Day 12 (Friday)



Day 13 (Saturday)

Looking at the data in Table 8 suggests the likely reason for Participant 12's persistently elevated stress levels on Day 9. Having slept for less than three hours is likely to interfere with emotional regulation and recovery, as well as making even the simplest tasks more difficult. Looking at the sleep data in general there is a consistent pattern over the thirteen days of observation of sleep insufficiency. There were only two nights of more than six hours of sleep (Days 5 and 6), and four nights where Participant 12 slept for less than four hours a night (Days 2, 4, 9, and 11).

Further, on four nights (Day 3, 9, 10, 13) there is evidence of low-quality sleep, with grey lines indicating movement and stress measurements spiking above resting levels during the sleep period. As a result, the percentages of REM sleep are also extremely low on most nights. As noted earlier for the average young adult 20 to 25% of total sleep should be REM sleep. However, on Day 13 out of 297 minutes of sleep only 10 minutes were REM sleep, just 3.37% of total sleep.

On Day 6 Participant 12 experienced closest to Reite, et al.'s (2008) suggested seven hours of sleep, with a total of six hours and fifty minutes of sleep, with no evidence of disturbed sleep. Participant 12 accumulated 114 minutes of REM sleep on Day 6, which is within Reite, et al.'s (2008) suggested range for healthy sleep for the average young adult. This indicates that Participant 12 is capable of healthy sleep when conditions are right.

This sleep data raises an interesting question regarding Research Question 2, how LLA affects language learning or language performance. Sleep was included in the variables considered in the case studies because it is an essential process for the regulation of

virtually every system in the human body and mind. Not only does sleep affect learning over time, but also performance on the day of a test (MacDonald, 2015), and other processes such as stress regulation and both mental and general health (Chiang, et al., 2019). Establishing causality in Participant 12's case is not possible. It may be that their somatic symptoms interfered with sleep, causing the slow sleep quantity and quality, or that low sleep quantity and quality caused or exacerbated their somatic symptoms. Similarly, the anxious component seen in Participant 12's GHQ-28 scores may be the cause of their sleep problems, or a result.

What is important though is that if the thirteen days of sleep data from Participant 12 are representative of their normal sleep patterns, then it suggests the below-average language performance and language learning could be sleep-related. When someone is having less than six hours of sleep a night for eleven out of thirteen nights, then the LLA construct is not required to explain why this individual is performing worse than average in language classes. Nor is changing the class environment necessarily going to help much. What Participant 12 most needs is to regularise their sleep schedule, although the mechanism for achieving that outcome that will depend on the underlying source of clinical distress. Until the underlying cause is addressed, it may be that the language learning and language performance impairment is merely one of many symptoms. While symptom alleviation has its place, one of the problems with LLA theory is that it risks misrepresenting the symptom as the cause and may thereby discourage individuals from seeking help in addressing the real cause.

A final variable was singled out for special attention in the case studies, namely activity data. As discussed in the literature review in Chapter 2, exercise can help with stress regulation.

6.3.3. Case Study 1: Participant 12 Activity

The *Garmin Vivosmart 3/4* was designed to capture activity data related to fitness and exercise. As exercise is a well-known moderating variable in both physiological and psychological health, including some sort of measure of the participants' activity levels seemed sensible (Salmon, 2001; Contrada & Baum, 2011).

Table 9

Activity Data: Participant 12, Days 1 to 13

Day 1 (Monday)	Day 2 (Tuesday)	Day 3 (Wednesday)
Steps: 13,193 Kilometres: 10.2	Steps: 15,908 Kilometres: 11.8	Steps: 8,518 Kilometres: 6.5
Day 4 (Thursday)	**Day 5 (Friday)**	**Day 6 (Saturday)**
Steps: 8,688 Kilometres: 6.6	Steps: 8,675 Kilometres: 6.4	Steps: 7,564 Kilometres: 5.9
Day 7 (Sunday)	**Day 8 (Monday)**	**Day 9 (Tuesday)**
Steps: 7,186 Kilometres: 5.4	Steps: 11,669 Kilometres: 9.3	Steps: 7,642 Kilometres: 5.7
Day 10 (Wednesday)	**Day 11 (Thursday)**	**Day 12 (Friday)**
Steps: 4,997 Kilometres: 3.7	Steps: 9,506 Kilometres: 7.2	Steps: 8,465 Kilometres: 6.5
Day 13 (Saturday)		
Steps: 10,827 Kilometres: 8.1		

Participant 12 showed consistently high levels of physical activity across almost the entire study period, only dipping below 7,000 steps on one day, and walked more than 10,000 steps on four of the thirteen days for which data was available. This is important in terms Salmon's (2001) research, covered in Chapter 2, which suggested that regularity was important in considering the effect of exercise on mood. Participant 12 seems to exercise regularly and vigorously.

Day 2, where Participant 12 walked nearly 16,000 steps was unusually intense, but as predicted in Salmon (2001), while the couple of stress measurements available during the evening walk suggest very high levels of stress, there seems to be a significant decrease in stress after exercise. Participant 12's stress levels slowly decrease after the walk, reaching

resting levels an hour after the walk and staying at that level until Participant 12 goes to sleep.

Participant 12 does experience more sleep than normal after the walk on Day 2, sleeping for just over five hours. The first half of the night appears to be undisturbed sleep, however the sleep shows signs of disturbances and moments of stress during the second half of the night.

The exercise data seems to suggest that Participant 12's high stress levels and poor sleep quality and quantity are not a result of a lack of physical activity. While Participant 12 does appear to experience some short-term stress relief after exercise, as predicted by Salmon (2001), the moderating effects of exercise do not appear to be sufficient to ensure good sleep for the whole night.

6.4. Case Study 2: Participant 10 (Above Caseness, Above Average Stress)

Participant 10, a twenty-year-old male, was the focus of Case Study 2. Their GHQ-28 scores are summarised in Table 10 below.

Table 10

GHQ-28 Scores for Participant 10

	Total	Sub-Scales			
		Anxiety and Insomnia	Severe Depression	Somatic Symptoms	Social Dysfunction
Week 1	43	10	9	6	18
Week 4	Did Not Complete				
Week 6	43	9	5	12	17

Participant 10 did not complete the mid-study GHQ-28, but their consistently very high GHQ-28 scores, scoring 43 on both the first and final assessments above the caseness threshold of 23, strongly suggests that the clinical distress was persistent for the duration of the study. The chief complaint on the GHQ-28 is some sort of social dysfunction, however there is a consistent element of anxiety, scoring 10 in the first assessment and 9 in the final assessment. Participant 10 reported no history of depression or anxiety, nor any current problems with depression or anxiety. Under the final question in the screening section, "Is there any other reason why you think your stress levels may be unusual?" they responded "Exhausted. Lack of sleep." in Japanese. This may suggest that Participant 10 had not

received a clinical diagnosis and was not seeing a mental health professional about their problem, if they considered the lack of sleep to be the primary issue.

Participant 10 scored 127 on the FLCAS scale. The average participant's FLCAS score was 117.7 out of a possible total of 198.

Participant 10 scored 585 on the TOEIC test on entering the university, against an average participant score of 487.06. One year later (the year of the study) Participant 10 scored 620 on the TOEIC test, against an average participant score of 641.18. In one year their TOEIC score increased by 35 points, which will be used as a reflection of their language learning during that period. The average participant's TOEIC score increased by 154.12 in the same period.

As with Case Study 1 the FLCAS appears of no predictive value in this case. While Participant 10's FLCAS score is slightly above the average their level of language performance and learning is quite extreme. Despite entering university with a TOEIC score well above average, nearly a hundred points higher than the average student, Participant 10 has slipped to twenty points below average after a year. In all fairness the FLCAS claims to measure foreign language class anxiety, and Participant 10's classroom stress scores are also slightly above average. Perhaps in this case the FLCAS is predicting the anxiety in English classes but is failing to predict the degree of language learning impairment in total.

Participant 10's average stress measurements are detailed below in Table 11.

Table 11

Participant 10 Average Stress Measurements by Activity

	Participant 10's Average Stress	Number of Measurements	Average Participant's Stress
English Classes	41.34	223	40.68
Other Language Classes	N/A	N/A	40.83
Non-Language Classes	40.32	334	46.14
Free Time	56.68	1538	44.63
Part-time Job	N/A	N/A	47.91
Overall Average Stress	52.44		45.09

Participant 10's average stress scores appear to be only a little above average, 7.35 points higher than the average of 45.09. Their stress levels are below average in non-language classes, very slightly above average in English language classes, and highest during their free time where they are well above average. The GHQ-28 scores indicate that the chief complaint is some sort of social dysfunction, and so these stress scores make sense. Language classes tend to involve more interaction than non-language classes, whether with the teacher or with other students, such as during group work or paired activities. However, classroom interactions occur in a structured setting with clear guidelines for expected behaviour, and with the teacher acting as a moderator in case of disagreements, misunderstandings, or other problems. For someone who finds social settings distressing, the classroom environment, even in classes where some interaction is required, may be less distressing than the unstructured interactions that occur outside of the classroom. In unstructured interactions there are no clearly stated rules, no clear goals, and no moderator in case of difficulties, and therefore the higher stress scores seen during unstructured time seem to make some sense in this case.

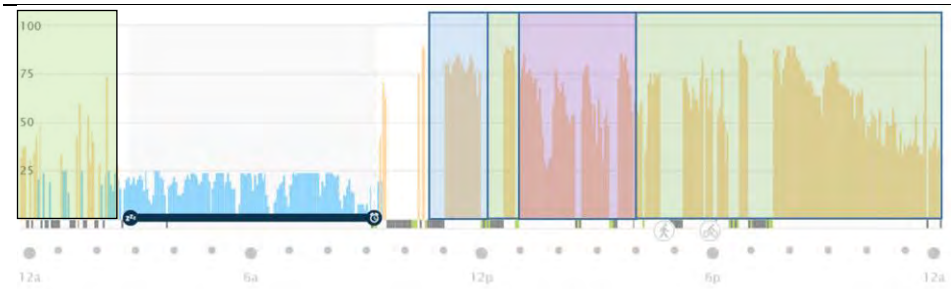
The following section will examine the stress scores in more detail, looking for elevated levels of stress in language classes, either in the form of persistent elevation or spikes.

6.4.1. Case Study 2: Participant 10 Stress

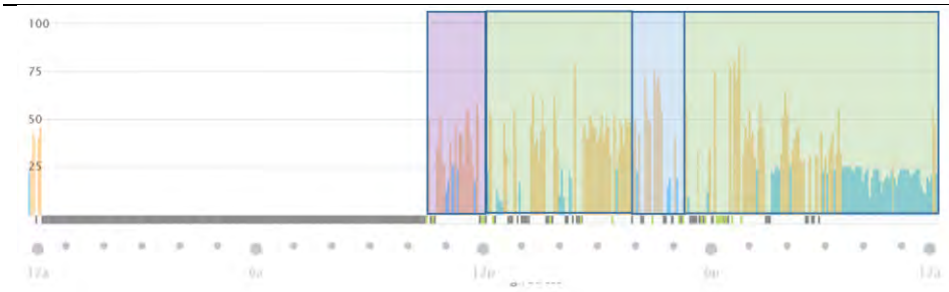
As with Case Study 1 the stress measurements for Participant 10 will be presented in a table below, and then discussed afterwards.

Table 12

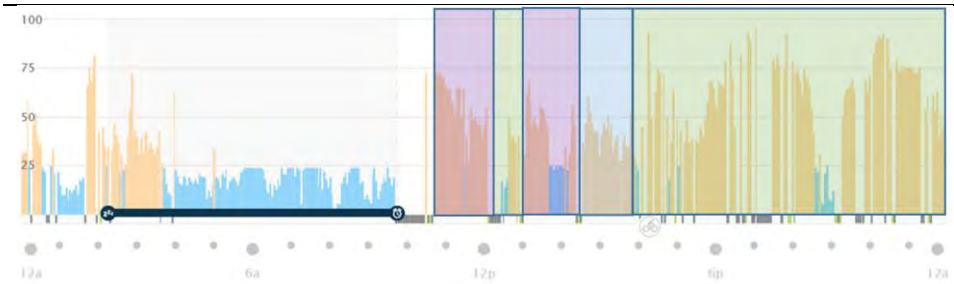
Stress Measurements: Participant 10, Days 1 to 14

Day 1 (Monday) (No data 10am to 11am)


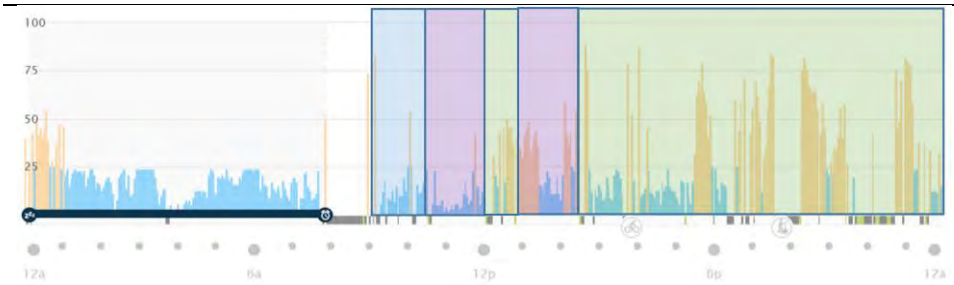
Day 2 (Tuesday) (No data from midnight to 11am)



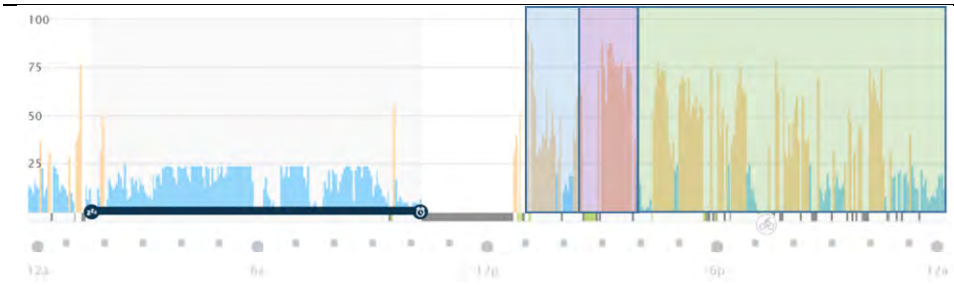
Day 3 (Wednesday) (No data from 10am to 11am)



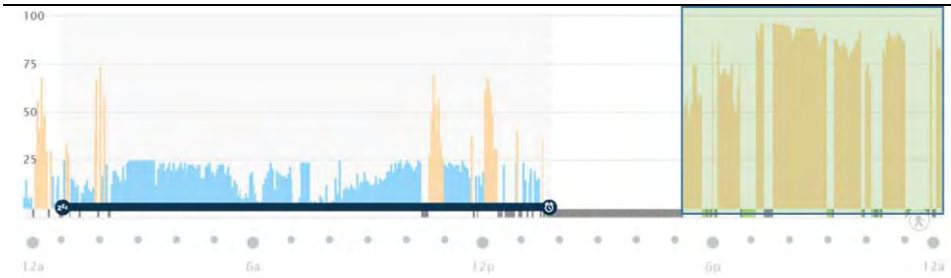
Day 4 (Thursday) (No data from 8am to 9am)



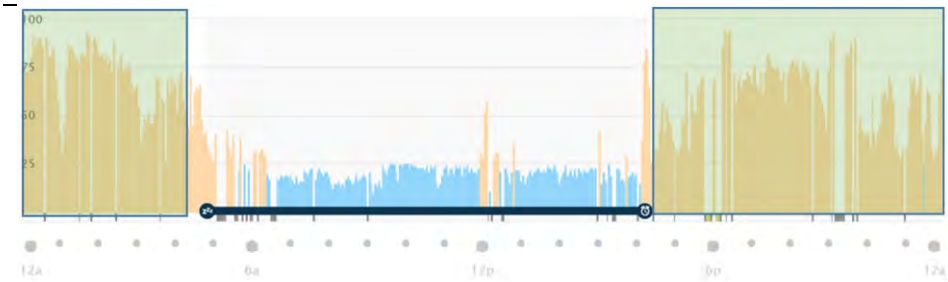
Day 5 (Friday) (No data from 10am to 1pm)



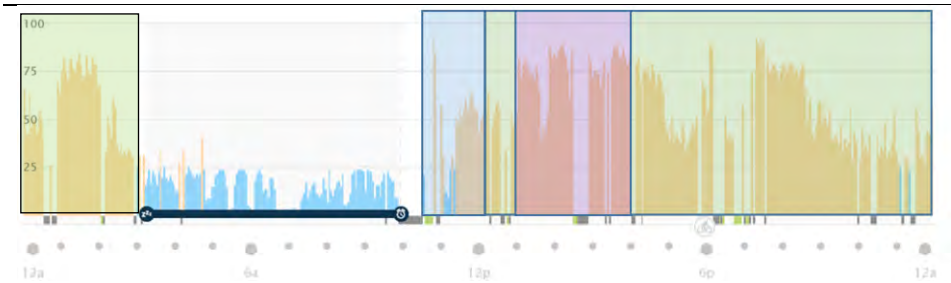
Day 6 (Saturday) (No data from 2pm to 5pm)



Day 7 (Sunday)

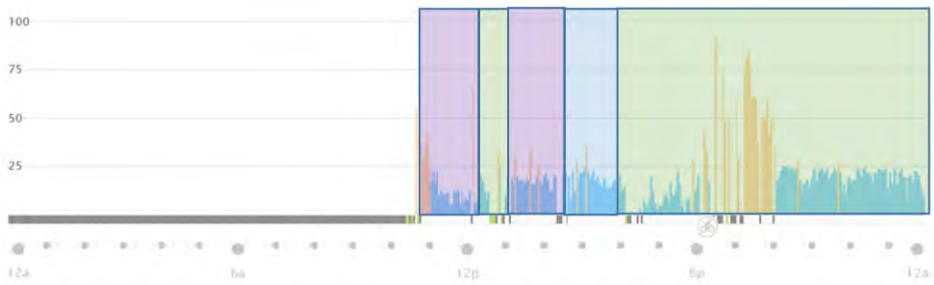


Day 8 (Monday)

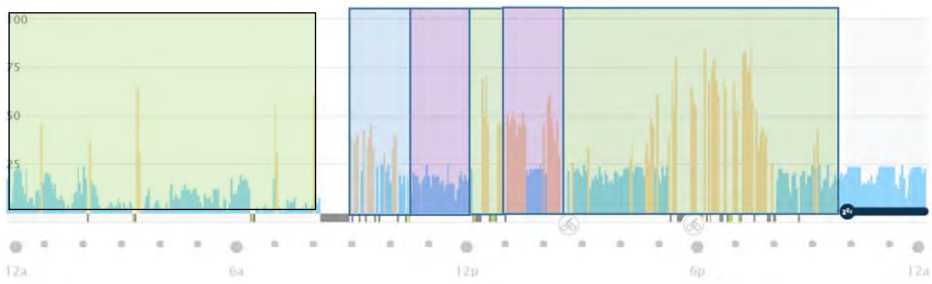


Day 9 (Tuesday) (No data from 3am to next day)

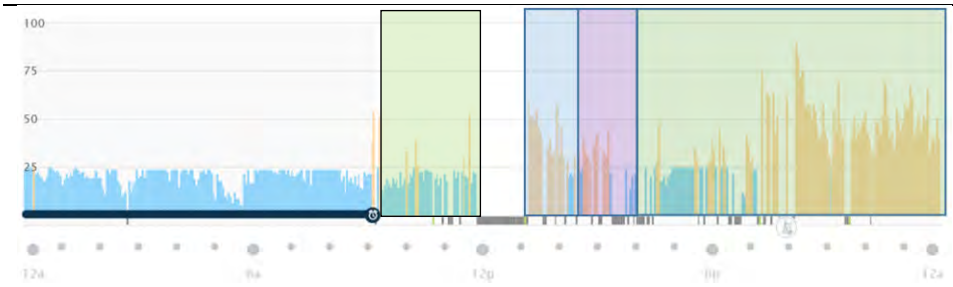
Day 10 (Wednesday) (No data from midnight to 11am)



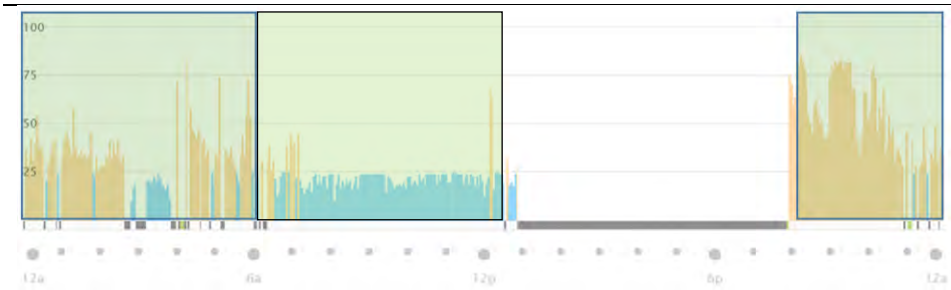
Day 11 (Thursday) (No data from 8am to 9am)

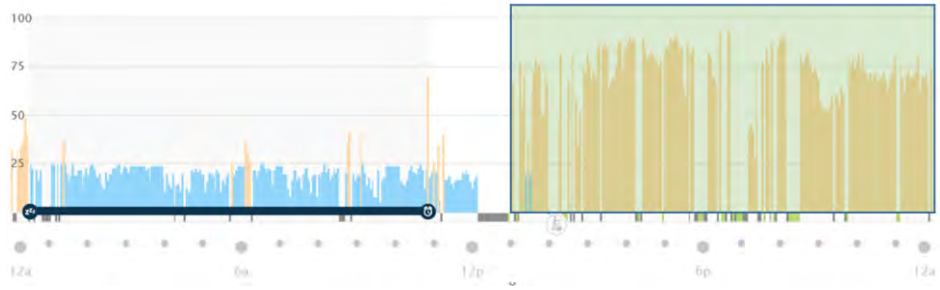


Day 12 (Friday) (No data from noon to 1pm)



Day 13 (Saturday) (No data from 1pm to 8pm)



Day 14 (Sunday) (No data from noon to 1pm)


Note. Data in Table 12 will be referred to by day. For example, “on Day 8”, rather than “in Table 12, Day 8”.

Participant 10 removed the *Garmin Vivosmart 3/4* several times the study. Most of these appear to be for brief periods after waking, such as on Days 1, 3, 4, 11, 12, and 14. The simplest and most likely explanation is that Participant 10 either did not understand that the *Garmin Vivosmart 3/4* was waterproof, and so removed it to shower after waking in the morning. It might also merely be Participant 10’s habit to remove all clothing before showering. These brief periods of not wearing the device did not appear to be significant as there was already a large amount of data on Participant 10’s stress patterns outside of classes. This also explains the gaps on Day 5 and Day 6, where the wearable device was removed after waking, and only put back on a couple of hours later. It is probable that Participant 10 only put the device back on when preparing to go out.

There were two longer gaps in the data. On Day 2, from midnight to 11am Participant 10 appears to have removed the wearable device. Again from 3am on Day 9 to the next day, Day 10, at 11am there is a significant loss of data. The *Garmin Vivosmart 3/4* advertises a five-day battery life, however use of the device suggests that battery life is closer to seven days. The timing would tend to suggest that Participant 10 removed the device to charge it overnight on Day 2, and then removed it again to charge on Day 10, and simply forgot to put it on in the morning, resulting in no readings over that period.

While the loss of data from classes on Day 10 is regrettable, there is still more than sufficient data for analysis. This has a bearing on Research Question 5, in terms of establishing how long the wearable device needs to be worn. The literature discussed in Chapter 4 suggested at least a week of data, however that was doubled in this study to allow for the sort of erratic wearing seen in Participant 10’s case, and in this case, it appears to be sufficient.

Examining Participant 10’s stress data there is again the same great deal of variability that was seen in Case Study 1, although Participant 10’s stress levels are in the

high (76 to 100) range more often, and more consistently. On Days 1, 3, 6, 7, 8, and 14 the majority of stress readings (outside of sleep) are in the medium (51 to 75) or high (76 to 100) range, with relatively few periods of rest. The periods of elevated stress seem to persist for hours on end, with no evidence of a return to normal, such as on Day 14 where from 2pm to midnight Participant 10's stress levels are in the upper medium to high range. This failure to recover from stress was noted earlier in the literature as a key determinant in identifying disfunction in normal stress regulation systems and as an indicator of clinical levels of distress (Dwight, et al., 2005; Evans, et al., 2005, Lupien, et al., 2006; American Psychiatric Association, 2013). Given that Participant 10's results on the GHQ-28 suggest clinical distress this is not unexpected but is important when considering the utility of wearable devices in research, Research Question 5.

What is interesting is that while Participant 10 shows a dysregulation in stress management on some days, it does not manifest as clearly every day. Consider Day 1's consistently medium to high stress levels, with no recorded periods of resting stress levels (0 to 25) from 11am to midnight. Then the next day, on Day 2, there are some periods of high stress, and a large proportion of medium stress, but also large portions of the day where Participant 10 returns to resting stress levels. While it would not be accurate to characterise Day 2 as a good day, it is a comparatively good day when compared with Day 1.

Regarding LLA, while there is evidence of periods of elevated stress and spikes in stress during English language learning classes, the same can be said for non-language classes and Participant 10's free time. As in Case Study 1 the stress data suggests that in clinical cases the impairment is global, not specific to language learning classes. While clinical distress does seem to impair language learning and language performance, it is not a situation-specific stressor as postulated in LLA theory. The data from Participant 10 adds to the suggestion that LLA is not a clinical phenomenon, as in both case studies involving participants with clinical levels of distress the data suggests global, rather than situation-specific, impairment of functioning.

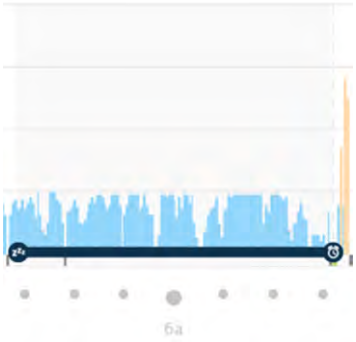
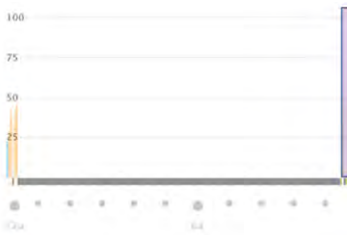
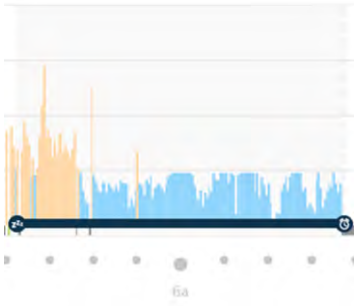
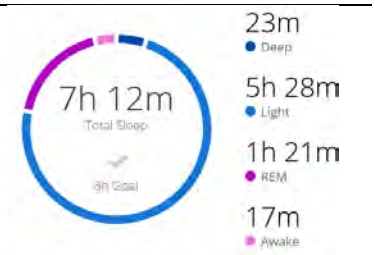
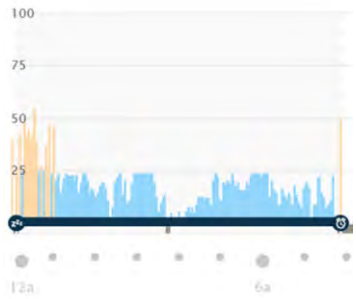
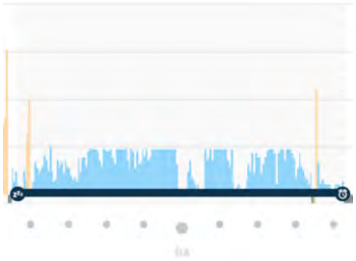
In the first case study the sleep data proved to be quite valuable in explaining the variations in daily stress measurements, and Participant 10 specifically noted that sleep insufficiency as a factor in their case, so the next section will explore the sleep data.

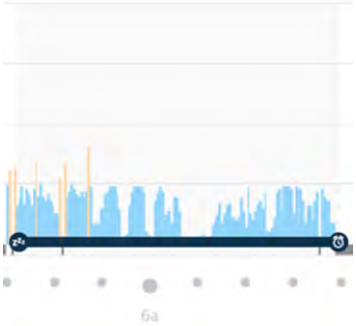
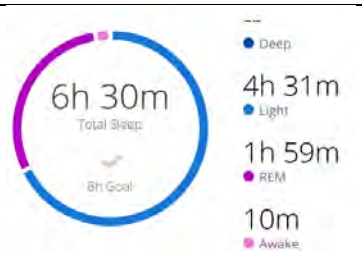
6.4.2. Case Study 2: Participant 10 Sleep

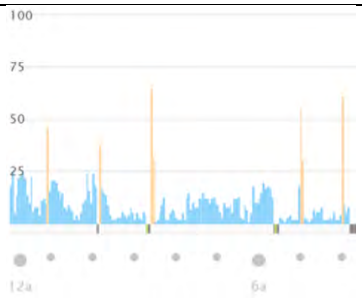
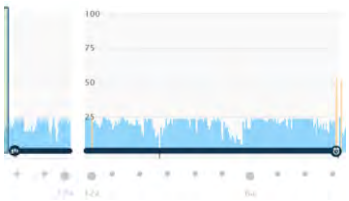
Participant 10's sleep data was incomplete, however there is still sleep data for nine out of the fourteen days, and possibly for two more days, for reasons that will be discussed below.

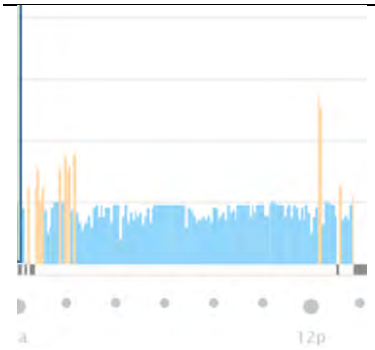
Table 13

Sleep Data: Participant 10, Days 1 to 14

Day 1 (Monday)	Day 2 (Tuesday)	Day 3 (Wednesday)
		
		
Day 4 (Thursday)	**Day 5 (Friday)**	**Day 6 (Saturday)**
		
		

Day 7 (Sunday)	Day 8 (Monday)	Day 9 (Tuesday)
		Fitness Tracker Not Worn No Data
		

Day 10 (Wednesday)	Day 11 (Thursday)	Day 12 (Friday)
Fitness Tracker Not Worn No Sleep Data		
		

Day 13 (Saturday)	Day 14 (Sunday)
	
	

Note. Day 11 and Day 13's data is not labelled as sleep by the fitness tracker but may be sleep that was not correctly identified by the software.

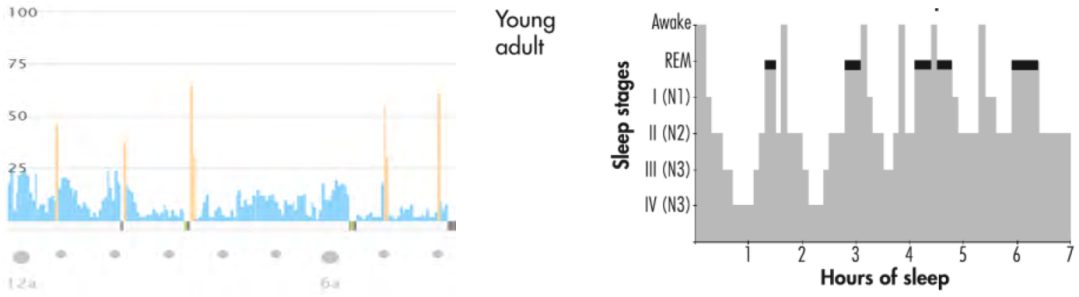
While Participant 10's sleep quantity does not appear to be insufficient, being consistently over 6 hours, their sleep patterns are irregular with sleep onset anywhere between 22h00 (10pm) and 05h00 (5am). What is immediately apparent is that weekend sleep quantities are significantly higher than weekday sleep quantities, with eight to eleven hours of sleep on Fridays, Saturdays, and Sundays, as compared to six to seven and a half hours of sleep on weekdays. These longer sleeping hours on the weekend suggest that Participant 10 was not sleeping enough during the week and was accumulating a sleep debt that needed to be repaid on the weekends.

This inconsistency in sleeping patterns highlights an important issue in the current limitations of wearable device technology that is relevant to Research Question 5. Mechanically, it is difficult to tell the difference between someone lying on a couch watching television and someone who is sleeping. As was noted in the previous case study, the assumption that sleep is characterised by resting (0 to 25) levels of stress is not necessarily true. The fitness tracker relies on the user inputting a normal window for when they sleep. The software then examines this window for patterns normally associated with sleep, as shown earlier in the sleep hypnogram, during the hours flagged for sleep to conclude that the individual was probably asleep during that period. However, there is the potential for the software to possibly err. Participant 10's data provides two examples of types of errors that may occur on Days 11 and 13.

Table 14 below contains a portion of Participant 10's data from Day 11 that is suspected to be sleep. It is presented alongside the typical sleep hypnogram from Reite, et al. (2008) so that an easy comparison can be made.

Table 14

Sleep Data: Participant 10, Day 11 Compared to Typical Sleep Hypnogram



Note. The Sleep Hypnogram to the right is from page 28 of "Clinical manual for evaluation and treatment of sleep disorders" by Reite, et al. in 2008, printed by American Psychiatric Publishing. Reprinted with permission.

A comparison of the stress data and hypnogram suggests that Participant 10 was probably asleep from shortly after midnight to about 08h00 on this day. Participant 10's stress data shows the characteristic V shape of classic sleep from midnight to 2am as seen in the sleep hypnogram to the right in Table 14 above. In addition, it seems unlikely that someone would be sedentary for seven and a half hours, moving only briefly four times during the whole period. There are good reasons to suspect that Participant 10 was asleep during this period, and that the *Garmin Vivosmart 3/4* simply failed to automatically flag the period of sleep.

What may have confused the fitness tracker's heuristics is that Participant 10's sleep cycles seem to be irregular in both duration and pattern. For example, in Reite, et al, (2008) a sleep cycle is typically approximately 90 minutes in duration, and in the early stages of sleep this can be described by a valley-like progression from light sleep, down into deep sleep. This V-shape can be seen in Participant 10's data, they complete the valley-pattern in less than the expected amount of time, in approximately an hour as opposed to the normal 90 minutes that might be expected. This may have thrown off the heuristics designed to track sleep.

There was also a second incident where sleep data may not have been correctly labelled on Day 13. Looking at the data in Table 13 for Day 13 it does not follow the V-shape predicted in Reite, et al. (2008), and so does not resemble classic sleep. Rather the stress data describes a five-hour period from shortly after 7am to noon of no movement, with stress levels barely within the resting (0 to 25) stress band.

The data does not follow the pattern for regular sleep, however it does resemble the pattern for light sleep or napping, although five hours may be stretching most people's definition of a nap. The lack of movement for five hours, combined with the lack of sleep since the day before, and the time period does tend to strongly suggest sleep though. Judging by the stress data it appears to be mostly light sleep, sleep stages 1 and 2 in Reite, et al.'s (2008) sleep hypnogram.

The stress pattern is atypical, and possibly more importantly it falls outside of the period that Participant 10 labelled as their typical sleeping time (which Participant 10 entered into the application as 23h00 to 07h00). Similar sleep patterns are seen on Days 1, 12, and 14. These are possibly identified as sleep because they fall within the period Participant 10 identified as their normal sleeping time.

Regarding stress and sleep, Participant 10's data suggests that their sleep issues may require more detailed analysis. Consider Day 1, where Participant 10 receives what on the surface appears to be six hours of relatively undisturbed sleep with only one brief period of movement. In terms of sleep quantity and movement it looks like good sleep. Yet when one looks at the stress measurements for the rest of Day 1 it is very clear that Participant 10 had a highly stressful day, with no periods of rest (0 to 25), and many periods of high (76 to 100) stress during the day. Compared to Participant 12 in Case Study 1, the data seems to make no sense.

However, when examining Participant 10's stress measurements during sleep on Day 1 they are tightly clustered towards the top of the resting (0 to 25) range. As noted earlier, this is less like healthy sleep and more like an extended nap. Looking at the detailed sleep breakdown by type there is no REM sleep in the entire six-hour period, and only about a quarter of the sleep is deep sleep. Reite, et al. (2008) suggest that REM sleep should account for about 20 to 25% of total sleep, or about 84 to 105 minutes in a typical seven-hour sleep period.

Looking at the nine sleep periods where detailed sleep data is available, the REM sleep quantities are irregular on Day 1 (0 minutes, 0% REM sleep), Day 4 (48 minutes, 10.53% REM sleep), Day 5 (68 minutes, 14.08% REM sleep), Day 8 (119 minutes, 30.51% REM sleep), Day 14 (181 minutes, 29.43% REM sleep). Participant 10's sleep quantities are also highly irregular, with sleep quantities ranging from just over 11 hours to no sleep (although possibly napping).

The averages are misleading in the case of Participant 10, as their average sleep quantity appears to be in a healthy range, however more detailed analysis reveals highly irregular sleeping patterns, and irregular sleep types. On four out of the nine nights for which

sleep data is available the REM sleep quantities are below Reite, et al.'s (2008) recommended ranges. This has obvious implications for the consolidation of memories formed during those days, and this logically has implications for learning, as well as processes like emotional regulation. On weekends, such as Days 8 and 14, Participant 10 sleeps for longer periods and receives higher quantities of REM sleep. Participant 10's stress data from the subsequent days suggests that these periods of longer sleep do not appear to restore normal functionality. It should also be noted that sleeping more on the weekends, days when there is new knowledge from classes to be consolidated, is unlikely to be helpful in learning, whether it is in language classes or in any subject.

As with Case Study 1, it is unclear whether the irregular sleep is a causal factor in the distress evidenced in Participant 10's GHQ-28 and stress measurements, an ancillary factor, or a resultant factor. What is important in the context of this study is that the sleep, stress, and GHQ-28 data suggest that in Case Study 2 the diminished language learning is not situation-specific, nor specifically related to anxiety, but may rather be related to global dysfunction. This has implications for Research Question 1.

Participant 10's activities seem to be highly irregular, so in the final section Participant 10's activity data was examined to see if a similar lack of regularity was present.

6.4.3. Case Study 2: Participant 10 Activity

Participant 10's activity data is summarised in Table 15 below.

Table 15

Activity Data: Participant 10, Days 1 to 14

Day 1 (Monday)	Day 2 (Tuesday)	Day 3 (Wednesday)
Steps: 5,843 Kilometres: 4.7	Steps: 2,753 steps Kilometres: 2.2	Steps: 4,049 Kilometres; 3.3
Day 4 (Thursday)	Day 5 (Friday)	Day 6 (Saturday)
Steps: 4,070 Kilometres: 3.3	Steps: 4,304 Kilometres: 3.4	Steps: 6,863 Kilometres: 5.5
Day 7 (Sunday)	Day 8 (Monday)	Day 9 (Tuesday)
Steps: 1,111 Kilometres: 0.9	Steps: 3,297 Kilometres: 2.6	Steps: 622 Kilometres: 0.6

Day 10 (Wednesday)	Day 11 (Thursday)	Day 12 (Friday)
Steps: 2,423 Kilometres: 2.0	Steps: 4,123 Kilometres: 3.3	Steps: 1,813 Kilometres: 1.5
Day 13 (Saturday)	Day 14 (Sunday)	
Steps: 297 Kilometres: 0.2	Steps: 4,011 Kilometres: 3.2	

Participant 10 displays lower levels of activity than Participant 12 in Case Study 1, and this may be an exacerbating factor in their high stress levels. Salmon (2001) suggests that regularity is a key factor, and with activity levels ranging from a low of 297 steps on Day 13 to a high of 6,863 steps on Day 6 (both Saturdays), there seems to be little regularity. This lack of regularity may be a factor in explaining why on Day 6, despite exercising for a considerable period Participant 10 shows no improvement in stress levels. As with sleep, the data from Participant 10 in Case Study 1 suggests that regularity may be an important factor to consider when analysing the data surrounding stress and learning.

6.5. Case Studies 3A and 3B: Participants 31 and 28 (Below Caseness, Below Average Stress)

Only two participants with accessible data from the *Garmin Vivosmart 3/4* met the criteria for both low stress and no GHQ-28 scores that exceeded the caseness threshold during the period of the study. Unfortunately, neither participants' fitness tracker data was complete.

Participant 31 reported problems with the fit of the fitness tracker, and discontinued use as a result of the device moving around too much on their wrist as the regular sized devices were too large. This issue was addressed when replacement trackers were ordered and some smaller trackers were included.

Participant 28 did not report any specific issues, neither did they withdraw from the study and they completed all the other elements of the research, including the portions after the period when the fitness tracker was worn, suggesting that the problem was probably with the fitness tracker itself, similar to the first participant.

In recognition of the fact that one participant's data might be insufficient, and in order to ensure parity with the amount of data in other case studies, the data from both participants will be presented, labelled as case studies 3A and 3B, respectively.

6.5.A. Case Study 3A: Participant 31 (Below Caseness, Below Average Stress)

Participant 31, a twenty-year-old female, was the focus of Case Study 3A. Their GHQ-28 scores are summarised in Table 16 below.

Table 16

GHQ-28 Scores for Participant 31

	Total	Sub-Scales			
		Anxiety and Insomnia	Severe Depression	Somatic Symptoms	Social Dysfunction
Week 1	12	4	0	4	4
Week 4	11	3	0	2	6
Week 6	4	1	0	2	1

Participant 31's GHQ-28 scores were well below the caseness threshold of 23 throughout the study, suggesting that there was no clinical level of distress.

Participant 31 scored 96 on the FLCAS scale. The average participant's FLCAS score was 117.7 out of a possible total of 198.

Participant 31 scored 705 on the TOEIC test on entering the university, against an average participant score of 487.06. One year later (the year of the study) Participant 31 scored 705 on the TOEIC test, against an average participant score of 641.18. In one year their TOEIC score increased by 0 points, which will be used as a reflection of their language learning during that period. The average participant's TOEIC score increased by 154.12 in the same period.

Participant 31's average stress measurements are detailed below in Table 17.

Table 17

Participant 31 Average Stress Measurements by Activity

	Participant 31's Average Stress	Number of Measurements	Average Participant's Stress
English Classes	19.25	129	40.68

Other Language Classes	N/A	N/A	40.83
Non-Language Classes	28.13	60	46.14
Free Time	25.87	1033	44.63
Part-time Job	44.86	55	47.91
Overall Average Stress	26.13		45.09

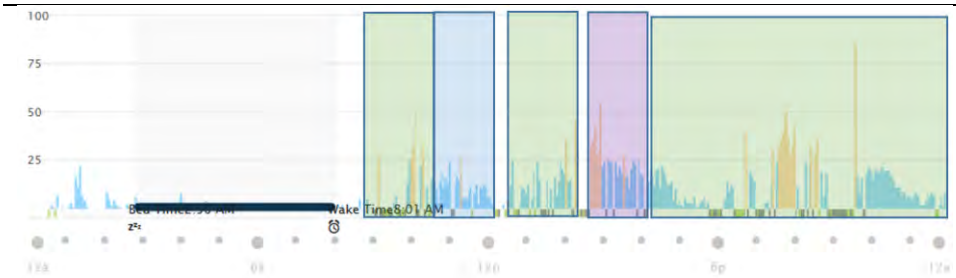
Participant 31's GHQ-28 scores and low average stress scores across most areas (with the exception of their part-time job), seem to be at odds with their lack of any progress in language learning, with zero improvement in their TOEIC score over a year of study. The detailed stress scores will be examined in the following section to see if they suggest an explanation.

6.5.A.1. Case Study 3A: Participant 31 Stress

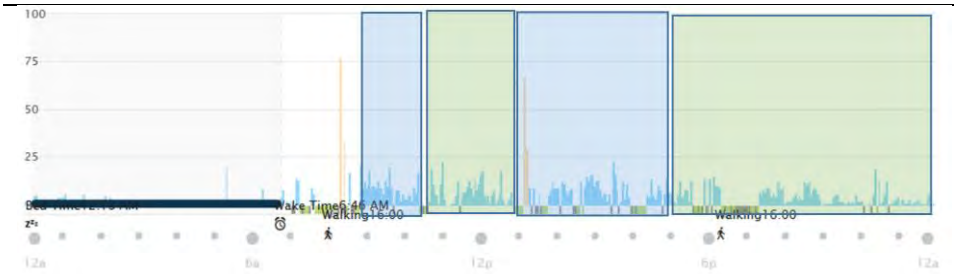
Participant 31 only provided partial stress data for the first eight days of the study. Despite this, there are 129 stress measurements from English language classes, 60 stress measurements from non-language classes, 1,033 stress measurements from free time, and 55 stress measurements from their part-time job. Further, studies referenced in the literature (Pakhomov, et al., 2020; Akbar et al., 2021) suggested that a week of data had proven to be sufficient for stress research in the past. Unfortunately, the data were incomplete in other respects because of the device fitting poorly. While 129 stress measurements were taken in English language classes these are from only four 90-minute classes, on two different days. This was part of the reason for including Case Study 3B.

Table 18

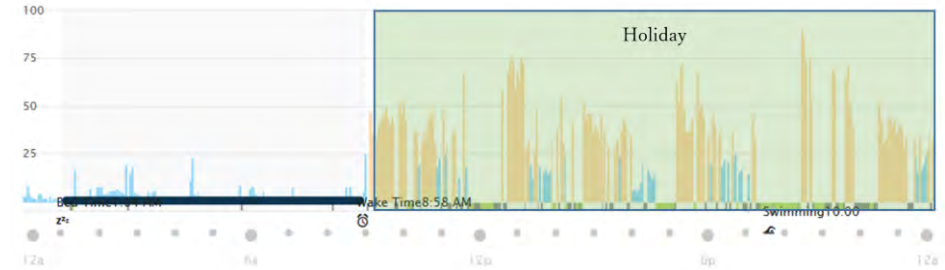
Stress Measurements: Participant 31, Days 1 to 8

Day 1 (Monday)


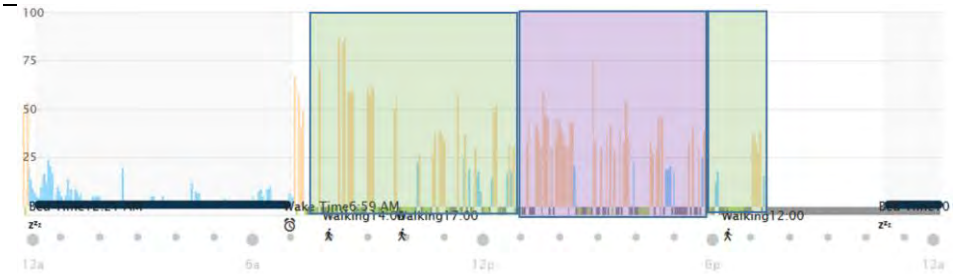
Day 2 (Tuesday)



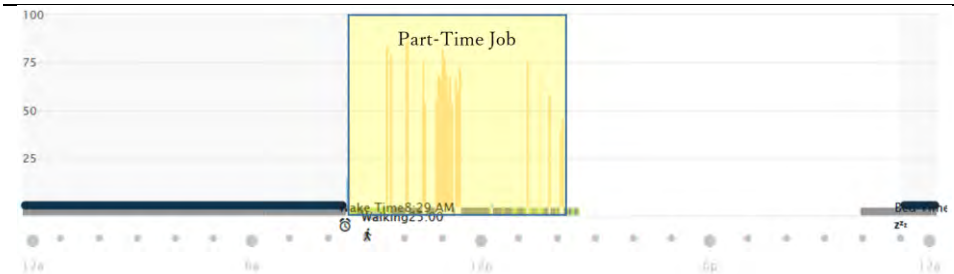
Day 3 (Wednesday)



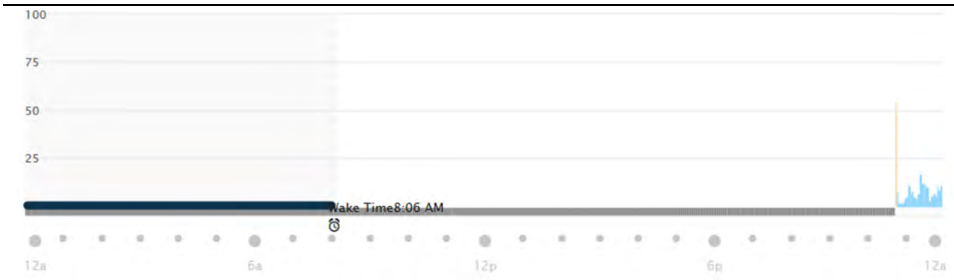
Day 4 (Thursday)



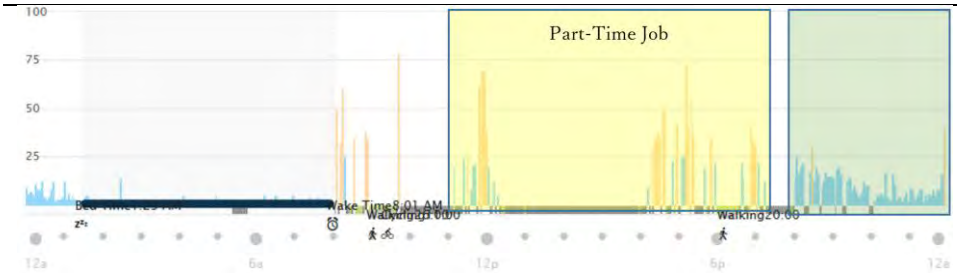
Day 5 (Friday)



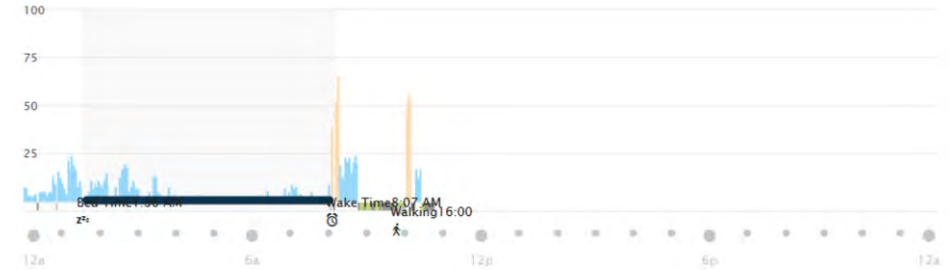
Day 6 (Saturday)



Day 7 (Sunday)



Day 8 (Monday)



Note. Data in Table 18 will be referred to by day. For example, “on Day 8”, rather than “in Table 18, Day 8”.

Looking at the difference in stress measurements between Day 4 and Day 2 demonstrates the degree of variability that the same individual can display just two days apart. This is a major limitation when considering the data from Participant 31, but also has a bearing on the importance of contextualised data. While some insights can be gleaned from the data available, this case study has a particular bearing on Research Question 5 regarding the use of wearable devices in research.

In the previous studies cited where one week of data was gathered, the phenomenon under investigation was constant. In Akbar, et al. (2021) the administrative work was performed every day. Similarly, in Pakhomov, et al. (2020) the study was concerned with stress in the seven days before examinations, and the devices were worn during this period. In this study the data was episodic as there were not English language classes every day.

This is important when one compares the English language classes on Day 2, against the non-language class on Day 4. On Day 2 the contextualised stress data from the whole day shows generally resting (0 to 25) stress levels, with only two relatively brief periods of higher stress. These may reflect moments of transient anxiety. However, on Day 4 Participant 31 is clearly experiencing a more difficult day, with elevated stress levels across all contexts and relatively few moments of resting stress (0 to 25). Comparing the non-language class data from Day 4 against the English language class data from Day 2 and concluding that non-language classes were more stressful would be an obviously invalid conclusion. Likewise, average data that included Day 4's non-language classes would skew the average higher.

In this case study the comparison of Participant 31's stress in language versus non-language classes must be restricted to Day 1. However, this comes with the caveat that in previous case studies it was demonstrated that stress levels may vary from class to class in the same type of class. What Case Study 3A suggests is that rather than simply counting days, there should also be some consideration of how many incidents of a phenomenon are being captured. In the case of this study that would be the number of each type of classes. In Case Study 1 data were gathered on 16 language classes, 10 non-language classes, and 13 days of data about stress during free time. In Case Study 2 data were gathered about 9 language classes, 15 non-language classes, and 13 days of free time data. The quantity of data in Case Studies 1 and 2 appears to be sufficient for meaningful analysis. Case Study 3A consists of four language classes, three non-language classes, and six days of free time data, and may be insufficient for the detection of broader trends in stress patterns. It should also be noted that three of the language classes fall on one day and should possibly be counted only once. What this suggests for further research is that the quantity of data required means considering not just the number of days or number of incidents, but a combination of the two.

Despite these limitations, Participant 31 does seem to display extremely low stress levels in English language classes. Day 1 provides a reasonable basis for comparing Participant 31's language versus non-language stress levels. If Day 1 is a reliable indication of their normal stress patterns, then their stress level in English classes seems to be markedly lower than their stress levels in non-language classes. Participant 31's stress levels in English language classes is typified by resting (0 to 25) levels of stress.

A possible explanation for this comes from considering Participant 31's TOEIC score of 705 and Csíkszentmihályi's (1997) flow model, shown in Figure 3 in Chapter 2.

A TOEIC score of 705 puts Participant 31's level of English proficiency 63.82 points above the average participant's TOEIC score of 641.18. Perhaps more importantly, Participant 31 started university with a TOEIC score of 705 over an average score of 487.06. Initially there would have been a large mismatch between the level of class content aimed at the average student and Participant 31's level of English proficiency. Csíkszentmihályi's (1997) flow theory shown in Figure 3 predicts that where the individual's skill level is high, but the level of challenge is comparatively low, then the individual will be relaxed.

The Hebbian (1955) model of stress and performance suggests that this relaxation may not be conducive to performance and learning. Participant 31's seeming lack of improvement in their TOEIC score suggests that Hebb's (1955) model may be a valid way of examining stress and language learning.

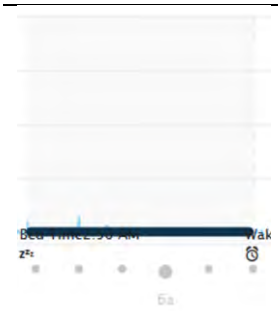
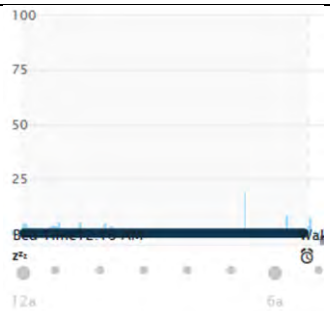
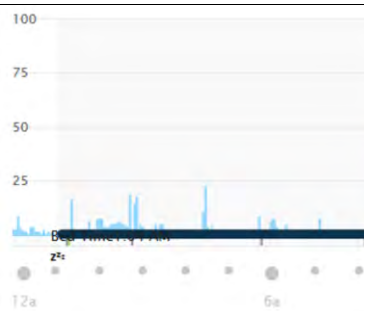
Looking at the stress data there seems to be a great deal of variation between Days 1 and 2, where Participant 31 shows very low stress levels, and Days 3 and 4 where stress levels are significantly higher. While Participant 31 shows higher stress levels on several days it is important to note that there are numerous balancing periods of resting stress, unlike the consistently high stress seen in Case Study 2. However, this variation is unexplained, so the next section will examine Participant 31's sleep data to see if perhaps there is some clue to the reason for this variation.

6.5.A.2. Case Study 3A: Participant 31 Sleep

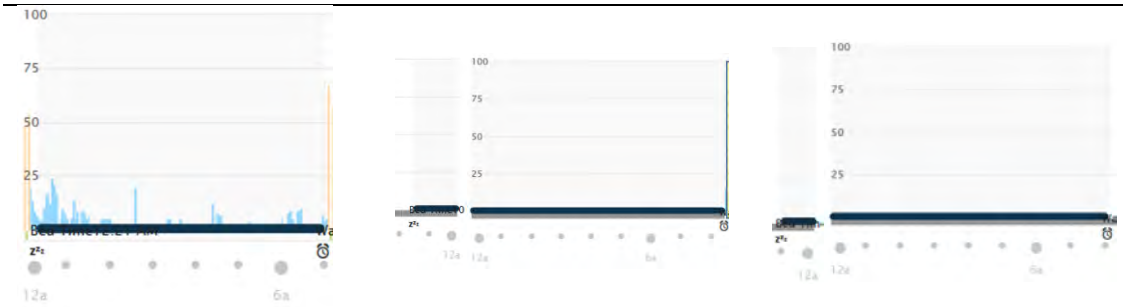
Participant 31's sleep data for eight days is presented in Table 19 below.

Table 19

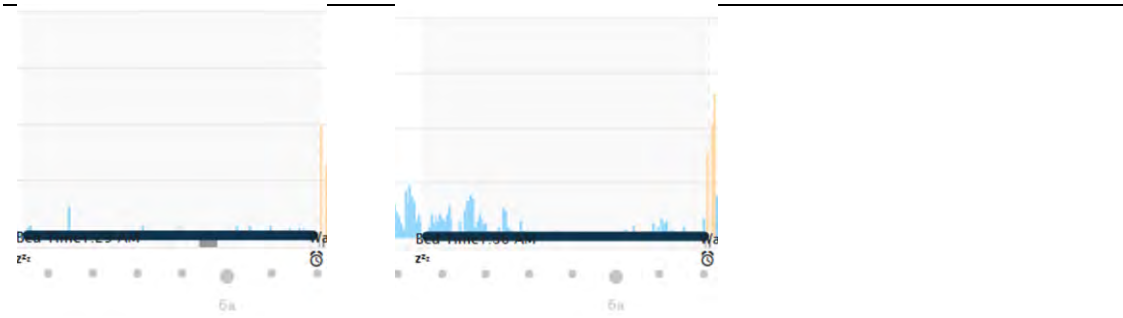
Sleep Data: Participant 31, Days 1 to 8

Day 1 (Monday)	Day 2 (Tuesday)	Day 3 (Wednesday)
		

Day 4 (Thursday) | Day 5 (Friday) | Day 6 (Saturday)



Day 7 (Sunday) | Day 8 (Monday)



Participant 31's sleep data indicates irregular sleeping times, with sleep starting from between 22h30 and 03h00, however waking times were fairly regular, within an hour

variation around 08h00 (07h00 to 09h00), and Participant 31 woke at roughly 08h00 on five of the eight days for which detailed data is available.

Participant 31 averaged approximately seven and a half hours of sleep a night, although again the issue of regular sleep patterns becomes important. On most weeknights Participant 31 received approximately six and a half hours of sleep (with a low of five hours and eleven minutes, and a high of seven hours and forty-five minutes). However, on the weekends Participant 31 slept for 9 and 10 hours, respectively. This tends to suggest that Participant 31's weekday sleep was insufficient, and that weekend sleep was being used to repay this sleep debt. Despite issues with sleep regularity, Participant 31's sleep quantity deficit appears mild.

Sleep quality appears to be good, with stress readings within the low resting range, and only one night with any significant movement (Day 7 around 05h30 to 05h50). The only day when Participant 31's sleep quantity is remarkably low is on Day 1, however there seems to be no accompanying increase in stress the following day.

Next, Participant 31's activity data will be reviewed to see if there are any clues there for the variation.

6.5.A.3. Case Study 3A: Participant 31 Activity

Participant 31's activity data is summarised in Table 20 below.

Table 20

Activity Data: Participant 31, Days 1 to 8

Day 1 (Monday)*	Day 2 (Tuesday)*	Day 3 (Wednesday)*
Steps: 6,355 Kilometres: 4.6	Steps: 7,986 Kilometres: 5.7	Steps: 10,361 Kilometres: 8.4
Day 4 (Thursday)*	**Day 5 (Friday)**	**Day 6 (Saturday)**
Steps: 10,188 Kilometres: 7.2	Steps: 6,166 Kilometres: 4.3	Steps: 0 Kilometres: 0
Day 7 (Sunday)*	**Day 8 (Monday)**	
Steps: 9,559 Kilometres: 6.7	Steps: 2,745 Kilometres: 1.9	

Participant 31's activity levels appear to be quite erratic, but this is a result of the inconsistent data from the wearable device due to poor fit. Those days when the fitness tracker was worn properly for almost the whole day have been marked with an asterix (*). Considering these five days, Participant 31 walked an average of 7,745 steps a day.

As with the sleep data, Participant 31's activity data suggests no underlying cause for the variations in stress levels. There is likewise no note in Participant 31's diary about the reason for the increased stress seen on some days.

While the lack of an answer to the variation in stress levels may feel unsatisfying, the sleep and activity data for Participant 31 are commensurate with what one would expect to see for a healthy individual experiencing low levels of stress. The answer to the variation may be as simple as Participant 31 having been thrown off their normal routine by the public holiday on Day 3.

6.5.B. Case Study 3B: Participant 28 (Below Caseness, Below Average Stress)

Participant 28, a nineteen-year-old female, was the focus of Case Study 3B. Their GHQ-28 scores are summarised in Table 21 below.

Table 21

GHQ-28 Scores for Participant 28

	Total	Sub-Scales			
		Anxiety and Insomnia	Severe Depression	Somatic Symptoms	Social Dysfunction
Week 1	14	3	8	0	3
Week 4	15	4	4	5	1
Week 6	8	0	3	3	2

Participant 28's average stress measurements are detailed in Table 22.

Table 22

Participant 28 Average Stress Measurements by Activity

	Participant 28's Average Stress	Number of Measurements	Average Participant's Stress
English Classes	18.05	56	40.68
Other Language Classes	12.97	59	40.83
Non-Language Classes	14.07	67	46.14
Free Time	30.02	738	44.63
Part-time Job	50.33	62	47.91
Overall Average Stress	28.64		45.09

Participant 28's GHQ-28 scores were well below the caseness threshold of 23 throughout the study, suggesting that there was no clinical level of distress.

Participant 28 scored 93 on the FLCAS scale. The average participant's FLCAS score was 117.7 out of a possible total of 198.

Participant 28 scored 600 on the TOEIC test on entering the university, against an average participant score of 487.06. One year later (the year of the study) Participant 28 scored 695 on the TOEIC test, against an average participant score of 641.18. In one year their TOEIC score increased by 95 points, which will be used as a reflection of their language learning during that period. The average participant's TOEIC score increased by 154.12 in the same period.

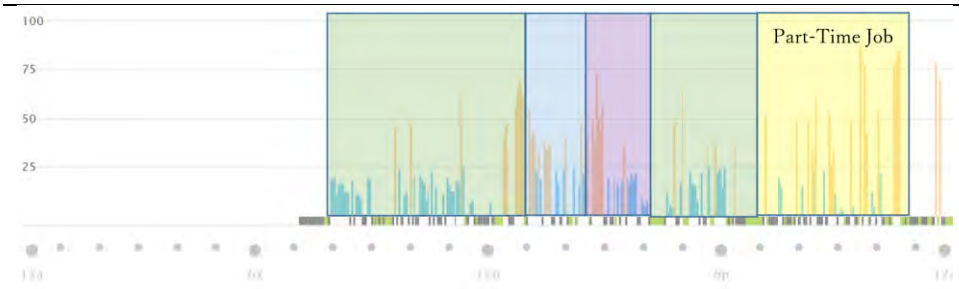
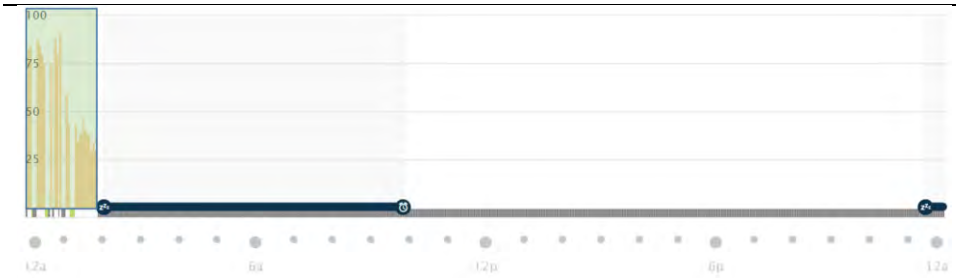
6.5.B.1. Case Study 3B: Participant 28 Stress

As in Case Study 3A the participant in this case study only provided partial stress data. In Case Study 3B no specific difficulty was reported. There is complete data for Day 5 and Days 7 to 11. Data is completely absent for Days 1 to 4, with the device appearing not to have been worn at all. On Day 6 there is a large gap in the data from 10h00 to approximately 23h30. On Day 12 to 14 only sleep data is present, but no stress readings. The erratic pattern is somewhat puzzling and one possible reason is that there may have been a malfunction of some sort in one of the *Garmin Vivosmart 3/4's* sensors. A simpler explanation might be that a loose fit meant that the sensor that monitors pulse and HRV was not in sufficiently close proximity to the participant's skin to get clear readings. Variations in

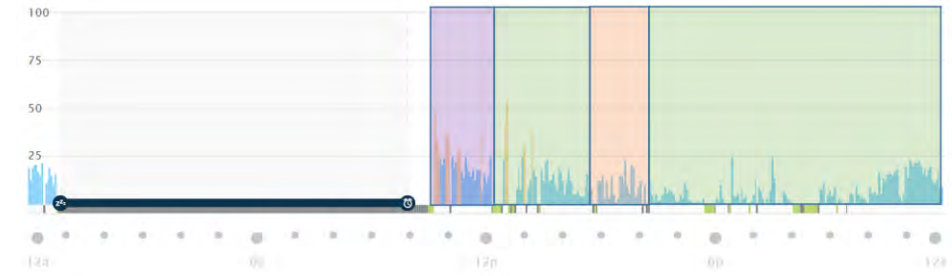
the participant's positioning might have erratically shifted the sensor close enough to take readings. Similarly, variations in how tightly the participant fastened the wristband might have made the difference between the sensor being able to take readings or being unable to take readings that day.

Table 23

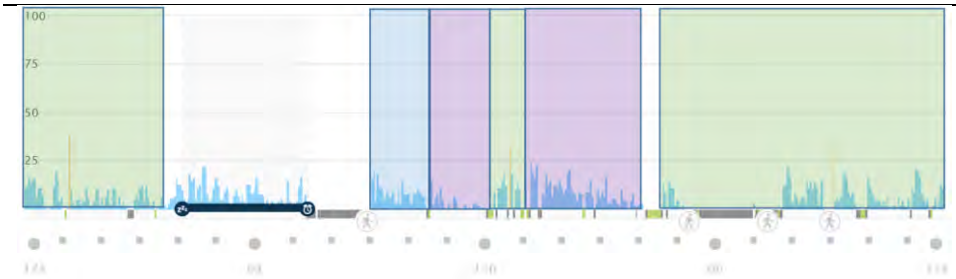
Stress Measurements: Participant 28, Days 5 to 12

Day 5 (Friday)

Day 6 (Saturday)

Day 7 (Sunday)

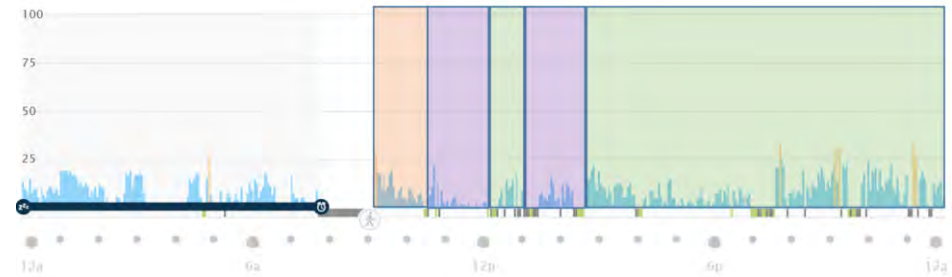

Day 8 (Monday)



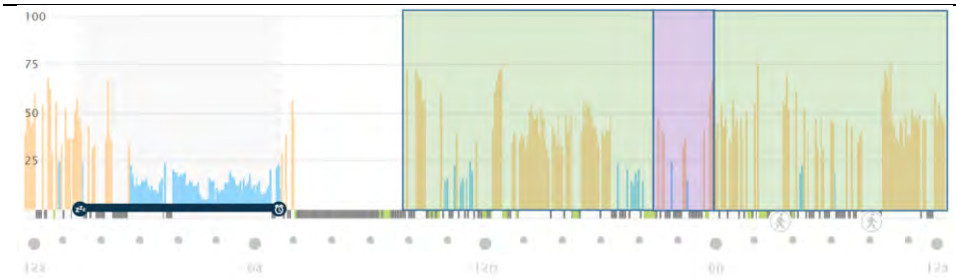
Day 9 (Tuesday)



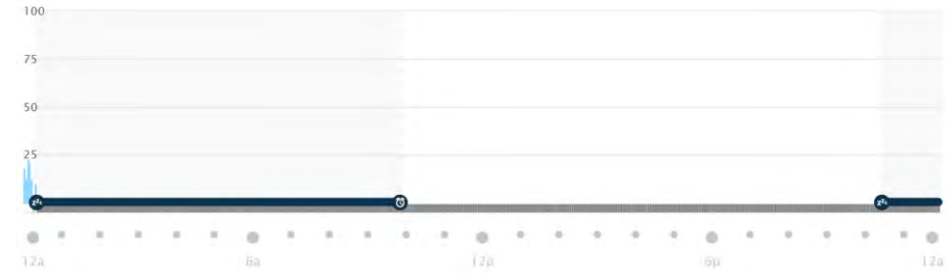
Day 10 (Wednesday)



Day 11 (Thursday)



Day 12 (Friday)



Note. Data in Table 23 will be referred to by day. For example, “on Day 8”, rather than “in Table 23, Day 8”.

Participant 28’s average stress level was below average, but this fails to capture the variation between days. To broadly group the days into three categories, on Days 6, 7 and 11, Participant 28’s stress levels tended towards moderate stress (51 to 75). On Days 8, 9 and 10, Participant 28 tended towards resting stress levels (1 to 25). On Day 5 stress levels were a mixture of resting with stress spiking into the moderate (51 to 75) range.

This variation in daily stress makes comparisons between days difficult, however Participant 28’s schedule fortunately permits clear comparisons between subjects on the same days. The stress patterns are very similar between classes, and while stress levels appear to be a little lower during language classes this depends on the day, and the difference is not marked.

This is another area where the averages are misleading, because one of the two English language classes falls on Day 5, which was a more stressful day than normal for Participant 28. However, there were seven non-language classes, four of which fell on very low stress days. As a result, the average English language class stress levels are higher than the non-language classes, even though comparisons of English class stress and non-language class stress suggest this is not an accurate representation of the data.

Participant 28 also studied Chinese as another language and shows similar patterns of low stress levels during Chinese classes, suggesting that their lack of stress and anxiety is not limited to English language classes.

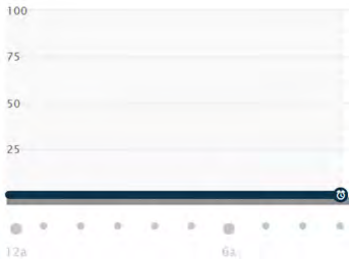
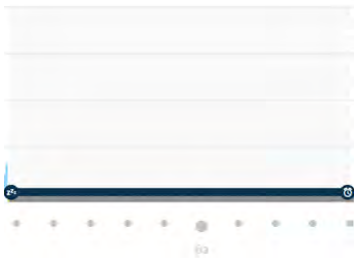
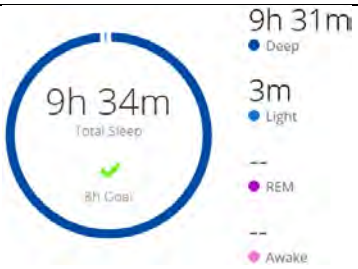
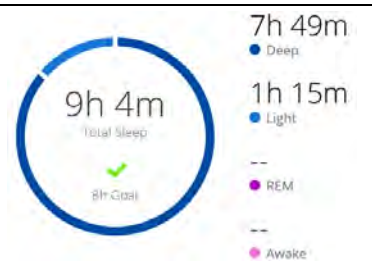
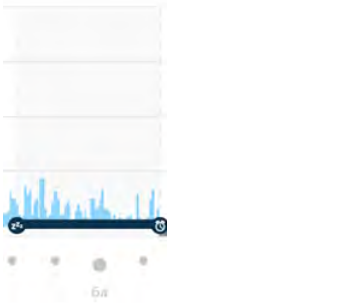
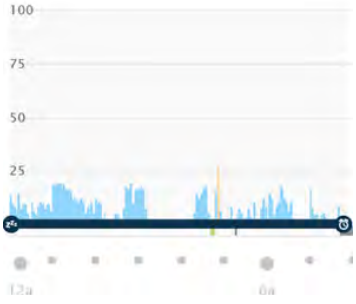
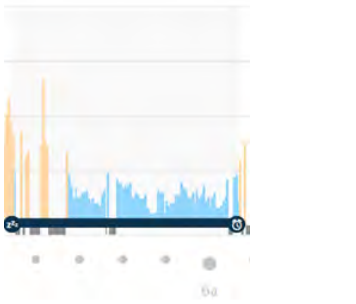
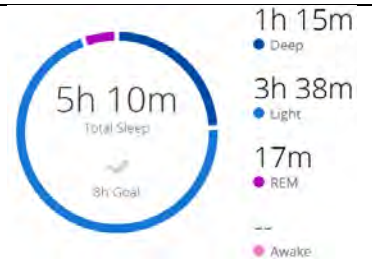
Similar to Participant 31, but less extreme, Participant 28’s starting TOEIC score was 600 against an average score of 487.06, which may have contributed to their resting stress levels in English language classes. However, in this case the sleep data may suggest an alternative hypothesis.

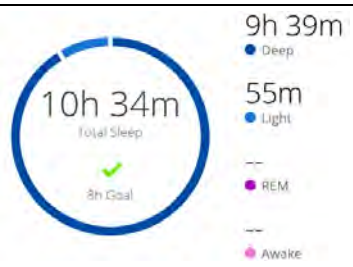
6.5.B.2. Case Study 3B: Participant 28 Sleep

Participant 28's sleep data for eight days is presented in Table 24 below.

Table 24

Sleep Data: Participant 28, Days 1 to 8

Day 6 (Saturday)	Day 7 (Sunday)	Day 8 (Monday)
		
		
Day 9 (Tuesday)	**Day 10 (Wednesday)**	**Day 11 (Thursday)**
		
		

Day 12 (Friday)	Day 13 (Saturday)	Day 14 (Sunday)
	No stress data	No stress data
		

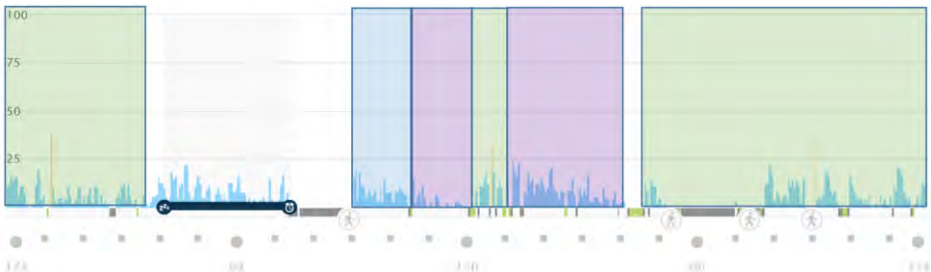
On Days 6, 7, 8, 12, 13, and 14 the fitness tracker identified sleep duration by lack of significant movement for a protracted period, and classified types of sleep in a similar manner, but was unable to record precise stress levels for some reason

Participant 28's average sleep quantity over the nine days for which there is sleep data is approximately 8 hours a day, which appears to be sufficient, however again when one considers the detailed data it becomes clear that Participant 28's sleep patterns are highly erratic, varying from three hours and fourteen minutes on Day 9 to ten hours and fifty-four minutes on Day 14. There is the familiar pattern, as in the other case studies, of lower sleep quantities on weeknights, followed by much higher sleep quantities on weekends as the individual attempts to make up for lost weekday sleep.

What is interesting is that when one cross-references the sleep data to the stress data, Participant 28's stress levels do not appear adversely affected by the single night of low sleep quantity on Day 9, as shown in Figure 16 below.

Figure 16

Participant 28 Day Nine (Wednesday) Detailed Stress Data



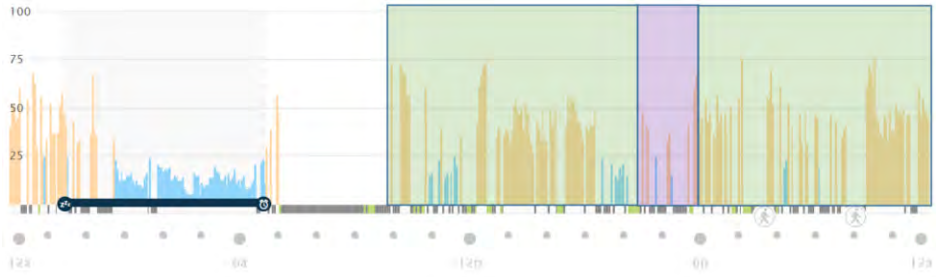
As has been noted before, fitness trackers are generally poor at flagging daytime napping. It is possible that Participant 28 compensated for their low sleep quantity on Day 9 by napping during the day. Looking at their stress data these would have to have been very short naps as Participant 28 was quite busy, and there do not appear to be long periods of inactivity, apart from in class. As a cultural note, in Japan it is not unusual for students to fall asleep in class. While this behaviour is less prevalent at a university level than it is at a senior high school level, it is still a possibility that the resting stress levels and lack of movement seen between 09h00 and 12h10 might represent periods of sleep. Students sleeping in class will generally not be disturbed, as the prevailing attitude is that if they need the rest so much that they have fallen asleep in class, then it would be both unkind and counter-productive to wake them. So long as they are not disrupting the class they will be permitted to sleep. There is therefore the possibility that the stress levels in these classes are misleading and do not reflect low stress because of the content, but rather because the individual was asleep.

The current inability of fitness trackers to correctly flag daytime napping is a potential area of concern for researchers using this technology to monitor both sleep and stress. There is the possibility that this individual did indeed nap for both classes, waking briefly at the end of class, which would neatly coincide with a natural waking point, being at around the 90-minute mark. This would change their sleep quantity for Day 9 from three hours and fourteen minutes to a number that might be closer to six hours of sleep. While the anatomy of napping is different from continuous sleep, being lower in REM sleep and higher in NREM sleep, this potential doubling of sleep quantity might significantly impact analysis.

As noted earlier, the single day of low sleep quantity on Day 9 did not appear to raise stress levels on Day 9, however by late on Day 10 one begins to see increasing stress levels towards the evening, culminating in unusually high stress levels during sleep on Day 11 and throughout the day.

Figure 17

Participant 28 Stress Data Day 11



The sleep data for Day 11 is irregular in several respects. The first is the total sleep quantity, which is 5 hours and ten minutes. There is only seventeen minutes of REM sleep. The sleep is also clearly disturbed, with the grey markers at the bottom of the stress graph showing at least three periods of movement that was sufficiently significant for the fitness band not to collect stress data during these periods. The next irregularity is that for nearly the first two hours of the sleep period, stress levels were noticeably higher than resting stress levels, spiking into the low and moderate stress ranges. This is quite unusual for sleep and were it not for the GHQ-28 results, one might suspect anxiety based on this sleep data.

Case Study 3B again illustrates that most common descriptive statistics, such as mean, median or average values would have failed to adequately describe this individual's sleep data. The average sleep quantity was a little over eight hours, the median sleep quantity was nine hours, and the modal sleep quantity was nine and a half hours. None of these numbers capture the irregularities in this individual's sleep data. While this data is represented in numbers on graphs, there is a strong qualitative element to the interpretation of each participant's data that cannot be ignored with risking a fundamentally flawed presentation of the data.

It also raises the issue of how sleep loss may have a cumulative, rather than immediate impact on stress. This consideration of cumulative sleep debt and its impact on stress is a potential area for future research and may partially explain some of the stress patterns seen in earlier case studies.

The final dataset to be examined in Case Study 3B is the Participant 28's activity data.

6.5.B.3. Case Study 3B: Participant 28 Activity

Participant 28's activity data is summarised in Table 25 below.

Table 25

Activity Data: Participant 28, Days 5 to 12

Day 5 (Friday)	Day 6 (Saturday) ▲	Day 7 (Sunday)
Steps: 10,158 Kilometres: 7.0	Steps: 1,180 Kilometres: 0.8	Steps: 14,362 Kilometres: 10
Day 8 (Monday)	Day 9 (Tuesday)	Day 10 (Wednesday)
Steps: 6,162	Steps: 10,570	Steps: 6,405

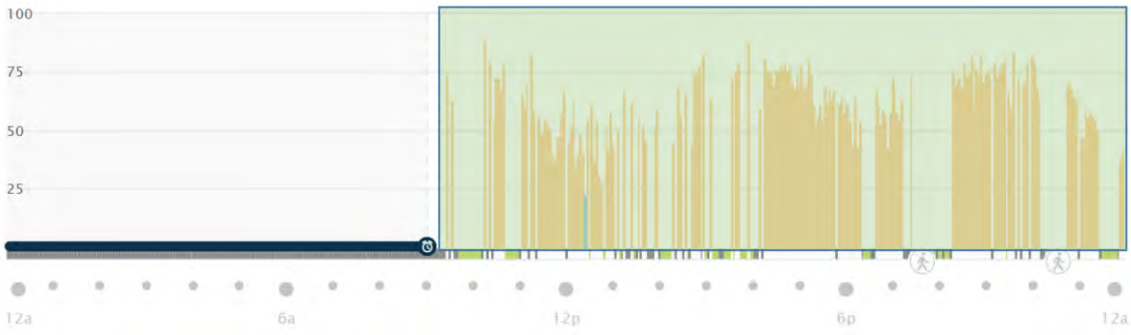
Kilometres: 4.3	Kilometres: 7.3	Kilometres: 4.4
Day 11 (Thursday)	Day 12 (Friday) ▲	
Steps: 7,802 Kilometres: 5.4	Steps: 36 Kilometres: 0.0	

Note. Days where the data is incomplete have been marked with ▲

On the days where complete data is available, Participant 28 averaged 9,243 steps a day, ranging from 6,162 steps on Day 8 to 14,362 steps on day 7. Day 7 is also the day that Participant 28 experienced their highest recorded stress levels, and it is unclear whether stress levels were higher because of the level of physical activity, or whether Participant 28 engaged in higher levels of physical activity as a coping mechanism because they were feeling stressed.

Figure 18

Participant 28's Stress Data from Day 7



Looking at the pattern of movement in Figure 18 above, marked on the stress graph by the lines where no stress readings were possible, it appears that the 14,362 steps were not continuous. There are three periods of extended movement, from about 09h30 to about 10h30, then from shortly after 19h00 to shortly after 20h00 and then again from shortly after 22h00 to just before 23h00.

However, the majority of the movement during the day appears to occur in short periods of movement followed by periods of non-movement. This is apparent because the fitness tracker will not take stress readings during movement as movement interferes with the measurement of inter-beat variance, which is the basis for measuring stress.

This pattern of movement tends to mitigate against the interpretation that Participant 28 felt stressed and then decided to take a walk. The data suggests that Participant 28 was busy during the day doing things like shopping and attending to chores, and other activities

that would fit the pattern of standing up and moving around for a while, and then being able to sit down between periods of activity.

6.6. Case Study 4: Participant 30 (Below Caseness, Above Average Stress)

Participant 30, a nineteen-year-old female, was the focus of Case Study 3B. Their GHQ-28 scores are summarised in Table 26 below.

Table 26

GHQ-28 Scores for Participant 30

	Total	Sub-Scales			
		Anxiety and Insomnia	Severe Depression	Somatic Symptoms	Social Dysfunction
Week 1	18	5	2	5	6
Week 4	16	5	1	4	6
Week 6	8	2	0	2	4

Participant 30's average stress measurements are detailed below in Table 27.

Table 27

Participant 30 Average Stress Measurements by Activity

	Participant 30's Average Stress	Number of Measurements	Average Participant's Stress
English Classes	63.31	178	40.68
Other Language Classes	N/A	N/A	40.83
Non-Language Classes	62.61	313	46.14
Free Time	66.32	925	44.63
Part-time Job	73.88	360	47.91
Overall Average Stress			45.09

Participant 30's GHQ-28 scores were below the caseness threshold of 23 throughout the study, suggesting that there was no clinical level of distress. However, Participant 30's week 1 score of 18 was concerning.

Participant 30 scored 122 on the FLCAS scale. The average participant's FLCAS score was 117.7 out of a possible total of 198.

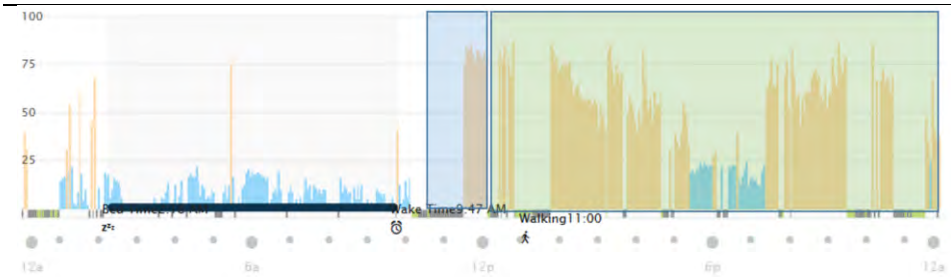
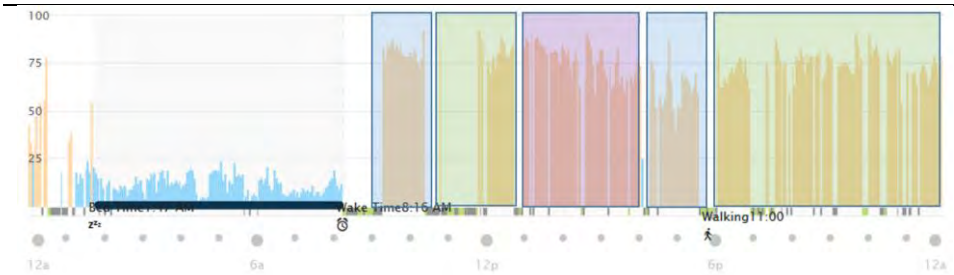
Participant 30 scored 520 on the TOEIC test on entering the university, against an average participant score of 487.06. One year later (the year of the study) Participant 28 scored 520 on the TOEIC test, against an average participant score of 641.18. In one year their TOEIC score increased by 0 points, which will be used as a reflection of their language learning during that period. The average participant's TOEIC score increased by 154.12 in the same period.

6.6.1. Case Study 4: Participant 30 Stress

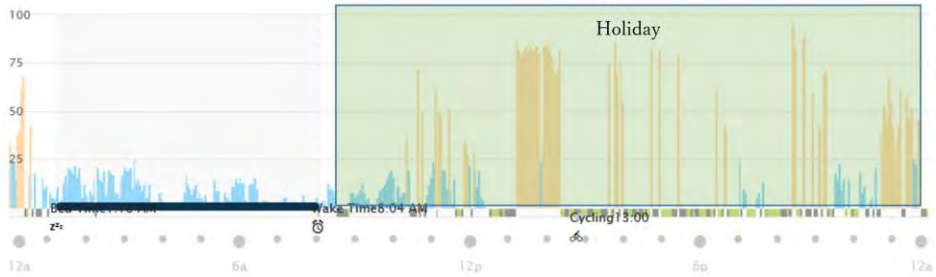
Table 28 below will present Participant 30's detailed stress scores.

Table 28

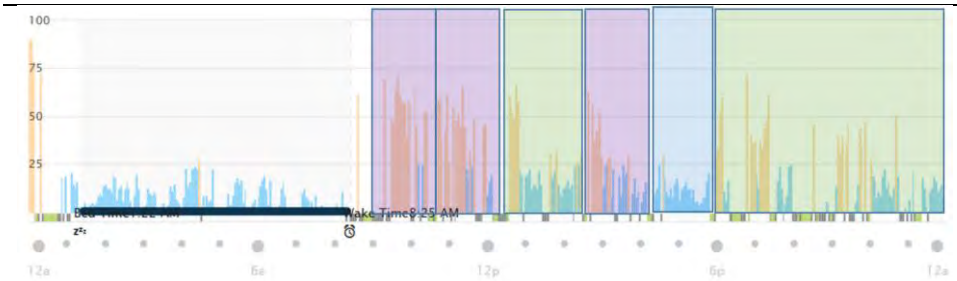
Stress Measurements: Participant 30, Days 1 to 14

Day 1 (Monday)

Day 2 (Tuesday)


Day 3 (Wednesday)



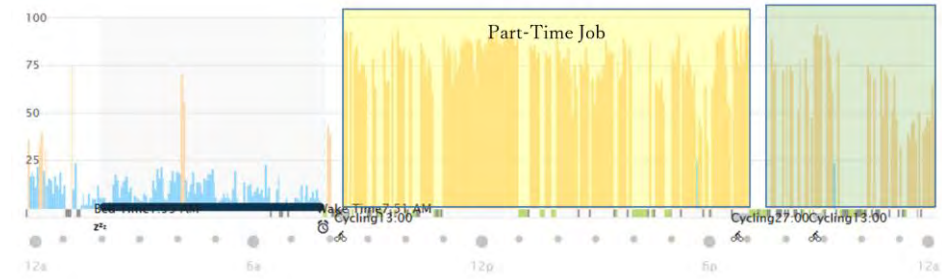
Day 4 (Thursday)



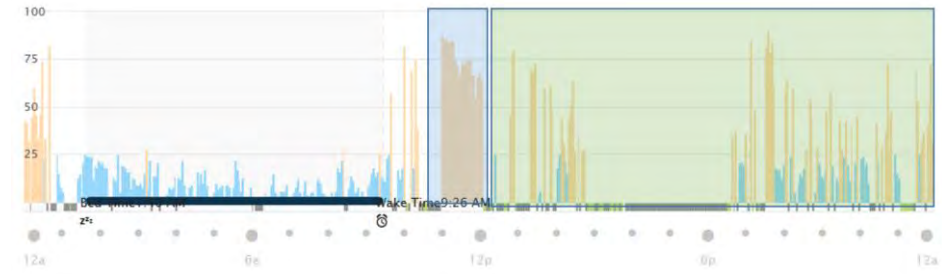
Day 5 (Friday)

Day 6 (Saturday)

Day 7 (Sunday)

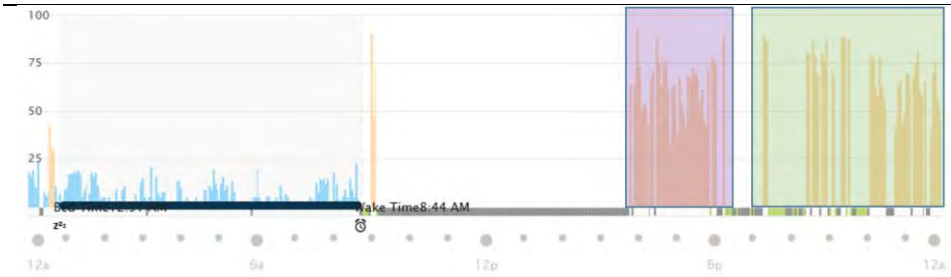


Day 8 (Monday)

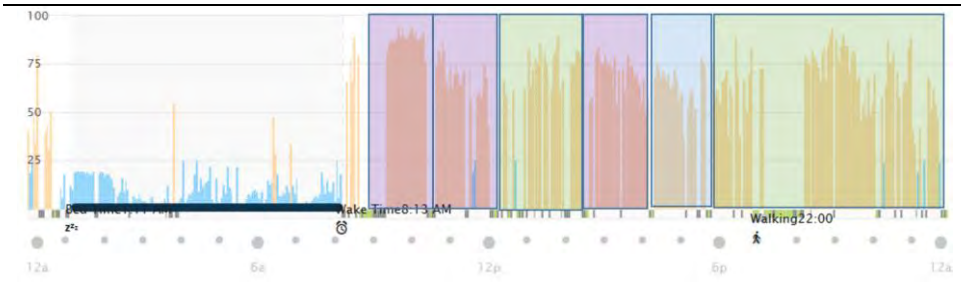


Day 9 (Tuesday)

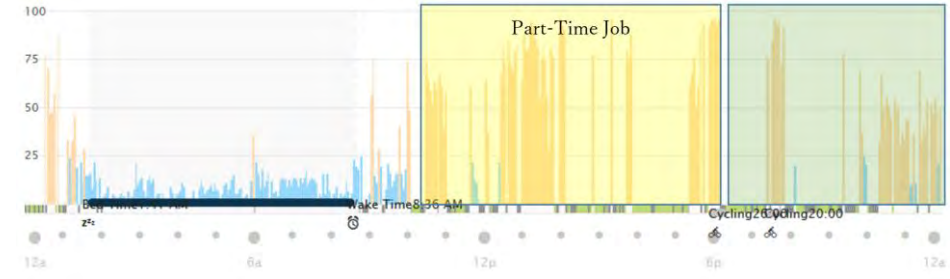
Day 10 (Wednesday)



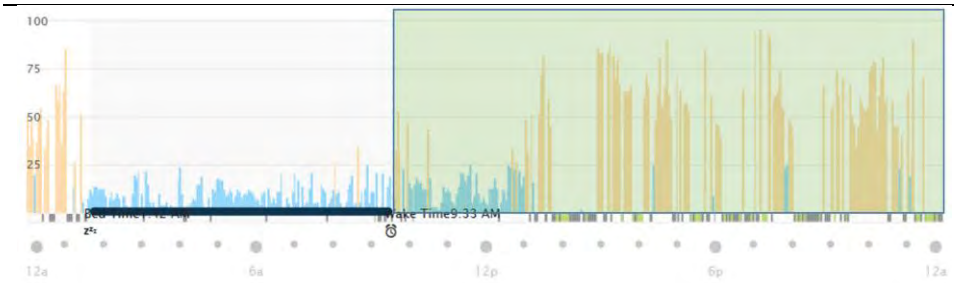
Day 11 (Thursday)



Day 12 (Friday)



Day 13 (Saturday)



Day 14 (Sunday)

Note. Data in Table 28 will be referred to by day. For example, “on Day 8”, rather than “in Table 28, Day 8”.

Participant 30’s stress levels were in the moderate (51 to 75) to high (76 to 100) range for most of the two-week period of the study. This sort of consistent elevated stress is worrying,

and while the GHQ-28 results seemed to indicate no clinically significant level of concern, it should be borne in mind that the test is reported to be 88% accurate.

Participants in the study were reminded of the presence of a student counselling office that offered free consultations, but due to the single blind nature of the study and, as an unavoidable result of the ethical boundaries placed upon the research, the researcher was unable to link participant data to individuals. Therefore, this did not allow any further intervention unless the participant in question choose to contact the researcher with a direct query and identify themselves and their participant number. It is hoped that the participant saw their constantly elevated stress levels and sought help, however there is no way to be sure.

Participant 30's constantly elevated stress levels make it difficult to distinguish between stress in various contexts, with periods of relaxation seemingly occurring randomly without perceivable links to context. On Day 4 Participant 30 experiences a rare period of resting levels of stress during an English class, but also shows sudden dips from moderate stress (51 to 75) down into resting stress (1 to 25) earlier in the day during two non-language classes. It has previously been noted that this illustrates a potential weakness with earlier stress studies and episodic data gathering.

The stress patterns seen in Case Study 4 are remarkably similar to those seen in Case Studies 1 and 2, where clinical distress was present. It is important to note here that while the DSM-5 has moved stress-induced conditions into a separate category related to the genesis of the condition. This is not the same as delinking stress as a symptom arising from these conditions, and in the words of Sapolsky (2015), stress has a "… ubiquitous, but non-specific, role … in psychiatric disorders" (p. 1347). It may be that in this specific case the GHQ-28 results are incorrect, or illustrate the 'fake good' risk in self-report questionnaires. It may also be that the condition was in the prodromal phase, and that it fell below clinical levels at the time of the study. Alternatively, this might merely indicate a particularly stressful period in Participant 30's life, such as the death of a grandparent or other major life event, and this was normal and non-clinical stress.

While the unusual pattern in the English class on Day 4 did shift the average stress in English classes down by nearly a full percentage point, it was not sufficient to skew the dataset because of the large number of measurements across the two weeks. Large datasets are not necessarily automatically superior to small datasets. However, this study has repeatedly demonstrated that when discussing the topic of stress, the degree of variability both within and between days argues in favour of a larger dataset. More frequent

measurements over an extended period capture the variability seen in stress levels, which episodic measurements might miss.

With regards to LLA theory, Participant 30's stress levels were not markedly higher in non-language classes than in language classes, with mean scores of 62.61 and 63.31 respectively, against an overall average stress level of 66.90. This suggests that language learning is not significantly more stressful than other learning or everyday stress.

Participant 30's starting TOEIC score of 520 was a little higher than the average starting score of 487. This difference was not so high that one would expect to see them feeling complacent due to their slightly higher level of ability. Nor was it so low that it could be argued that they would be out of their depth in English classes when their classmates were not.

Over a year Participant 30 showed no (zero) improvement in their TOEIC score, as compared to an average improvement by all participants of 154 points. This would tend to indicate that high stress levels may have a negative impact on language learning. The average participant, who had moderate stress levels of about 40, showed the greatest improvement, with an average improvement of about 150 points. Data from Case Study 3A and 3B suggests that resting levels of stress seemed to be associated with a lack of improvement (in 3A), or lower levels of improvement than average (in 3B). This seems to describe the inverted U shape predicted by theories such as Yerkes-Dodson's (1908) stress curve, where at extremes of the stress scale (either high or low stress), performance is impaired. This would suggest that there is some legitimacy to LLA theory, but that the simple straight-line relationship where high stress is bad and low stress is good is an over-simplification of a more complex phenomenon.

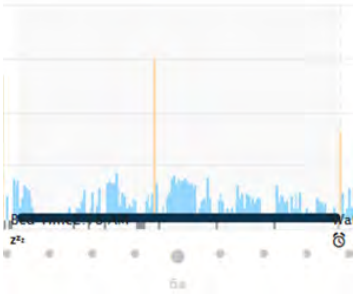
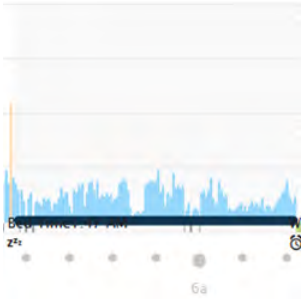
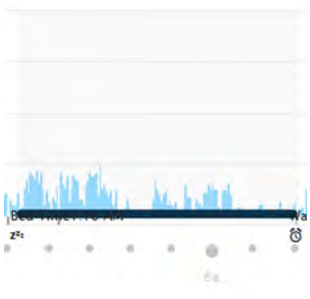
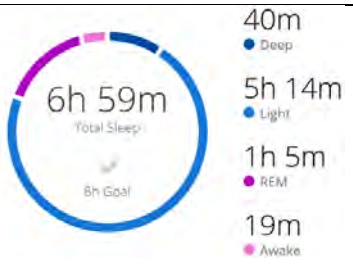
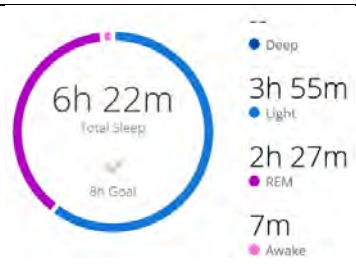
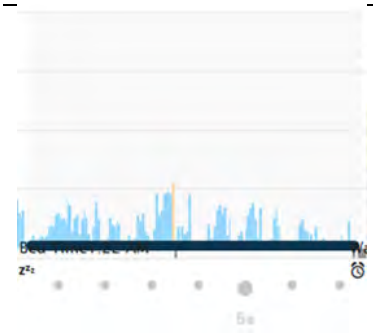
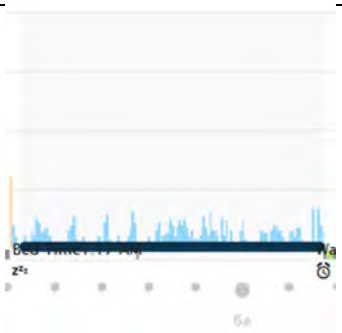
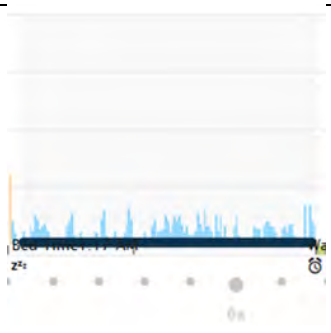
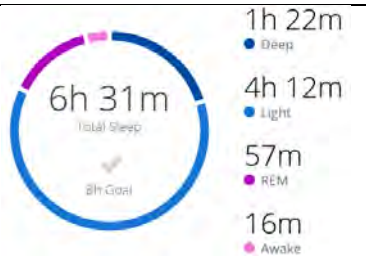
Participant 30's FLCAS score did not predict the apparent limitations in their language learning, with Participant 30 scoring 122 on the FLCAS against an average score for all participants of 117.7. This is not a sufficiently large difference to explain the large difference in performance between Participant 30 and the other participants in this study. Nor did the slightly higher FLCAS predict Participant 30's elevated stress levels in class and in general.

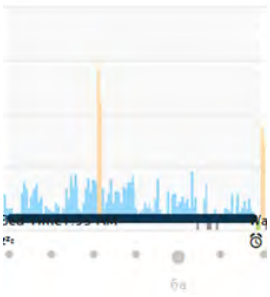
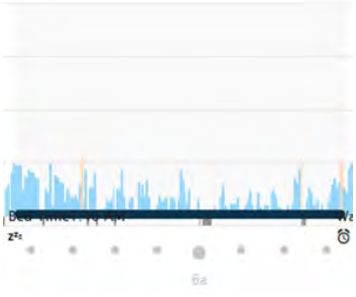
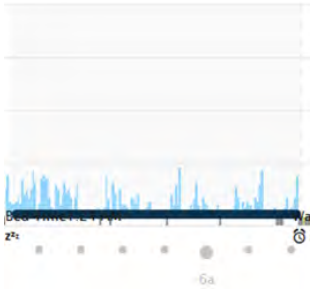
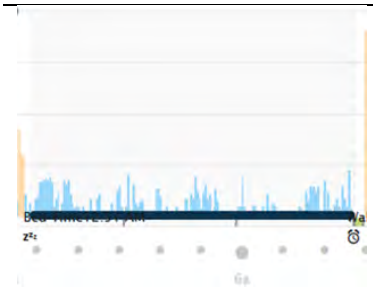
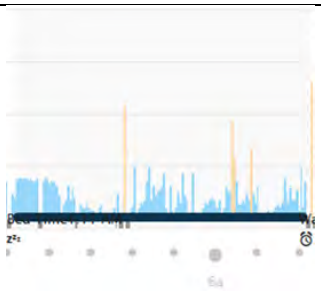
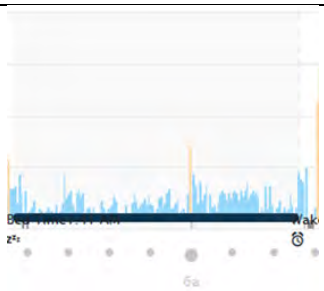
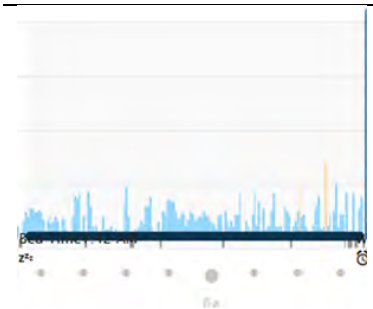
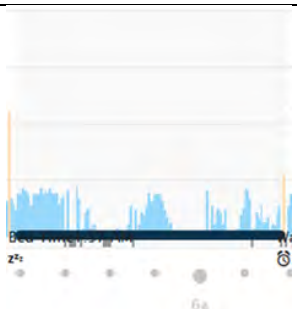
6.6.2. Case Study 4: Participant 30 Sleep

Participant 30's sleep data for fourteen days is presented in Table 29 below.

Table 29

Sleep Data: Participant 30, Days 1 to 14

Day 1 (Monday)	Day 2 (Tuesday)	Day 3 (Wednesday)
		
		
Day 4 (Thursday)	**Day 5 (Friday)**	**Day 6 (Saturday)**
		
		

Day 7 (Sunday)	Day 8 (Monday)	Day 9 (Tuesday)
		
		
Day 10 (Wednesday)	**Day 11 (Thursday)**	**Day 12 (Friday)**
		
		
Day 13 (Saturday)	**Day 14 (Sunday)**	
		

Participant 30 averaged about six hours and forty minutes of sleep a night (6h41m). Sleep quantities were in a fairly narrow range, from a low of five hours and fourteen minutes (5h14m) on Day 14 to a high of seven hours and thirty-two minutes (7h32m) on Day 13. Sleep quantities were within half an hour of the mean sleep quantity of six hours and forty minutes on nine of the fourteen nights.

This is notable because the longer sleeping time on weekends seen in some other participants is not present in Participant 30's sleeping patterns. This raises the question of whether Participant 30 naturally required less sleep because of their individual biology, was unable to sleep because of insomnia or anxiety, or was limiting their sleep quantity for some reason.

Participant 30's consistently elevated stress levels tend to indicate some sort of persistent stressor in their lifestyle, and insufficient sleep is a logical contributing factor. The issue of sleep sufficiency is a complex topic as criteria differ. For language learning there is evidence that suggests that six hours of sleep is sufficient, so lack of sleep does not explain the lack of language learning (MacDonald, 2015). For general health there is evidence that suggests that less than seven hours of sleep was associated with a rise in all-cause mortality (Tamakoshi & Ohno, 2004). The research into sleep and stress suggests a bidirectional relationship. Higher stress levels lead to reduced total sleeping time and lower sleep quality. Reduced total sleeping time and lower sleep quality could similarly lead to higher stress levels (Yap, et al, 2020). Causality is unclear, whether the lack of sleep is causing stress or the stress is causing lack of sleep. It might also be both the lack of sleep and high stress levels are symptoms of an underlying clinical condition. Regardless of the genesis, Participant 30's sleep quantity is unhealthy. This adds a fourth potential factor, health concerns.

What is concerning is Participant 30's sleep quality. Looking at the grey lines at the bottom of the sleep graphs there is evidence of movement during the first 3 hours of sleep almost every night. This occurred every night for approximately the first three hours of sleep, showing disturbed sleep during the first two sleep cycles, and strongly suggesting low sleep

quantity. In addition, Participant 30 woke for at least some portion of the night every night, ranging from 1 minute on Day 4 to 42 minutes on Day 14.

In a systematic review of the literature around sleep and anxiety Papadimitriou and Linkowski (2005) report that, “In a meta-analysis based on 177 studies with 7151 psychiatric patients, including those suffering from anxiety disorders, compared to normal subjects it was demonstrated that there was disrupted sleep continuity with significant reduction of total sleep time (TST)” (p. 229). They also noted that in 18% of cases insomnia appeared before the anxiety, although in 82% of cases it was at the same time or after the onset of anxiety (Papadimitriou & Linkowski, 2005).

6.6.3. Case Study 4: Participant 30 Activity

Participant 30’s activity data is summarised in Table 30 below.

Table 30

Activity Data: Participant 30, Days 1 to 14

Day 1 (Monday)	Day 2 (Tuesday)	Day 3 (Wednesday)
Steps: 6,259 Kilometres: 4.6	Steps: 8,172 Kilometres: 6.1	Steps: 10,116 Kilometres: 7.8
Day 4 (Thursday)	**Day 5 (Friday)**	**Day 6 (Saturday)**
Steps: 6,716 Kilometres: 4.9	Steps: 14,851 Kilometres: 10.9	Steps: 3,889 Kilometres: 2.9

Day 7 (Sunday)	Day 8 (Monday)	Day 9 (Tuesday)
Steps: 12,662 Kilometres: 9.3	Steps: 7,047 Kilometres: 5.2	Steps: 9,901 Kilometres: 7.4
Day 10 (Wednesday)	Day 11 (Thursday)	Day 12 (Friday)
Steps: 5,085 Kilometres: 3.8	Steps: 8,142 Kilometres: 6.0	Steps: 12,283 Kilometres: 9.1
Day 13 (Saturday)	Day 14 (Sunday)	
Steps: 3,715 Kilometres: 2.7	Steps: 13,279 Kilometres: 9.8	

Participant 30's activity levels appear to vary widely from day to day, ranging from 3,715 steps on day 13 to 14,851 steps on day 5. Over the fourteen days of the wearing the *Garmin Vivosmart 3/4* Participant 30 walked 122,117 steps, averaging 8,722.64 steps a day. It thus seems unlikely that Participant 30's high stress levels arose from insufficient exercise. While Participant 30's activity levels do vary widely, they are regular in the sense of following a weekly pattern. If one compares Day 1 (Monday) with Day 8 (Monday) the step quantities are quite similar. This holds for all the days except Day 3 (Wednesday) where Participant 30 walked 10,116 steps, against Day 10 (Wednesday) where they walked only 5,085 steps. However, this may be a result of a chunk of missing data from 09h30 to 15h30. Judging by the timing, the most likely explanation seems to be that Participant 30 removed the *Garmin Vivosmart 3/4* on waking to shower, then forgot to put it back on until later. Regardless of the reason, this is the most likely reason for the disparity between these two days.

Adopting this method of comparing the same day of the week across weeks changes the assessment about the regularity of Participant 30's regularity in activity, with it appearing that Participant 30 had a regular exercise routine. This has a bearing on the recommended time period for the use of wearable devices in research.

6.7. Case Studies Closing Comments

As discussed at the beginning of this chapter the quantity and level of detail in the data gathered by wearable devices represents both an opportunity and a challenge. Wearable devices present the researcher with data that gives rich and detailed insights into the participants' daily lives. However, this wealth of data also presents challenges to traditional

methods of analysis and communicating these insights. This section has presented some of the data from just five of the eleven participants who submitted wearable data, highlighting issues that require further investigation and discussion.

In Chapter 7 some of the issues highlighted for consideration in the case studies will be investigated using statistical methods to see if the relationships suggested in the case studies might be further elaborated through statistical analysis.

Chapter 7: Statistical Analysis

This chapter will present the quantitative analysis of the data gathered in this study. In some areas the sample size available is on the borders of what is required for statistical analysis, as a result of the low return rate of data from the wearable devices. However, the analysis conducted here should be read in conjunction with the detailed analysis from the case studies section and is primarily confirmatory and works to triangulate aspects of the data. This section follows up on issues raised in the case studies, to see if the statistical evidence supports the case-by-case observations.

For this analysis, the GHQ-28 results from the mid-study (week 4) test will be used, as these results are from the period during which the *Garmin Vivosmart 3/4s* were being worn and stress measurements were being taken.

7.1. The FLCAS

Chapter 2 presented a critique of the FLCAS, challenging its validity and reliability. This is important, since the FLCAS is the current dominant measure of LLA and provides the statistical basis for a number of contentions that will be challenged in this thesis. The FLCAS is not only cited as support for the notion that LLA exists, but also for the contention that it is situation specific. Further, the mathematical modelling of the FLCAS suggests a straight-line negative correlation between the FLCAS and LLA.

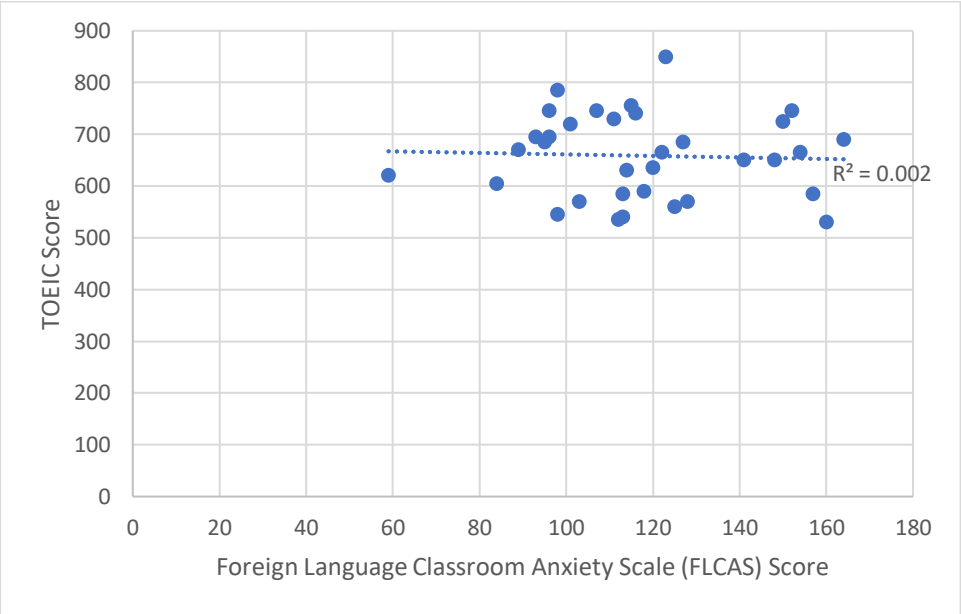
The challenges raised to the FLCAS's reliability and validity in Chapter 2 should be sufficient to seriously question the evidence accumulated under the FLCAS. The detailed case studies provided some evidence that the FLCAS was not a reliable predictor of language proficiency, language learning, or stress. This section will present more evidence in support of these challenges.

7.1.1. The FLCAS and Language Proficiency

In this study the TOEIC was used as a measure of language proficiency, providing a reliable and validated third party measure of language proficiency that was not subject to issues of inter-rater reliability. Scores were also normed within the population of test takers, allowing meaningful comparisons between individuals.

Figure 19

TOEIC Scores versus FLCAS



The data presented in Figure 19 above suggests that there is no relationship between performance on the TOEIC (as a measure of English proficiency) and FLCAS score. Participants with similar TOEIC results (the left axis) score very differently on the FLCAS (the bottom axis), creating the broad band across the top of Figure 19. This band creates a linear trend line with an R^2 value of 0.002, an almost flat line, suggesting no correlation between the two variables.

To confirm the result, a Pearson's Product-Moment Correlation Coefficient was calculated for the two variables, FLCAS score and TOEIC score, and produced a result of r= -0.088, p=0.619, n=34, indicating that the probability is very high that there is no statistically significant relationship between FLCAS scores and English performance, with the normal threshold being regarded as a p value of 0.05 in the social sciences.

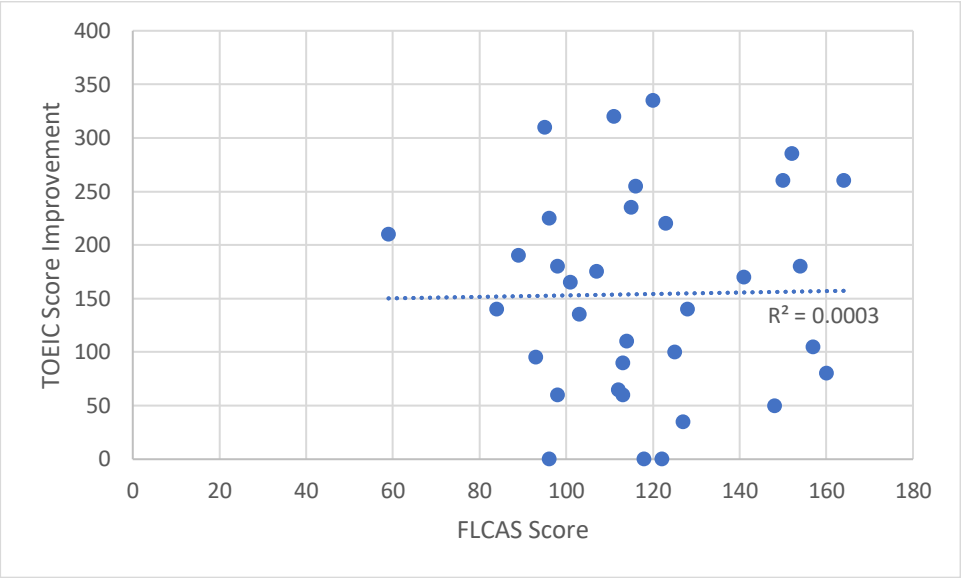
This result lends supports to the observations in the case studies that the FLCAS is not a predictor of language proficiency. The next step was to examine whether there is any correlation between the FLCAS and language learning.

7.1.2. The FLCAS and Language Learning

In this study participants' language learning was represented by the difference between their TOEIC scores on entering university and their scores one year later. If the FLCAS is a predictor of language learning then there should be some correlation between FLCAS scores and the degree of improvement shown by the participants over a year.

Figure 20

FLCAS and English Language Learning



The data points in Figure 20 are all broadly distributed with no pattern being immediately apparent. The R^2 value of 0.0003, and an almost flat trendline confirms that there is no pattern hiding in the seemingly random distribution.

A Pearson's Product-Moment Correlation Coefficient was calculated with a result of r=0.17, p=0.922, n=34. There is a very high probability that there is no statistically significant correlation between language learning and the FLCAS measure.

As some students showed zero improvement in their TOEIC scores over the period, it was not possible to calculate alternative trendlines for this dataset. However, looking at the scatter plot of the data in Figure 20, it seems clear that there is no trendline that would cover a significant number of the datapoints.

This lends support to the observations in the case studies section, that there seemed to be no significant relationship between FLCAS and English language learning.

7.1.3. The FLCAS and Anxiety

The FLCAS purports to be a measure of anxiety, claiming strong correlations to several measures of anxiety, but does not distinguish between clinical or non-clinical anxiety. This section will investigate whether there is any relationship between the FLCAS and GHQ-28 scores as an indicator of clinical distress of any kind. Next, what will be established is whether there is a relationship between the FLCAS and the GHQ-28 anxiety subscale. This second analysis is included with the caveat that this is not the intended purpose of the GHQ-

28 anxiety subscale, and that this analysis is included in the interests of completeness. For this section, the mid-study GHQ-28 total scores were again used.

Figure 21

FLCAS and GHQ-28 Total Scores

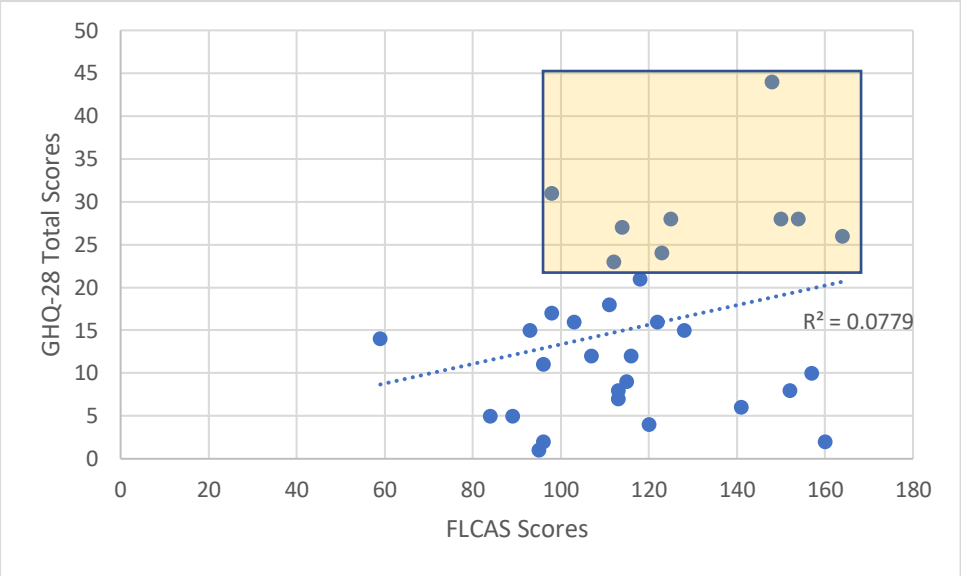


Figure 21 above presents the results of the total GHQ-28 scores graphed against the FLCAS scores. When graphed against one another and analysed with a linear trendline (R^2=0.0779) it can be seen that most values do not fit to along the trendline.

A Pearson's Product-Moment Correlation Coefficient was calculated for the two variables, FLCAS and GHQ-28, producing a result of r=0.279, p=0.122, n=32. This indicates that there is no statistically significant relationship between GHQ-28 scores and FLCAS scores in general.

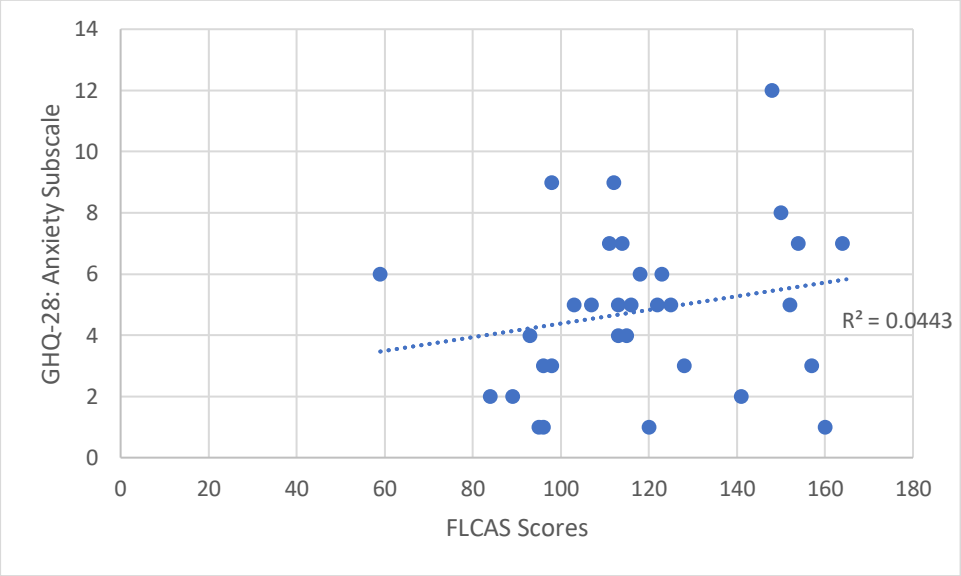
However, only GHQ-28 scores over the threshold of 23 indicate clinical levels of distress. These values have been highlighted in an orange rectangle. While four of the data points in the orange rectangle do score above average on the FLCAS, equally so five of the data points score in the normal range on the FLCAS. The pattern in the data suggests that there is no statistically significant relationship between GHQ-28 scores indicating clinical distress and the FLCAS. A line drawn to incorporate most of the points in the orange rectangle would be close to a flat line.

No statistical analysis was attempted because the sample size falls below the subminimum for valid statistical analysis. However, the data available, while admittedly insufficient, suggests no relationship between clinical distress and the FLCAS. This appears to corroborate what the case studies suggested.

Proponents of the FLCAS might argue that this is an unfair comparison as the GHQ-28 is a screening measure for general clinical distress, not specifically anxiety. Therefore, a further analysis of the anxiety and insomnia subscale was conducted. As noted earlier, this is not a use for which this subscale has been validated, and so it is included purely for interest's sake, and undue weight should not be attached to the results.

Figure 22

FLCAS and GHQ-28 Anxiety Subscale



Looking at Figure 22 there is no clear link between anxiety and FLCAS scores. The trendline in Figure 22 shows some clustering of values around the trendline with an R^2 value of 0.0443, which indicates a better fit than for the total GHQ-28 score in Figure 21. However, most of the values do not fit around the trendline and there are numerous outliers.

A Pearson's Product-Moment Correlation Coefficient was calculated for the two variables, FLCAS score and the GHQ-28 anxiety subscale score, producing a result of r=0.211, p=0.247, n=32. This suggests that the GHQ-28 anxiety subscale score and FLCAS do not covary.

This result is not valid or reliable though, as the GHQ-28 anxiety subscale was not intended for this sort of use, and there is no subminimum on each scale to distinguish between clinical and non-clinical distress by sub-scale. A more reliable measure would be to investigate whether there is any relationship between stress and the FLCAS.

7.1.4. The FLCAS and Stress

If the FLCAS is a measure of anxiety, and theories of stress and anxiety such as Hebb (1955) are correct, then there should be a relationship between stress and anxiety. Since the

FLCAS has no minimum threshold for anxiety, there should be a straight-line relationship between the FLCAS and measures of stress. This analysis will rely on the data from the *Garmin Vivosmart 3/4s*, and while the sample size is smaller than hoped there, the data are sufficient for statistical analysis.

Figure 23

FLCAS and Average Stress

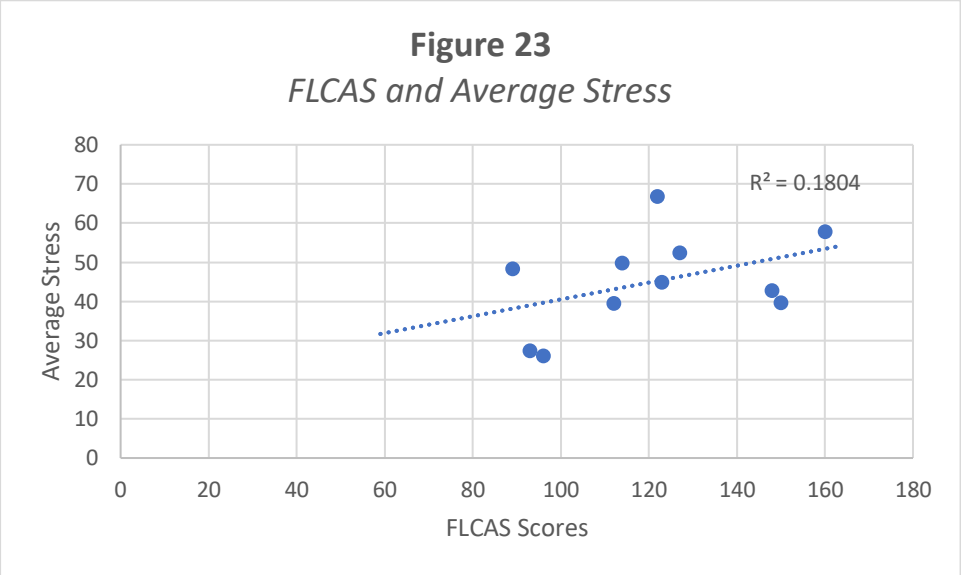


Figure 23 shows a straight positive trendline between stress and FLCAS scores. While not particularly surprising this is useful in suggesting that the FLCAS is probably measuring stress, not anxiety as the name of the construct suggests. However, this relationship is quite loose. A good example of this can be found by considering the participants with average stress ratings of 40 to 50. Despite these individuals scoring in quite a narrow 10-point (40 to 50) stress there are a wide variety of FLCAS scores, ranging from 89 to 148, indicating no real predictive relationship between FLCAS scores and average stress. The trendline confirms this poor fit for the data (R^2=0.18).

A Pearson's Product-Moment Correlation Coefficient was calculated to check if there was a statistically significant correlation; resulting in r=0.425, p=0.193, n=11. The significance (p value) is well over the threshold of 0.05. The statistical analysis tends to confirm what the case studies suggest, which is that the FLCAS does not predict stress or stress-related anxiety.

Again though, proponents of the FLCAS might challenge this analysis based on the argument that the FLCAS was never intended to be a measure of general stress, but rather is a measure of anxiety specifically in the foreign language classroom. A final analysis will be conducted on the FLCAS and stress as measured by HRV in English classes.

Figure 24

FLCAS and Stress in English Classes

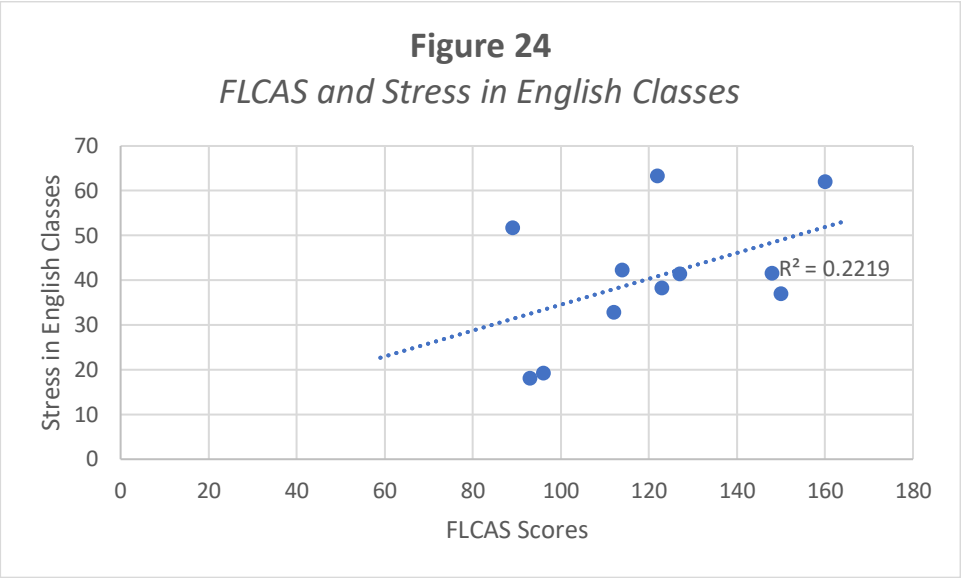


Figure 24 shows no clear relationship between FLCAS scores and average stress in English classes. More than half of the values are not near the trendline, with an R^2 value of 0.2219.

A Pearson's Product-Moment Correlation Coefficient was calculated to determine if there was a statistically significant correlation between FLCAS and stress levels in English classes, resulting in r=0.471, p=0.144, n=11. The p value falls above 0.05, indicating no statistically significant relationship between the FLCAS and stress levels in English classes.

7.1.5. Conclusions regarding the FLCAS

Both the case studies and the subsequent statistical analysis suggest that there is no evidence that the FLCAS covaries with language performance, language learning, clinical distress, clinical anxiety, stress in general, or stress in language learning classes. This tends to suggest that the answer to Research Question 4 is that the FLCAS is not a valid predictive measure of LLA.

The questions raised about the evidence for the FLCAS raised in Chapter 2, combined with the evidence presented here showing no significant correlations between the FLCAS and any of the expected variables, have serious implications for LLA theory. Before the FLCAS, the evidence for the existence of LLA was mixed, as shown in Scovel (1978). The evidence presented here sets LLA back to Scovel (1978), before Horwitz's (1986) preliminary evidence for the FLCAS as a measure of LLA.

Therefore, the question now moves to Research Question 1, namely whether any credible evidence can be presented that there is a situation-specific stressor that exists only

in the language learning environment and results in elevated stress levels in this environment. The logical starting point is to look at stress averages in different environments. This is an area where prior research into LLA has failed to provide evidence, merely asserting that the learning environment is an especially stressful and anxiety-provoking place.

7.2. Stress Average by Context

The detailed data presented in the case studies suggested that while there might be moments of anxiety and stress in language learning classes, there were similar patterns in non-language classes. This similarity in patterns does allow for meaningful comparisons of average data.

Figure 25

Stress Averages: English Classes versus Non-Language Classes

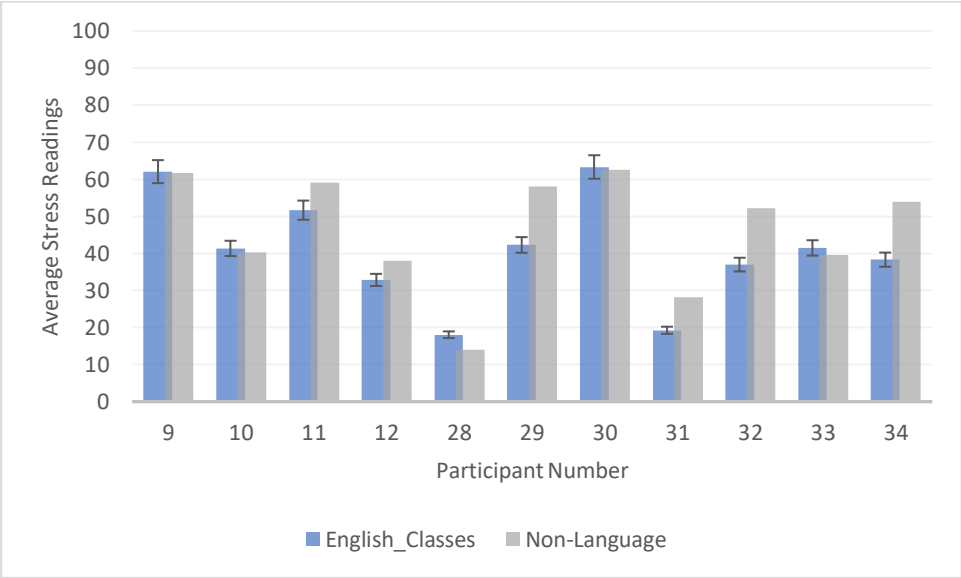


Figure 25 compares the average stress levels between English classes (blue) and non-language classes (grey). English class stress averages range from a low of 18.05 to a high of 63.31, while non-language classes range from 14.07 to 62.61. Five participants (Participants 9, 10, 28, 30, and 33) found English classes more stressful than non-language classes, while six participants found non-language classes more stressful than English classes.

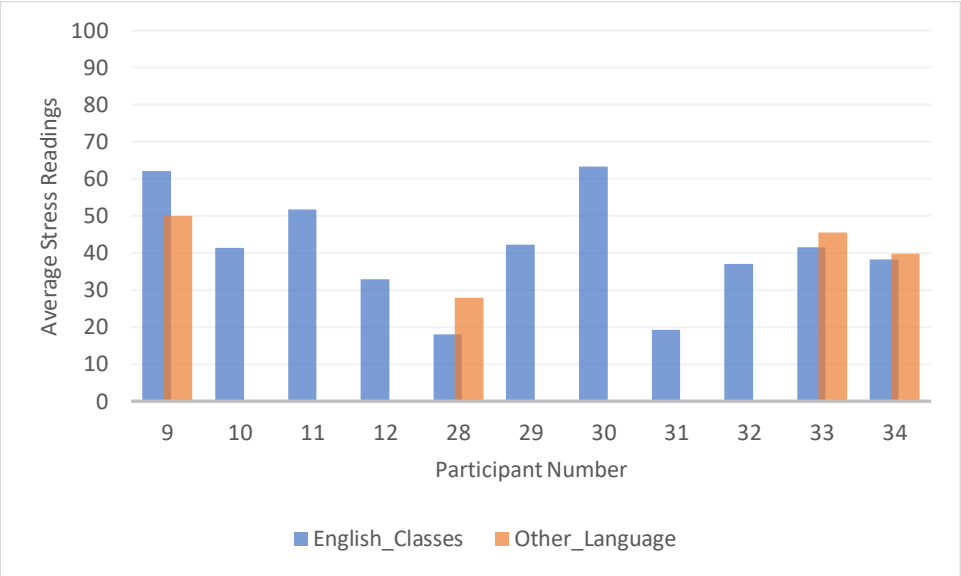
Standard error bars were added to Figure 25 set at 5% of the total stress value to illustrate that in those cases where stress levels in English classes were higher than non-language classes. The difference was less than 5% variance in all cases except Participant

28. Participant 28's average stress levels were in the resting stress (0 to 25) range in both English classes (18.05) and non-language classes (14.07).

There is no clear or consistent pattern of English classes producing appreciably more stress than non-language classes. This tends to suggest that the notion of special heightened levels of anxiety in language classes is not supported.

Figure 26

Stress Averages: English Classes versus Other Language Classes



Only four participants took both English language classes and other language classes. Figure 26 shows that the 3 out of 4 participants found other foreign language classes more stressful than English classes. However as in Figure 25 the difference in stress levels are quite similar in two of the cases (participants 33 and 34).

When stress levels in classes were compared to stress levels outside of classes (as in Figure 27 below), there was a pattern of class time being generally less stressful than time outside of class.

Figure 27

Stress Averages: English Classes vs. Free Time

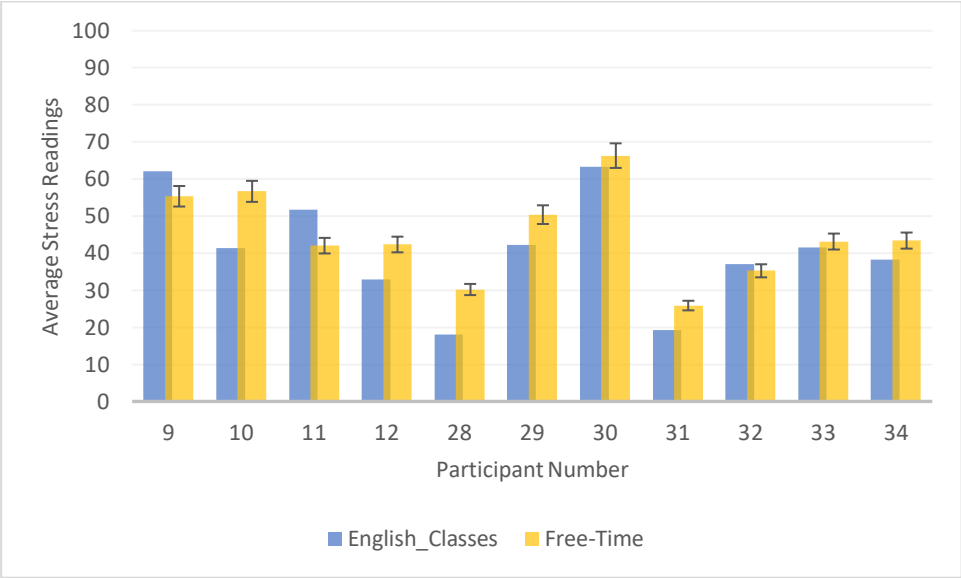


Figure 27 compares average stress measurements in English classes (blue) against free time. In eight of the eleven (72.73%) cases, free time was recorded as being more stressful than English class time.

Standard error bars have been added to Figure 27 to show differences of more than 5%. These error bars show that in six of the eleven cases (Participants 10, 12, 28, 29, 31, and 34) free time was more than 5% more stressful than English class. In only two of the eleven cases (Participants 9 and 11) was English class time more than 5% more stressful than free time.

Referring back to Figure 25 both participants 9 and 11 experienced similar or higher levels of stress in non-language classes, indicating that these higher stress levels are not likely to be specific to language classes. This is consistent with the data presented in the case studies, where free time was generally more stressful than class time, and where heightened stress tended to be global rather than situation specific.

This tends to argue against the assertion by LLA theorists that foreign language learning classes can be assumed to be especially stressful and anxiety-provoking environments. The data presented in this section represent a substantial challenge to the notion of LLA and to other notions of academic anxiety such as those presented in Cassady (2020). The data tends to suggest that where anxiety is present, it is relatively persistent across contexts. If anything, the learning environment seems to represent a less stressful environment, on average, for most participants than daily life. Again, this is not to argue that learning environments are free of stress or anxiety, merely that the notion of learning

environments as especially anxiety-provoking or stressful places is not supported by the data.

Nonetheless the idea of LLA has remained popular over a protracted period and attracted a great deal of scholarship. It is supported by numerous interviews with students, observations from teachers, and has a great deal of face validity for a reason. Therefore, the analysis in this chapter will continue by examining the data to see if there is a link between stress experienced by students in language learning classes and their levels of language proficiency or language learning. This analysis continues with the caveat that it is not premised on situation-specific anxiety as postulated in classical LLA theory.

7.3. Stress and Language Proficiency

A logical place to start is with the direct measures of stress from the fitness trackers during English classes, and to compare those to the participants' TOEIC scores, in order to assess whether there is any meaningful correlation. These are shown in Figure 28.

In this section and the section that follows, trendlines were used to display the underlying patterns in the data. These trendlines were calculated by Excel and aim to generate a line that comes closest to the largest number of points in the dataset. The goodness of the fit is measured by the R^2 value, with 1 being a perfect fit with all the data points touching the line, and 0 being a complete lack of pattern. Two different types of trendlines were contemplated, namely linear and curved. Horwitz, et al. (1986) postulated a linear (straight-line) relationship between stress and performance, while Hebb (1955) postulated a curved relationship between stress and performance. With the R^2 value from the trendline and knowing the sample size it was then possible to calculate the significance (p) value. For the linear trendlines the formula for Pearson's Product Moment Correlation Coefficient was used to calculate the significance. For the curved trendlines the F-test was used to calculate the significance.

Figure 28

TOEIC Scores and English Class Average Stress (Linear Trendline)

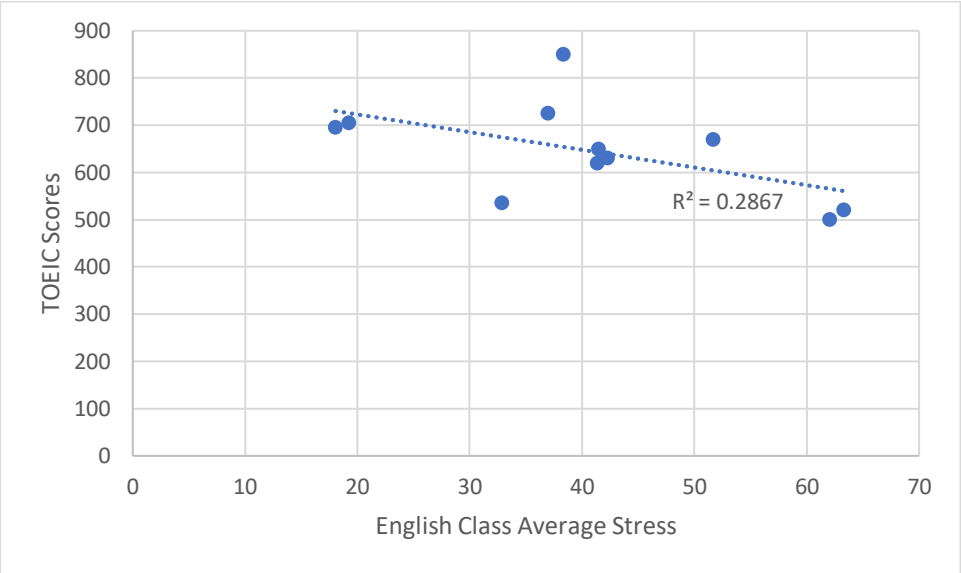


Figure 28 shows a loose fit with the available data with an R^2 value of 0.2867, and the linear trendline showed a negative slope. This suggests that if there is a correlation it will be negative, in other words as English class stress levels increase, so TOEIC scores decrease.

A Pearson Product-Moment Correlation Coefficient was calculated for the available data, and indicated r=-0.535, p=0.090, n=11. This indicates a negative correlation of moderate strength and marginal significance.

The use of marginal correlations is controversial. A limitation in this section of the analysis is the small number of participants who submitted sufficiently complete stress data from the wearable devices. This imposes limitations on the types of statistical analysis possible and will tend to result in less certainty over the significance (p value) of the findings. However, the sample sizes are sufficient for the type of statistical analysis conducted as long as one bears these limitations in mind (Kirk, 2007). As a result, in this section some latitude will be given in the interpretation of significance to account for the small sample sizes, such as considering marginally significant correlations (p>0.05<0.10) (Pritschet, et al., 2016).

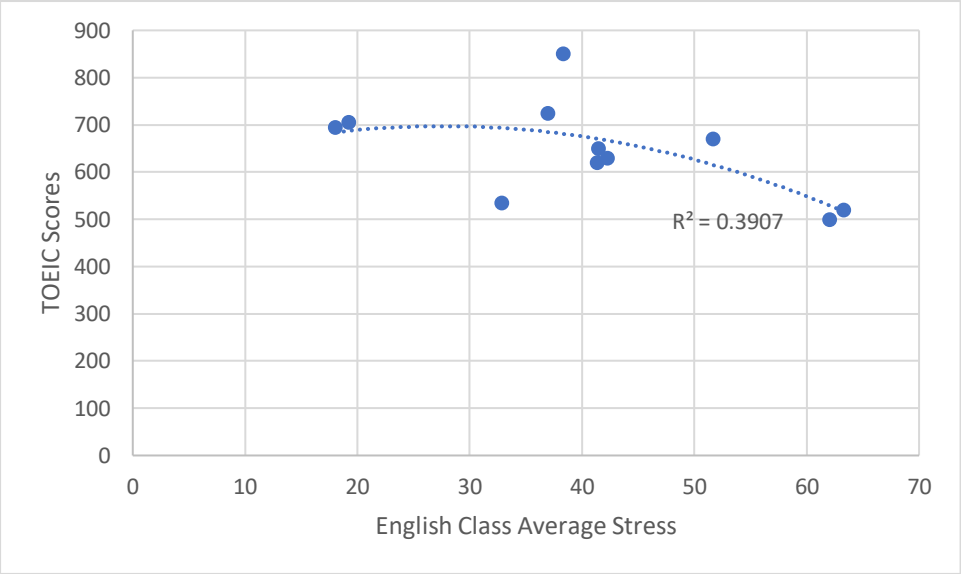
The marginal correlation reported in Figure 28 is included to highlight that there may be something there that deserves further consideration. As a result of this marginal correlation further trendlines were investigated.

According to the Hebbian (1955) model of performance and stress/arousal the expected relationship is an inverted U shape, with lower levels of performance at low stress, peak performance at optimal stress, and lower performance at high levels of stress.

Therefore, a curved trendline was investigated, asking Excel to calculate the best fit for the data points on the assumption that the line curved around one point.

Figure 29

TOEIC Scores and English Class Average Stress (Curved Trendline)



When this type of curved trendline was examined for fit, the R^2 value rose from 0.2867 to 0.3907, indicating a better fit for the data. This is to be expected given that this sort of relationship fits with well-established theories regarding stress and performance, such as Hebb (1955).

The p value for this curve is below 0.05 (p=0.0397) indicating that the correlation is statistically significant. There does seem to be some support for the notion that stress and language performance are related.

However, this comes with several caveats regarding the impact of stress and language performance. The first is that this stress is not necessarily situation specific, with evidence suggesting similar levels of stress in all classes, and that stress, rather than language performance, is the independent variable. As was seen in the case studies where clinical distress was present, the distress was global rather than situation specific.

Secondly, there is the issue of prior learning. Both of the data points to the right of Figure 29 were for participants with below-average initial TOEIC scores, and their higher stress levels in English classes may be a result of this mismatch between their English ability and the level of the classes, which are aimed at the level of the average student. Rather than stress or anxiety creating lower levels of language performance the opposite may be true: that lower levels of initial language performance may be creating higher stress levels.

Similarly, the two data points to the left of Figure 29 should be familiar to the reader, representing Case Studies 3A and 3B. As was discussed in these case studies there is the possibility that these participants' lower stress levels are a result of lack of stimulation in classes aimed at the average student.

Thirdly, the relationship illustrated in Figure 29 may be an artifact of the TOEIC test being used as a measure of language proficiency. The students with already high TOEIC scores may have felt no pressure to improve their performances because they had already achieved the required level for graduation. On the opposite end of the spectrum, the students with high stress may have found it more difficult to perform well on a timed test like the TOEIC test because they were stressed, which led to errors and a lower resultant score.

The three notes above offer potential explanations for why this pattern might be seen in the data, and all of them fit the data without needing to assume the existence of an unproven situation-specific stressor such as LLA. These explanations are also commensurate with established theories regarding stress and performance, such as Hebb (1955) and Csikzentmihalyi (1997). The curved line, which is the relationship between stress and performance predicted by Hebb (1955), fits the data better than the straight-line predicted by Hortwitz, et al. (1986). The curved line also generates a statistically significant result at $p \leq 0.05$ ($r=0.63$, $p=0.0397$, $n=11$), whereas the straight-line did not.

The discussion about prior versus present language proficiency leads then to the core issue in LLA, namely language learning, and whether there is a link between stress and language learning.

7.4. Stress and Language Learning

As the name implies, LLA is concerned with the theorised link between language learning and anxiety. This section will analyse the dataset in order to explore any links between the two factors.

Figure 30

English Class Stress and English Proficiency Improvement over Time (Linear Trendline)

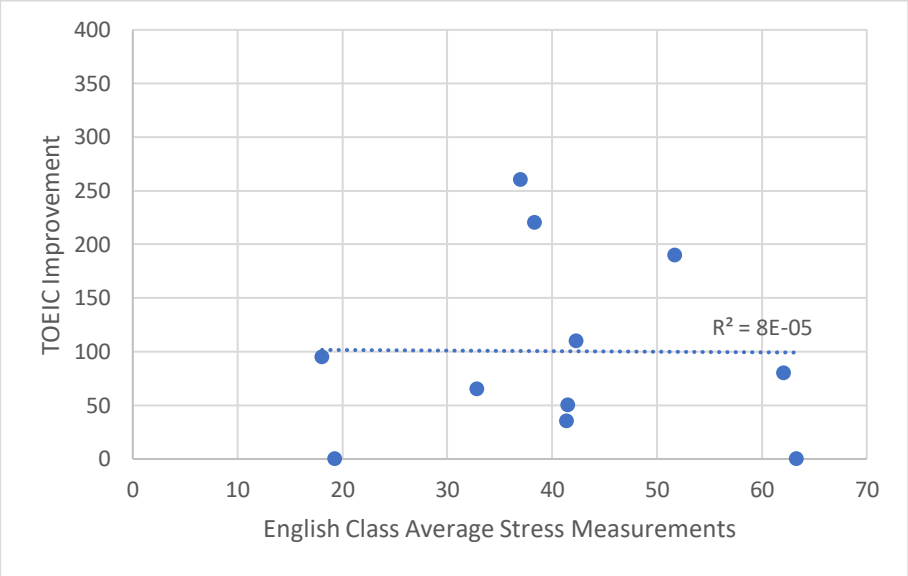
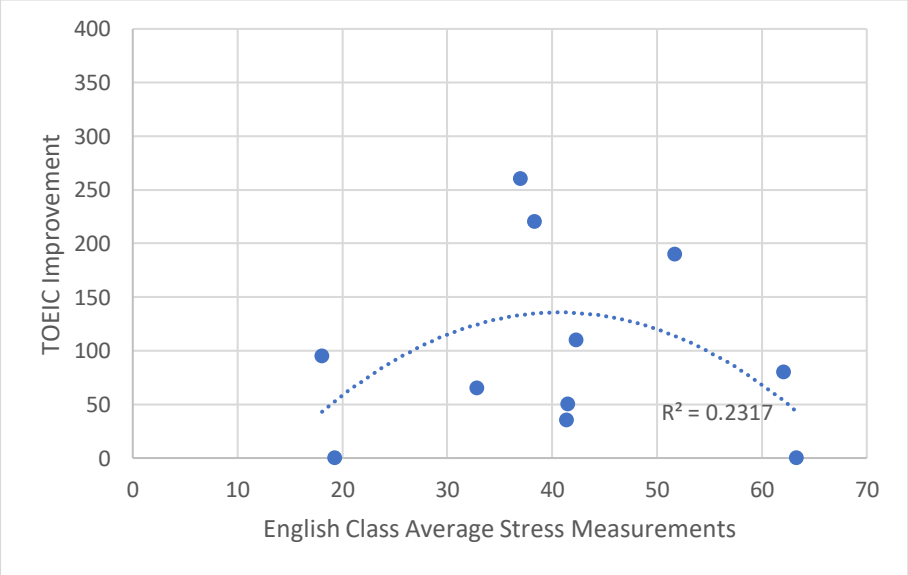


Figure 30 shows that there is no straight-line correlation between stress and language learning. This possibility was investigated first because it is the relationship predicted in current LLA theory.

Figure 31

English Class Stress and English Proficiency Improvement over Time (Curved Trendline)



A curved trendline, as shown in Figure 31, was a better fit for the data and describes the inverted-U predicted by Hebb (1955). Figure 31 shows significant levels of language learning impairment at both the upper (stress 60 or higher) and lower (stress 30 or less) ends of the

stress scales, with varying levels of language learning happening towards the mid-range (between 32.83 and 51.66) of the stress measurements. Peak learning seems to occur around average stress measurements of about 40, with a single outlier at 51.66.

The correlation seen in Figure 31 is not statistically significant, with a p value of greater than 0.05 (p= .13382). This is to be expected given the complex range of factors that influence language learning, such as variations in stress levels over time. It cannot be assumed that stress levels were stable for the entire year. While two weeks of measurement may give some indication of general stress, it would not necessarily give an accurate indication of stress over a whole year. The stability, or lack thereof, of stress over protracted periods is something that merits further research, with the data in the case studies shows considerable variation from day to day.

Another potential area for critiquing the methodology in this study is that all learning was weighted equally, so if a participant increased their score from 500 to 600 this was treated the same as a participant who increased their score from 800 to 900. Yet the TOEIC is a normed test, so to achieve a score of 600 might require a test taker to score only a little above the average test taker. However, achieving a score of 900 might require the test taker to score in the top couple of percentage of test takers. In this study an improvement of 100 points was treated as the same, regardless of whether it was from 500 to 600, or from 800 to 900. Some sort of weighting system was contemplated as a means of acknowledging that an improvement from 800 to 900 is not the same as an improvement from 500 to 600, but was ultimately rejected as requiring too many unsupported assumptions about the norming process used by the TOEIC.

While the trendline in Figure 31 is not statistically significant, the p value could be interpreted as reflecting about 86.62% certainty that the trendline shown in the figure is non-random. This is a fair value to achieve when placing a single variable against as complex a factor as learning. It suggests that stress is probably a factor in English language learning, but is likely to be only one of many.

7.5. Conclusions

The FLCAS does not bear up under any type of analysis based upon the data collected in this study. Curved trendlines were also attempted on the FLCAS data, but the R^2 values were lower than those for straight-line correlations. This suggests that all the literature based on the FLCAS may likewise be in doubt.

Statistical tests aside, Figures 31 and 32 show what accepted models of stress and learning such as Hebb (1955) predict; and agree with the evidence presented in the case

studies. Perhaps most importantly this arrangement explains the link between stress and anxiety, as seen by language teachers and reported by students that LLA sought to explain. It does so without the need for the construction of a special type of anxiety specific to the language classroom. Occam's razor, a philosophical principle first advanced by William of Ockham in the 14th century, and often used in scientific enquiry states that, "Entities should not be multiplied unnecessarily." (Gibbs, et al., 1996, p. 1). As discussed in Chapter 2, LLA has its foundations in the observations by teachers that anxious students seem to perform worse than non-anxious students. As a result of this, a special type of anxiety present only in the language learning environment, LLA, was theorised. What should be considered is that LLA is an unnecessary complication that violates Occam's razor. It tries to introduce a new entity without sufficient evidence and without the construct being necessary. Hebb's (1955) model not only provides the best fit for the data, but also fits with all the other data presented here.

It also makes sense, particularly in cases where clinical levels of distress are present. As was shown in the case studies, such individuals show elevated levels of stress across all contexts. Similarly, at the lower end of the stress curve, the results fit predictions in terms of existing models, showing lower levels of language learning.

This inclusion of stress measurements from a wider range of contexts appears to be the decisive factor, showing that for the majority of participants (eight out of eleven, or 72.73%) stress levels were higher during their free time than during English language classes. In all three of those cases where stress levels were higher in English language classes than during their free time, those participants' stress levels were similar or higher in non-language classes. This sort of measurement of stress across a wide range of contexts was something that was extremely difficult to do in the past. Laboratory measurement of stress was naturally restricted to the laboratory. Questionnaire methods imposed an undue burden on participants. Interview methods relied on recollections. The type of measurement of stress levels across a broad range of contexts as in this study was enabled by wearable devices, and suggests that a fundamental paradigm shift is necessary regarding perceptions around stress levels in learning environments.

In the following chapter of this thesis the full implications of what has been presented so far will be discussed, and then the final conclusions presented.

Chapter 8: Discussion of Results

The case studies presented in Chapter 7 evidence certain common patterns, which will be discussed collectively in this section alongside the results from the statistical findings from Chapter 7. This discussion will commence with the research questions posed in this study and conclude with additional findings that were not the primary focus of the research but are nonetheless interesting and important.

8.1. Research Question 1

Is there evidence of a situation-specific stressor that exists only in the language learning environment? This question is central to the theory of LLA. Without evidence of specificity the phenomenon of LLA becomes a question of the impact of distress in general on performance and learning, rather than something specific to the language learning.

In Case Studies 1 and 2, the participants' (Participants 12 and 10 respectively) GHQ-28 scores suggested that there was possibly clinical distress, with both participants scoring consistently above the caseness threshold. In both Case Study 1 and 2 anxiety is an implicated factor, however not the chief complaint. This is an important consideration since, as discussed in Chapter 2, the LLA literature is replete with anecdotes from LLA theorists citing examples of anxious behaviours in language learning students and associated diminished language performance and language learning. Case Studies 1 and 2 seem to support these observations, with both Participant 12 and Participant 10 performing markedly below average in language performance, as measured by the TOEIC test. Participants 10 and 12 also showed well below average improvement in their TOEIC scores, which in this study was used as a measure of language learning. Therefore, it should be stated from the outset that previous LLA theorists' observations of anxious behaviour being linked to lower levels of language performance and language learning seem to have some basis. However, the presence of anxious behaviour does not mean that it is the chief complaint, nor the genesis of the effects being observed, nor does it mean that this effect is specific to the language learning environment.

What was missing in previous studies was broader contextualising data about participants' stress levels in other environments. Both Participants 10 and 12 show evidence of moments of distress (defined as stress readings in the medium range of 51 to 75), and possible anxiety (defined as stress readings in the high range of 76 to 100), across all environments. There is no clear pattern of noticeably higher stress readings in language classes versus non-language classes, or of distress and anxiety being more consistently more prevalent in one type of class or the other.

Figure 25 in Chapter 7 presents a broader comparison of average stress between language and non-language classes and tends to support the notion that there is no pattern of markedly elevated stress in English classes when compared to non-language classes. Where participants did display higher levels of stress in English classes, the difference is less than 5% in all but one case. In that single case (Participant 28) the participant's stress readings were in the resting stress range in both English (18.05) and non-English classes (14.07). Given the low stress values involved it hardly seems supportable to cite this case in support of the notion that English classes are especially stressful.

However, when comparing class time against both participants' unstructured time (free time) the data seems to suggest that free time was more stressful than either type of class time. Considered the stress data in Case Studies 1 and 2 from both other non-language classes and the broader contextualising from the participants' free time, it does not seem supportable to claim the presence of a situation-specific stressor producing additional distress and anxiety in language learning classes. By comparison with the stress readings obtained from Participant 10 and 12's free time a more supportable conclusion seems to be that class times represented relatively less stressful environments. This statement comes with the important caveat that environment seems to be a moderating factor in determining resultant stress, rather than a primary factor. As discussed in Chapter 2, stress is a complex phenomenon with multiple moderating variables, and sleep was also implicated in Case Study 2 as an important consideration. In cases where clinical distress is suspected, it is only reasonable to expect that environment would be a secondary consideration in resultant stress; therefore Case Studies 3A and 3B will be considered as these participants scored below the caseness threshold on the GHQ-28 and exhibited stress readings that were below average.

As in Case Studies 1 and 2 the participants in Case Studies 3A and 3B (Participants 31 and 28 respectively), showed higher stress levels during their free time than during either language or non-language classes. While both participants show some stress readings that may indicate transitory distress (medium stress) or possible anxiety (high stress), these are not notably more prevalent in either language or non-language classes. Case Studies 3A and 3B are also exemplars of how stress readings can vary from day to day. In Case Study 3A in Table 18 the difference between Day 2 and Day 4 is striking, as is the difference between Day 10 and Day 11 in Table 23 in Case Study 3B. Comparing and contrasting these differences between days is important because it lends support to the notion that stress readings in classes are misleading without context from outside of class. Consider Case Study 3B where on Day 10 Participant 28's stress readings in non-language class are at resting levels (0 to 25), while on Day 11 they are largely at low levels (26 to 50). Without

considering the rest of the day this difference may be misleading, leading to the conclusion that the class on Day 11 was more stressful. However, when the contextualising stress data from the rest of the day is included, the strong influence of outside stress becomes apparent. The almost doubling of stress readings between Day 10 and Day 11 and the dissemination of this stress into classes on these days provides an example of the importance of contextualising stress.

In Case Study 4 comparing the English class from 16h20 to 17h50 on Day 4 and Day 11 in Table 28 further illustrates the importance of contextualising stress with broader measurements from the participants' daily lives. Without this contextualising stress data, the lower stress readings from the class on Day 11 might easily be misinterpreted as being indicative of some sort of evidence for the existence of a situation-specific stressor in the class on Day 11. Participant 30's stress readings show evidence of distress (medium stress) with moments of possible anxiety (high stress). Compared this against Day 4 where Participant 30's stress readings are mostly in the resting stress range, with a few moments of low stress. Yet this is the same class with the same teacher in the same venue, and presumably similar content to other classes. When the contextualising stress data from other classes and Participant 30's free time activities is considered the strong effect of outside stress is apparent.

The findings from the case studies are supported by the analysis in Chapter 7 as shown in Figure 27. In eight of the eleven participants who submitted stress data, their average stress readings during free time were higher than their stress readings during English classes, and in six of those eight their free time stress readings were more than five percent higher. It should be noted that the use of percentages in this case may be misleading. For example, in Case Study 3A Participant 31's free time average stress reading was 25.87, while their average stress reading in English classes was 19.25. This is only a difference of 6.62 points in absolute terms, but in terms of percentage difference English classes were 25.59% less stressful than free time activities. Looking at Figure 27 another way to interpret the results may be to look at the stress readings in terms of categories. In most cases the moderating influence of environment did not change participants' stress category, for example participants in the moderate stress category (51 to 75) stayed in that category regardless of environment. There were some exceptions, namely Participants 10 (free time 56.68; English classes 41.34), 11 (free time 42.03; English classes 51.66), and 28 (free time 30.02; English classes 18.05). The case of Participant 11 is an edge case, moving from the upper end of the low stress category (26 to 50) into the lower end of the moderate stress category (51 to 75). Similarly, Participant 28's stress level moves from resting stress

(0 to 25) to low stress (26 to 50). Generally, the trend holds that context did not alter the participants' stress category for the majority (8 out of 11, 72.72%) of participants.

Considering the case studies presented in Chapter 6 there are two noteworthy findings regarding the hypothesised existence of a situation-specific stressor in language classes as theorised in LLA. The first is that when comparing stress readings between similar types of activities, such as language and non-language classes, the evidence seems to suggest that the differences are quite small. Were there to be a situation-specific stressor associated only with language classes then the expectation would be that stress readings would be markedly higher in language classes than in other non-language classes. This statement comes with the qualifier that this does not necessarily indicate that there are not stressors specific to each learning environment. For example, someone experiencing social anxiety might find classes that require public speaking to be more stressful. There is no evidence that the language learning environment is markedly more stressful than other learning environments. Where differences are exhibited, they may be attributable to personal factors, such as a personal preference for a certain type of class or instructor. This raises the question of how learning environments compare to other non-learning environments. As discussed briefly in Chapter 2, there is the notion of a more general academic anxiety that may mean that the reason for similar stress readings between environments is that all learning is stress-inducing (Cassady, 2010).

This leads to the second note, a comparison between the participant's stress readings inside of classes and outside of classes. If there are situation-specific stressors in learning environments that are reason for concern, then stress readings in classes should be considerably higher than participants' stress readings outside of classes. The evidence suggests the opposite of this scenario. Rather than classes evidencing higher stress readings than free time, the majority of participants (eight out of eleven) exhibited higher stress readings outside of class, both on average and when the detailed stress readings in the case studies are considered. There might be many reasons for this difference. For example, the routine and structure in the classroom environment may be less stress-provoking than daily life.

The evidence suggests that the notion of a special stressor in language learning environments that makes these environments especially stress-provoking does not seem to be supported. The inclusion of contextualising data from other participants' classes suggests that language learning classes are not markedly more stressful than non-language classes. This should not be read as claiming that participants did not experience moments of distress or anxiety in language classes. When the detailed stress readings are examined, there is

evidence of moments of elevated stress, however these patterns are also present in other classes and in daily life.

The implications of these finding are important for educators. The evidence presented in Chapters 6 and 7 suggests that while environments may have a moderating effect on stress, the participant's stress readings could largely be explained in terms of stress external to the learning environment. This finding should not be surprising as stress tends to be pervasive, spreading from one context to the another. What this means for educators is that while the classroom environment can moderate students' stress levels it is not the primary determinant of student's stress levels. This was the error in earlier LLA theory, attributing all stress and anxiety to the learning environment without any consideration of outside stress. This effect can be seen most markedly in those participants experiencing clinical distress (Case Studies 1, 2, and possibly 4), where stress showed a global effect that was largely context independent. In lower stress individuals, such as in Case Studies 3A and 3B, the percentage change between environments was more marked, with English classes showing marked reductions in stress, but as noted earlier this may be an artefact of using percentages rather than absolute scores. As a general trend, in most participants context did not alter their stress category (resting, low, moderate, or high). This is important in the context of LLA theory in that it tends to mitigate against the notion that context exerts an influence strong enough to cause or remove distress or anxiety.

This raises the next question regarding the effects of stress on language performance and learning. This is important because it may inform teaching practice in terms of achieving educational outcomes.

8.2. Research Question 2

The question of whether stress has a debilitative, facilitative, or mixed effect on language performance and language learning has been debated by LLA theorists and has been discussed in Chapter 2. In this section the evidence from the case studies will be considered to see what it suggests about the relationship between stress and language performance and language learning, as evidenced in participants' TOEIC scores.

In both Case Studies 2 and 4, which were included as exemplars of high stress, there was evidence of below-average TOEIC scores in the latest TOEIC test, and below-average improvement in TOEIC scores. Participant 10 (in Case Study 2) started university with an above-average TOEIC score, but a year later had slipped to a below-average TOEIC score. Participant 30 (in Case Study 4) exhibited a similar pattern, with a slightly above-average score on their TOEIC test when entering university but showing no improvement in their

TOEIC score over a year, which resulted in a below-average TOEIC score at the time of this study.

Participant 31 and Participant 28 (Case Studies 3A and 3B respectively) were exemplars of low stress and scored much higher than average on the university entrance TOEIC, tending to indicate higher initial levels of English proficiency. However, these participants' scores showed below average improvement over the intervening year, with Participant 31 showing no improvement in their TOEIC score. Improvement in TOEIC scores over the year was used as a proxy for language learning. Therefore, both Participant 31 and 28 evidenced some degree of diminished language learning. Despite this diminished language learning both Participants 31 and 28 had TOEIC scores that were above-average at the time of this study, as a result of their much higher initial scores.

Case Study 1 was the only case where the participant entered university with a below-average TOEIC score, and below-average improvement on the TOEIC score widened this gap considerably. As discussed earlier in this chapter, the combination of very high and very low stress readings in Participant 12's stress data combine to produce an average stress level that seems to be below average. A closer examination of the detailed stress graphs reveals that in this case the averaging of Participant 12's stress readings masks evidence of frequent distress (medium stress) and possible anxiety (high stress). These higher stress readings are counterbalanced by frequent moments of resting stress readings that, when the data is averaged, mask the higher stress readings.

The case studies raise two important issues. The first is the difference between language performance and language learning. This difference between language performance (as measured by the participants' latest TOEIC score), and language learning (as measured by improvement in the participant's TOEIC scores over a year) is a non-trivial point of contention in LLA theory. If just language performance is considered, then the data from Case Studies 2, 3A, 3B, and 4 would seem to conform to LLA theory as hypothesised. Participants with below-average stress (Case Studies 3A and 3B) showed above-average levels of language performance in the TOEIC test, while participants with above-average stress (Case Studies 2 and 4) show below-average language performance in the TOEIC test.

The relationship between stress and language proficiency (as measured by TOEIC score) was investigated further in Chapter 7 using statistical methods. Figure 28 displays the results if a straight-line correlation is hypothesised, while Figure 29 displays the results for a correlation curving around a single point. The straight-line correlation predicted by Horwitz (1986) produced a marginally significant correlation, however the curved correlation was a

better fit for the data from this study and produced a result that was statistically significant with a p value of 0.04. The curve also matched the inverted U similar to that predicted by Hebb (1955), and which is similar to more recent research (Sapolsky, 2015), however the curve was skewed with no evidence of diminished language performance at lower stress levels. Not too much should be read into this curve for reasons to be discussed below.

When the participants' TOEIC scores from a year earlier are considered, the contaminating effect of including prior learning in a comparison with current stress becomes clear. In all four of the case studies mentioned above (2, 3A, 3B, and 4) the participants exhibited above-average TOEIC scores in the test a year earlier, and this earlier learning influences the current TOEIC scores. Considering just language performance without adjusting for prior language learning may be the reason why earlier studies found evidence for a straight-line negative correlation between language proficiency and stress. In Chapter 2 this point was discussed as a hypothetical weakness in previous studies such as those of Horwitz, et al. (1986). In Horwitz, et al. (1986) it was assumed that learners had no prior knowledge of the language, and that their language performance in the class was indicative of all of their language learning, despite the study including Spanish classes in a state in the US where Spanish is widely spoken. The case studies presented here provide evidence that failing to consider prior learning may, at least in part, explain the reason for the straight-line correlations found in earlier LLA research. Despite the straight-line pattern predicted by Horwitz (1986) showing a marginal correlation, the inverted U pattern predicted by Hebb (1955) and Sapolsky (2015) showed stronger statistical significance, suggesting that it is more likely to be accurate.

It should be noted that the statistically significant inverted-U curve linking language proficiency and stress should not be misinterpreted as representing a modified form of LLA theory. The reasons for this type of relationship between stress and performance are adequately explained by Hebb (1955) and Sapolsky (2015) without the need for LLA.

Next the relationship between stress and language learning (as measured by improvement in TOEIC score over a year) was investigated. These results are displayed in Figures 31 and 32, showing the straight-line versus curved line correlations, respectively. While neither pattern produced a statistically significant result the curved line is a better fit for the data. The lack of a statistically significant correlation may be an artefact of the small sample size and the inherent difficulty of trying to assess the impact of a single variable, stress, on a complex phenomenon such as language learning. Alternatively, stress may have varied considerably over the year of study, and the two weeks of data gathered in this study may not be presentative of participants' stress levels over the year.

What can be stated is that in both cases the straight-line model proposed by LLA theorists produced a weaker fit for the data than the curved line inverted U model. The inverted U model is also supported by the literature and the data presented in the case studies, with both high and low stress cases exhibiting lower levels of language learning.

While the results of this section were not as conclusive as hoped, the data from both the case studies and the statistical data tend to suggest that the inverted U model probably holds in the case of language learning. If this is the case then it has serious implications for LLA theory. As discussed in Chapter 2, the straight-line negative correlation between stress and learning proposed by LLA theorists suggests that peak learning happens at low stress and performance diminishes as stress increases. The inverted U model proposed by Hebb (1995) and Sapolsky (2015) suggests that learning peaks at mid-range stress, with learning declining as stress either increases or decreases away from this peak. If the inverted-U model is correct then, and if the goal is to maximise learning, then rather than aiming for low stress in the classroom the goal should be to aim for stress at the very upper edge of the low stress range (26 to 50). The problem with this is that, as discussed in the previous section, the data suggests that the language learning environment is not the primary determinant of stress, but rather a moderator of stress. Students do not enter the classroom as blank slates with the environment setting their stress level, but rather enter with pre-existing stress, which the environment has a role in moderating up or down.

The question therefore becomes what goal the teacher is pursuing. It is maximising learning, or reducing stress? The negative effects of stress reduction on high-proficiency individuals can possibly be seen in Case Studies 3A and 3B. These individuals entered university with TOEIC scores that were well above average, and what may have happened was that they were placed in classes where material was aimed at the average student. Csikzentmihalyi's (1997) flow model would predict that the combination of high proficiency and comparatively low challenge level would produce the boredom and resting stress levels seen in these cases during English classes. When read in combination with Hebb (1955) this low stress level explains the diminished improvement in TOEIC scores seen in these cases. However, Csikszentmihalyi's (1997) model equally applies to low proficiency cases, such as Case Study 4, where below-average initial proficiency combined with a relatively high level of challenge may be anxiety-provoking. There are proponents of resolving this issue by separating students into different streams or tracks depending on their level of ability.

Johnston and Wildly (2016) conducted a meta-analysis of the research on streaming/tracking, considering the social, psychological, and academic outcomes of this practice from seventy-one studies on the subject. Johnson and Wildly conclude that

streaming/tracking, "… is contentious and often contested, the literature generally shows that streaming impacts negatively on student learning outcomes." (2016, p. 1). Therefore the research suggests that streaming/tracking is not an effective solution to this issue as it has unintended consequences that negatively impact students. However, the reason that streaming may have proven unsuccessful is that streaming focuses on performance, not on stress. The impact of stress on learning is particularly evident in the cases studies where clinical distress was evident. In Case Studies 1 and 2 the participants entered university with above-average TOEIC scores and evidenced caseness and probable clinical distress that may have been associated with their TOEIC scores slipping to below average over the course of a year. Shifting these distressed students to a different stream would not necessarily address the problem of their distress. It might hypothetically also worsen distress because of the negative associations with being shifted to a "lower" stream.

Perhaps most importantly is the evidence in this study that distress was global in cases where clinical distress was evident, occurring across all environments. The mitigating effects of classroom environment cannot be dismissed, however the evidence suggests that the primary source of distress is internal. What this suggests is that if educators are seriously concerned about students' performance and learning, then given the global effect of distress on performance and learning, measures to safeguard students' mental health should be a primary area of concern.

This raises the question of how prevalent clinical distress is in student populations. The case studies presented contain a disproportionate number of individuals who may be displaying caseness. Case Studies 1 and 2, and potentially 4, may exhibit clinical levels of distress. The validation data in the manual for the GHQ-28 suggests that in an average population of university students in Japan 13.6% of should test above the caseness threshold (Goldberg, 2013). Five out of thirty-four (14.71%) of participants in this study showed evidence of prolonged caseness. The percentage of participants showing clinical distress in this study is a little above the average, but close to what would be expected. The prevalence of caseness is sufficiently common that the effects of clinical distress on language learning and language proficiency are factors that should be considered in this discussion and were the subject of Research Question 3.

8.3. Research Question 3

This section will discuss the findings relating to clinical distress and its effects on language proficiency and language learning. Both current LLA theory and the models of other stress theorists, such as Hebb (1995) and Sapolsky (2015), agree that individuals experiencing distress should show signs of diminished language performance. What was primarily of

interest in this section was whether there was also evidence of diminished language learning. This is an area of uncertainty since language learning is an endeavour that takes years, yet stress in this study was only measured over a period of two weeks, representing a relatively short period. It could not be assumed that distress had been present for a protracted period before this relatively brief period of monitoring.

There are two notable features in these cases. The first is the global nature of the distress experienced by participants who scored above caseness on the GHQ-28. Medium to high levels of stress were present in all environments. The patterns of stress are different in Case Study 1 and 2, with Case Study 1 showing balancing periods of low stress that resulted in a deceptively low average stress rating. Case Study 2 was typified by fewer periods of resting stress and more consistently elevated stress readings. However, despite these differences distress was evident across all environments to such an extent that it might be characterised as global in nature.

The second feature that is noteworthy is that in both Case Study 1 and Case Study 2 the participants exhibited both below average TOEIC scores in the latest TOEIC test. Both participants also exhibited below average improvement in their TOEIC scores, which in this study was taken as evidence of below average language learning. This finding comes with the caveat that distress may have impaired test performance during the participant's latest attempt at the TOEIC, and thus created the appearance of impaired language learning where there is none. This would be a fair critique of the methodology in this study however it is one that could be levelled at any assessment of language performance. Clinical levels of distress are likely to interfere with performance on any assessment. As such the term "language learning" is used with the understanding that the difference may be an artefact of distress rather than a lack of learning. In defence of this methodology all that can be stated is that this methodology is at least as valid, and arguably more so, than the methodology used in previous studies into LLA. It is not reasonable to generalise excessively from these two findings, but the findings are in line with both LLA theory and other stress theorists.

In Chapter 7 Figure 31 suggests that high stress readings were associated with lower levels of language learning. Interestingly in Figure 31 peak performance occurs at closer to 40 on the stress scale used for this study, rather than at the exact mid-point as suggested in Hebb (1955). This difference may be a result of the small sample size in this study, or differences in the measures used, but may be an area for further investigation.

As a proviso to the results presented in Chapter 7, average stress scores were used, and as noted in relation to Case Study 1 these may not be accurate reflections of the degree or frequency of distress. As was discussed earlier the quality and quantity of data from

wearable devices permits new and different types of analysis, however at present the analytical tools necessary for these types of analysis are not available. The tools may evolve in response to further research using these devices. Alternate forms of analysis were attempted, but the sheer quantity of data surpassed this researcher's skills at database management.

As a concluding note regarding Research Question 3 this agreement between LLA theory, other stress theorists, and the findings in the case studies for moderate to high stress participants should not be read as a confirmation of LLA theory in general. Diminished performance is predicted at high stress levels, such as was evident in individuals who exhibit caseness, however the key difference is in the question of where peak performance occurs, and what occurs at lower levels of stress. As discussed in Chapter 2, the FLCAS predicts a straight-line negative correlation between stress and language performance and learning, in which low stress should produce peak performance. The next section will examine what the evidence from the case studies suggests about the accuracy of the model presented by the FLCAS.

8.4. Research Question 4

This section is concerned with what the evidence suggested about the accuracy of the FLCAS as a predictor of clinical distress, stress, language performance, and language learning. The FLCAS predicts a straight-line negative correlation between stress and language performance and learning, therefore low stress should be associated with a markedly high level of language performance and language learning, and high stress should be associated with markedly diminished levels of language performance and language learning. In this study the participants' TOEIC test results were used as indicators of language performance and their relative improvement over a year on the TOEIC test was used as evidence of language learning.

In Case Study 1, despite markedly diminished language learning and performance Participant 12's FLCAS score was only slightly below average. Participant 12 was experiencing clinical levels of distress, primarily in the area of somatic symptoms but with evidence of frequent moments of high stress, possibly indicative of anxiety, which the FLCAS score did not predict. While Participant 12's stress levels were also slightly below average a more detailed analysis of the stress data revealed that this average was a result of highly variable stress readings across all contexts, which masked periods of high stress.

In Case Study 2 the slightly above average scores on the FLCAS did not predict clinical levels of distress, nor did they predict the degree of reduced language performance or language learning. An argument might be made that the slightly above average FLCAS

scores seems to correlate with Participant 10's slightly above average stress scores in English language learning classes. However, these slightly above average stress scores are not a reliable indicator of the presence of LLA or some other persistent anxiety related to language learning classes as they are artefacts of daily variations in stress. Participant 10's scores vary widely from day to day, showing more distress in non-language classes on some days, and more distress in language classes on other days. As it happened, the averages came out slightly higher for language classes because more English classes happened to occur on days where Participant 10's stress readings were higher. This should not be taken as an indication of the existence of some sort of specific anxiety in English classes. The difference in averages is so small as to be insignificant and is arguably most probably the result of factors unrelated to the content of those classes. Nor is this stress anticipatory, as an examination of the contextualising stress data shows that these elevated stress levels persist after class, as well as before class.

In Case Study 3A, Participant 31's score on the FLCAS was a little lower than average at 96 against an average of 117.7 and this covaried with their higher language performance (Participant 31's TOEIC score was 705, as compared to the average participant's score of 641.18). Participant 31's average stress is likewise below average, as was their FLCAS score. Case Study 3A provides one data point that suggests that the FLCAS may have some validity in predicting language performance and stress levels in non-clinical cases. However, the FLCAS did not predict the degree to which language learning was diminished in this case, with Participant 31 showing no improvement in their TOEIC score over a year.

In Case Study 3B, Participant 28 exhibited a similar pattern to 3A, with a slightly below-average FLCAS score (93 as compared to an average of 117.7), and an above-average level of language proficiency. As in Case Study 3A the above-average score on the TOEIC test was a result of starting university with an above-average TOEIC score a year earlier, and Participant 28 showed a markedly below-average improvement (95 points against an average of 154.12) in their TOEIC score over the intervening year. Assuming the straight-line negative correlation predicted in Horwitz (1986), the difference in the FLCAS score does not accurately model the degree of diminished learning evident in Case Study 3B.

Again, it should be noted that there are problems with assuming that any test accurately reflects an individual's knowledge and learning. It may well be possible that Participant 31 learned a great deal in the intervening year, but that it simply fell outside of the scope of the test. However, it should be noted that the FLCAS was developed using this

same test-based methodology, and thus it is fair to assess the FLCAS in the same manner in this study.

In Case Study 4 Participant 30 scored 122 on the FLCAS against an average of 117.7, which is only a little above average. However, Participant 30's language performance seemed to be quite far below average, with a TOEIC score of 520 on their university entrance TOEIC and the TOEIC year later, compared to an average participant score of 641.18 in the latest TOEIC results. In this case the FLCAS does not appear to predict language proficiency. Neither did the FLCAS appear to predict Participant 30's lack of improvement in their TOIEC score, which was used as a metric for language learning, nor their high levels of stress.

In Chapter 7 the FLCAS was analysed for possible correlations, but showed no statistically significant relationships with total GHQ-28 scores (Figure 21), the GHQ-28's anxiety subscale (Figure 22), average stress (Figure 23), or stress in English classes (Figure 24).

Overall, the evidence seems to suggest that the FLCAS, and the associated straight-line negative correlation between stress and language learning, are not supported. The FLCAS seems to loosely predict language proficiency in terms of broadly indicating above or below average performance, however the FLCAS did not show evidence of being a good predictor of TOEIC scores. It may be, as hypothesised in Chapter 2, that the FLCAS is measuring some sort of propensity in individuals towards negative attributions. This may be exerting a mild long-term influence on language learning. This should not be confused or conflated with stress, distress, or anxiety.

The following section will examine what the case studies suggest regarding the use of wearable devices in stress research.

8.5. Research Question 5

As discussed in Chapter 3, the use of wearable devices as research instruments such as the *Garmin Vivosmart 3/4* is a relatively new area, where consensus is still emerging regarding how these sorts of devices can and should be used. In addition to the expected findings there were several areas where the data gathered in this study suggested other areas where these devices might be useful. Of equal importance though were the potential pitfalls inherent in using these devices, such as the issues of poorly fitting devices seen in Case Study 3A and probably 3B. As evident in the case studies in this section the sheer quantity and level of detail in the data creates problems using traditional analytic approaches

associated with quantitative data, such as averaging. The discussion in this section will start with this issue.

8.5.1. Average (Mean) Data and Wearable Devices

As has been shown in the case studies the quantity and level of detail in the data available from wearable devices offers potentially unprecedented research opportunities. As discussed earlier in Chapter 4 this is not just more data but is also qualitatively different. This difference is evidenced in the problems with average (mean) data in Case Study 1 (Participant 12).

Examining the detailed stress graphs in Table 7, Participant 12's average stress levels seem quite misleading. Case Study 1 was included as an exemplar of "Above Caseness, Below Average Stress", with participant 12's average stress level being 39.56 compared to an average participant stress level of 45.09. However, on examining the stress data Participant 12's stress levels seem highly variable across all contexts. Participant 12's below average stress score is misleading, in that it is comprised of a large number of medium (51 to 75) to high (76 to 100) readings counterbalanced by resting stress (0 to 25). On looking at the average stress at the beginning of this case study the impression is created that Participant 12, on average, experienced low stress in the 26 to 50 range. However, the data in Table 7 shows that Participant 12's stress levels varied widely, often shifting rapidly from resting to medium or high stress.

The degree of variability across contexts shown in the wearable device data is something that would have been impossible to capture in the past. Laboratory studies could capture this sort of detail, but only in one context, and at the risk of contaminating the measurements as a result of the artificial setting. Diary, questionnaire, and interview approaches could not obtain this sort of fine detail without imposing an undue burden on research participants. This would also potentially contaminate their results by asking participants to stop every few minutes to respond to reflect on their stress levels, which could introduce an introspective element that might change the participant's stress levels. These data were gathered in a naturalistic fashion, without interruption to the participants' daily routines or imposing an undue burden on the participant.

When subjected to detailed analysis the wearable devices deliver data that is meaningful, contextualised, and nuanced. This demonstrates the utility of wearable devices in stress research. However, this comes with the caveat that this case also demonstrates how easy it is to misinterpret or misrepresent that data, by relying exclusively on traditional quantitative approaches to analysing the sort of data presented by wearable devices.

The degree of variation present in the stress data in Case Study 1 that caused the difficulties with average data, raises the issue that much of the historical research into stress is based on either episodic measurement, decontextualised laboratory research, or other research that may have missed this variation. This will be the subject of the next point for discussion.

8.5.2. Persistence of Distress in Clinical Cases

The detailed and contextualised longitudinal data provided by the *Garmin Vivosmart 3/4* raises some interesting possibilities regarding the notion of persistence and constancy in clinical cases. The DSM-5 repeatedly uses the word persistent 457 times (and persistently a further 32 times), and the word consistent 72 times (and consistently a further 22 times), thus many diagnoses hinge on these notions (The American Psychiatric Association, 2013). In practice, however, as the data from Participants 12 (Case Study 1) and 10 (Case Study 2) shows, there is far more variability day-to-day than these notions may suggest. This is not to imply that there is simplistic notion that persistence implies that individuals can't have good days, however establishing clinical guidelines is difficult without data. Wearable devices may offer a means of gathering data to support more clear definitions in terms of concepts like persistence and constancy in clinical cases.

Further to this point, as discussed earlier in Chapter 4 the role of stress in psychological dysfunction has thus far been difficult to pinpoint. Part of the reason for this may be the quantity and quality of stress data available to researchers. Case Study 4 is an interesting example as Participant 30's stress readings, sleep data, and diminished improvement in TOEIC their score are similar to Case Studies 1 and 2, where clinical distress were evident. Despite this, Participant 30 scored below the caseness threshold on the GHQ-28. This may be evidence of elevated stress with a lack caseness, a clinical condition in the prodromal phase, or a case where the participant was clinically distressed but "faking good" on the GHQ-28. All three of these possibilities are of potential interest for further research. Ethical restrictions placed on the design of this research did not allow for contacting the participant after the study to clarify the issue. However, this again demonstrates the potential utility of wearable devices in stress and associated research. One of the areas that deserves special mention is the issue of sleep and stress.

8.5.3. Sleep and Stress

The sleep data in Case Study 4 was useful in identifying a possible discrepancy between the GHQ-28 results and the other data. However, wearable devices still require considerable improvement before the sleep data can be regarded as reliable. As discussed earlier in Case Studies 2 and 3B the incorrect labelling of sleep data and inability to identify periods of sleep

that do not confirm to routine, such as napping, are important limitations in using wearable devices to investigate the link between sleep and stress.

The importance of sleep as a variable in language learning, emotional regulation, and health was discussed earlier in Chapter 6. While the mention here is quite brief, the implications of this limitation in current wearable technology may be profound or might be inconsequential. The implications of napping are an area that has proven difficult to investigate in the past, precisely because it represents sleep that falls outside of routine sleep patterns. Napping proved difficult to identify in the current study, but wearable technology is still very new, and if it advances to the point where it can identify napping this might be a valuable contribution to future research.

8.5.4. Stress and Routine

Regarding the issue of routine, the case studies suggest that wearable devices present promising research opportunities. Participant 31's (Case Study 3A) data following the public holiday provides an example of the effects of a break in routine.

It is common knowledge that some holidays, such as Christmas, are associated with higher levels of mortality. Wallet, et al. (2017) analysed eight years (2006 to 2013) of hospital admission data on myocardial infarctions (heart attacks) from the national Swedish database (SWEDEHEART) and found a connection between the Christmas and New Year public holidays and increased heart attacks. Wallet, et al. (2017) attributed this increase to psychosocial stress, and changes in routine. The degree of change in routine does not have to be large. Costa-Font, et al.'s (2021) review of the literature surrounding the 1-hour shift in routine from daylight savings, suggests negative impacts such as higher risks of car accidents, workplace injuries, heart attacks, depressive symptoms, stress.

While intuitively it feels like public holidays should reduce stress, it may be that the cost of breaks in routine caused by short-term changes such as public holidays may outweigh the stress reduction benefits of the day off. Past research, such as that by Wallet, et al. (2017) and Costa-Font (2021), has necessarily focused on severe adverse phenomena such as heart attacks, car accidents and workplace injuries as metrics for the effects of stress. However, this data is not contextualised and longitudinal, and is often retrospective, measuring stress only after the negative consequence. This leads to the next point for discussion, namely the metaphysics of stress.

8.5.5. The Metaphysics of Stress

Metaphysics concerns itself with determining what is and what it is like. As discussed at length in Chapter 2 and 3 the history of stress research is rooted in metaphysical uncertainty

regarding the nature of stress. While this research has settled on one definition of stress, namely resultant stress as measured by HRV, Sapolsky (2015) notes the lack of consensus regarding what stress is, or is not, and what it is like. A factor in this lack of consensus presents difficulties in gathering data on stress.

The problems evident in this lack of data are demonstrated in the data presented in these case studies. The notion that academic environments are more stressful than day-to-day life is an underlying unproven assumption in both LLA theory and in similar theories such as academic anxiety, as discussed in Chapter 2. What the data from these case studies suggests is that when the data is contextualised against daily life, it seems that learning environments may be less stressful. The importance of this finding cannot be understated, as it presents a fundamental challenge to many assumptions surrounding education and educational environments.

This evidence has implications not only for education, but also for stress research in general, as it presents data that suggests that there may be more areas where foundational assumptions about stress may need to be reviewed. The lack of definitional consensus that still surrounds stress may, at least in part, be a product of researchers in different areas operating under different basic unproven assumptions about the nature of stress.

It should be noted that this is not advocating for a position where resultant stress is measured solely in terms of the biomedical model, using HRV as measured by wearable devices. As discussed in Chapter 2 the concept of stress is complex and multi-dimensional. However, wearable devices present researchers into stress with an empirically defensible touchstone that could be used in addition to other measures in order to facilitate multidisciplinary discourse around stress. Sapolsky's (2015) advice to, "let a thousand flowers bloom" (p. 1348) is a beautiful image. However, it is problematic in that it does tend to perpetuate the primacy of the biomedical model of stress and may limit contributions from other fields and other measures of stress, by perpetuating incompatible definitions of stress. As noted in Chapter 3 the use of wearable devices is becoming more common, which presents stress researchers with opportunities to incorporate wearable stress data, in addition to their preferred stress measures, as a means of allowing interdisciplinary comparability between results to facilitate discourse. The incorporation of stress data from wearable devices into findings from other research may present a possible way to help more clearly delineate different, but equally valid, meanings of the word stress.

In particular, the area of subclinical stress and the distinction between stress and distress was an area where relatively little literature could be found, while there is a great deal of literature on clinical stress and anxiety. However, texts on clinical stress such as the

DSM-5 make references to concepts such as excessive stress, developmentally appropriate stress, and other notions regarding differences from normal subclinical stress, for which little basis could be found in the literature. Instead, these notions are left to the clinician's judgement. This element of subjectivity may be at the root of many of the problems in stress and anxiety research.

Another area where wearable devices may be of profound importance is in the study of stress and the genesis of psychological disorders. Past studies into the genesis of psychological disorders have faced numerous difficulties in tracing the path of the disorder, relying primarily on self-reported recollections from people who are presently suffering from clinical levels of distress. However, even at this point there are many individuals who have been wearing wearable devices consistently for years, providing a pool of data that could offer potentially paradigm-altering insights into the genesis of psychological conditions.

Further to this point, the identification of patterns of stress or sleep that are associated with the development of clinical distress opens the potential for early interventions before the individual reaches the point where clinical intervention is required.

This raises the issue of definitions of harm. As noted earlier in Chapter 4 when discussing stress, the definition of harm tends to be defined in clinical, typically psychiatric, terms. The lack of definitional agreement and common terms of reference in stress research, makes it difficult for researchers into subclinical stress to demonstrate evidence of harm and dysfunction at subclinical levels. The evidence in this study suggests that diminished language learning and language performance, begins to take place at stress levels well below those found in the case studies where clinical distress was evident. It could be argued that this diminished learning and academic performance is a form of harm that affects the individual's long-term academic prospects and needs to be taken seriously and addressed.

The data presented in this study suggests that wearable devices may present researchers with opportunities for an evidence-driven re-examination of the foundational concepts around stress. The inclusion of contextualising longitudinal data, that allowed comparisons between stress readings and patterns in particular, raises questions about the validity of certain core assumptions regarding stress and context.

However, the data presented in this study is deficient in many regards. The low response rate from the wearable devices led to a small sample size. The richness and detail of the data made analysis with traditional tools difficult. The question therefore becomes what can be concluded with relative confidence, and what this research has contributed to the field. This will be the subject of final chapter, which will discuss what conclusions can be reached.

Chapter 9: Conclusions

In this chapter the implications of the findings presented in the previous chapter will be discussed under three general headings, namely education (9.1), stress research (9.2), and the use of wearable devices in future research (9.3). While the findings were not as statistically significant as hoped, the implications of what was suggested in the data are arguably paradigm-altering, and important and interesting for reasons that will be explored in this section.

When reading what follows, it is important to recall the discussion in Chapter 2, where the previous research into LLA was presented, and where it was shown that the previous research into LLA was founded on methodologies that were questionable. Therefore, while the evidence presented in this study is not as robust as hoped, this is not a situation of competing well-supported theories. Rather the historical evidence for LLA is based on the questionable use of psychometric tests (Dunkel, 1947), measures that conflated language performance with language learning without considering prior learning (Horwitz, 1986), and similar problematic methodologies. This historical research is often based on unclear and conflicting definitions and unproven assumptions. Many of these problematic issues are products of the state of knowledge at the time, and while this research may contradict past research, it is with an awareness that their contributions spurred this research and without their contributions this research would have been impossible. However, respect for previous researchers and their work should not stop theories from being revised, improved, and even rejected as incorrect.

9.1. Educational Implications

A prime example of one of these unproven assumptions is the notion of language learning environments being more stressful, which was traced back as far as Dunkel (1947) in this research, although has probably existed for far longer than that.

The contextualised data in this investigation of LLA shows that for the majority of participants lower levels of stress were present in language learning environments, and academic environments in general, when compared with most participants' free time. This points to an area where foundational assumptions may be flawed. Over time this unproven assumption about learning environments being more stressful has been parlayed into the existence of some sort of special stressor, hypothesised as LLA in the case of language learning, existing in the learning environment. This unproven assumption then magnified until the work of Horwitz and Horwitz, et al. in 1986, where an attempt was made to prove that this special stressor is a predictor of language learning.

This is where the contextualising data presented in this study is potentially paradigm-altering and incredibly important. As discussed in the previous chapter the case studies and statistical data pointed towards learning environments (language or otherwise) being a moderator of stress, not the primary determinant. The data suggested that most of the stress experienced by participants was attributable to their own internal processes and everyday lives. Day to day variations shown in the case studies, possibly partially in response to sleep and changes in routine, showed the degree to which resultant stress could vary from day to day. In the majority of participants environment did not change their stress averages sufficiently to alter their general stress category (resting, low, moderate, or high). When the data was examined most closely in the case studies, stress varied widely and rapidly within contexts depending on what seemed to be the participants' perceptions and reactions to stressors.

LLA theory, which is based on the notion that environment exerts a sufficiently strong influence on resultant stress to substantially alter participants' performance, did not seem to be supported. Note once again that the notion in LLA of a situation-specific stressor that exists in the learning is an unproven assumption. This is not the case of this study contradicting the results of prior research, but rather providing evidence where there was none before. This is an example where a previously unproven assumption has been tested and the evidence suggests that the assumption is unfounded. The removal of this assumption has a profound effect on LLA theory, toppling the house of cards on which the theory is built. At this point, as suggested by Occam's Razor, the assumption of an additional situation-specific stressor becomes an unnecessary complication that is not evinced in the data. This should not be misrepresented as stress having no role in language learning.

The data did suggest a link between stress and academic performance and learning, although not in the straight-line negative correlation suggested by LLA theory. Instead, the relationship between stress, language performance, and language learning was a better fit for the inverted U shape predicted by Hebb (1955) and Sapolsky (2015), although with a slighter lower point of optimal performance than was predicted than Hebb (1955). Given the suggested link between stress and learning, and the pervasive nature of stress it is reasonable to postulate based on the evidence that the effects seen in language learning would also affect other subjects similarly, although this is an area that requires further research.

What are the educational implications of the findings of this study? What follows must be preceded with the important caveat that the data in this study suggested that at the

extremes of very high and very low stress the data was more tightly clustered, possibly proposing a stronger relationship between stress and learning. However, in the mid-range scores varied considerably. Therefore, stress is not some magic bullet that guarantees enhanced or diminished performance, but rather is just one of many factors that influence educational outcomes. While the data suggests that the effect may be stronger at extremes, this is based on a very small sample size and requires further investigation.

Having issued that warning, the data implies that pursuing a low stress classroom environment, as advocated in LLA theory, is unlikely to substantially alter learner's total resultant stress. The data in this study suggest that learning environment did not exert a strong influence on total resultant stress. Variations in sleep and routine seemed to exert more noticeable effects on resultant stress; and educating students about the importance of maintaining regular sleep patterns is an area that the literature suggests might be a profitable use of educators' time.

The data also suggested the inverted-U model of stress and learning is probably correct, and that levels of stress substantially above or below the optimal point could negatively impact learning. Regarding stress above the optimal point, the first thing to note is that this does not necessarily mean clinical distress. This speaks to the debates around the notion of "harm" in relation to stress, and the prevalence of the psychiatric model and definitions. It could be argued that diminished learning is a form of harm that could negatively impact on learners' lives and careers over the long term, in numerous ways ranging from the impact of lower performance on self-esteem to diminished educational and financial opportunities. While this harm is stronger at higher levels of stress, this is a continuum where harm is not limited to just the extremes, although further research is required to model the optimal range. In this study the data suggested that the optimal point may be lower than depicted in models such as Hebb (1955), not occurring at the mid-point of 50, but rather closer to 40. Identifying where peak performance occurs and the ideal range was not possible in this study because of the small sample size, but is an area that the evidence in this study suggests may need further research.

This broadened definition of harm should not be read as diminishing the importance of students experiencing high levels of distress, such as is often associated with clinical conditions (The American Psychiatric Association, 2013). It should be noted that the evidence presented in this study suggests that these individuals may show lower levels of performance and learning. The importance of correctly identifying the cause of this diminished performance and learning is of critical importance. Practices such as streaming/tracking where these students are moved to lower-level classes are

contraindicated by the findings in this study, which suggest that changing the learning environment will not substantially alter their levels of distress. This notion is supported by the meta-analysis conducted by Johnston and Wildly (2016) that showed negative outcomes from streaming in Australian schools. A missing piece of the puzzle may be that shuffling students experiencing distress, and possibly clinically distress, into lower-level classes may not only be harmful to the self-esteem and self-image of already vulnerable students. Furthermore, the data in this study suggest that the change in learning environment is unlikely to have any substantially beneficial effects as the data suggests that learning environment does not exert a strong effect on stress. What the data suggests might be a more productive approach is an engagement with the source of their stress, be it clinical or subclinical, and steps to address the stress in their lives, preferably by a qualified professional.

The role of high stress, in the form of distress and anxiety, is an area where both LLA and other stress theories agree, however the evidence in this study suggested that low stress may also diminish language learning. This is an area where LLA is at odds with other stress theorists, and this difference has important implications. This is an area where the findings may seem to be mixed, with Figure 29 in Chapter 7 showing a skewed inverted-U with little evidence of lower levels of performance at low stress levels. However as discussed in Chapter 7 and 8 the language performance figures are based on historical learning that occurred before the participants entered university. When language learning (as reflected by improvement in TOEIC scores) is considered, the evidence in the case studies and statistics suggests that low stress is linked with lower levels of improvement in TOEIC scores. This link was not statistically significant, but this may be a result of the mismatch between the relatively short two-week period of stress monitoring being compared against a full year of learning. The short-term stress monitoring was motivated by ethical concerns about potential harms in stress monitoring, for which no evidence was found in this study. This opens the possibility of future research where the learning period and stress monitoring period could be more closely matched, and which might produce a clearer link. However, the evidence available does correspond with theorists such as Hebb (1955) and Sapolsky (2015), suggesting that low stress is associated with diminished language learning.

In the case studies this pattern of low stress and diminished learning affected participants who entered university with TOEIC scores well above the average. The implications of this pattern tie into important debates around the treatment of students who perform above average. The phenomenon of seeking to bring those who perform above average down to the average level can be seen in many countries around the world, such as “Tall Poppy Syndrome” in Australia, the “Law of Jante” in Nordic countries, and in Japan is

epitomised in the phrase “the nail that sticks up gets hammered down”. In cultural anthropology the phenomenon is referred to as a levelling mechanism (Eller, 2020).

According to Csikszentmihalyi (1975) high proficiency combined with low challenge results in boredom, which would tend to match what was seen in this study. Participants with high levels of proficiency appeared to demonstrate low stress levels, possibly indicative of boredom, and an associated low level of language improvement. It was probably not the intention of LLA theorists to propose a levelling mechanism, however the evidence in this study suggests that might be what is happening. Participants with higher levels of starting proficiency experienced low levels of stress and this seemed to be associated with lower levels of learning. If the broadened definition of harm proposed earlier in this section is applied, then the push in LLA theory towards lowering stress might be characterised as harmful. There may seem to be a contradiction here between earlier statements that environment did not substantially alter stress categories for most participants and this notion that there might be real harm to these students.

While the data is sparse, with just two examples of high starting proficiency, it seems that in low stress participants, the effects of environment were more notable because of the low values. In Case Study 3A average English class stress was 19.25, while average free time stress was 25.87. In Case Study 3B average English class stress was 18.05, while average free time stress was 30.02. In both Case Study 3A and 3B the absolute difference in stress measurements is small, 6.62 points in Case Study 3A and 11.97 points in Case Study 3B, but when considered as percentage changes the changes are large, 25.59% and 39.87% respectively. Therefore, it might be argued that while generally environment did not exert a marked influence on stress in most cases, very low stress cases may be special areas for concern.

The theories of Csikszentmihalyi (1975) and Hebb (1955) suggest that a degree of stress may be desirable to achieve peak performance, and the drive towards low stress classroom environments, while well-intentioned, may be detrimental to students’ learning. However, as stated earlier, in this study in most cases the learning environment was not the main factor in participants’ total average resultant stress. The contextualising data from participants’ free time suggested that if peak learning is the objective, then stress management education and the identification of students experiencing stress at either extreme (high or low) is advised. This may seem like a controversial proposition, however the use of Intelligence Quotient testing is a common practice, despite its many problems (Contrada & Baum, 2011). The data from this study suggests that as wearable devices become cheaper and more common, it may be educationally desirable to include them as a

means of allowing students to monitor their stress levels, and when accompanied by education on how to interpret the stress scores this data may help students to manage their own stress, or to seek help if they are unable to do so on their own. This point does not wish to focus exclusively on clinical distress, but according to Goldberg (2013), 13.6% of an average student population are likely to score above caseness. Given this information it seems like a reasonable and defensible proposition that stress management should be viewed as important as sports education, and that the phrase *mens sana in corpore sano* (a healthy mind in a healthy body) is too often used to justify the existence of sports education classes, while ignoring the need for mental health classes and mental health support in educational institutions. This study suggests that providing students with the means to monitor their own stress may help to prevent harm, as well as providing both parents and educators with important information about students' wellbeing.

This leads to the next section, exploring the implications of this study for stress research.

9.2. Stress Research

This research started with what seemed like a relatively simple initiating question. Why was there a mismatch between the straight-line negative relationship between stress and performance depicted in LLA and the inverted U proposed in other stress theories such as Hebb (1955) and supported in more recent work in neurology (Sapolsky, 2015)? The researcher naively imagined that they had a firm grasp on the meaning of words such as stress and anxiety, but in reading about this topic discovered the depth of confusion and debate about the meanings of these words. In this thesis some of these meanings have been explored, such as the distinction between different types of stress and distress, whether clinical or sub-clinical, and the associated debates around the meaning of other words such as "harm". As the researcher read more about stress and anxiety, the number of questions multiplied as the complexity of the debates around stress and anxiety became more fully evident.

In this thesis many of these debates have been touched on briefly, and equally many have been excluded. The researcher has chosen a single type of stress, resultant stress, supported by a single measure, HRV. The researcher is acutely aware that this is a simplification of an enormously complex and multi-faceted area of study that touches on debates that are tremendously important. Ultimately this is the problem faced by any researcher, the necessity of drawing a ring around the research in order present a coherent argument in support of an achievable research goal, without losing focus.

It is therefore with an awareness of the inherent irony that arguably the most important conclusion from this research was regarding the criticality of context, outside of the primary area of investigation. What was evident in the case studies was the manner in which context altered the interpretation of all the surrounding data. The inclusion of contextualising data provided the means for comparing stress readings and patterns of stress between and within contexts. This inclusion of context has profound implications for the study of stress, not just within language education, but in general stress research.

Moreover, this study constitutes proof of concept regarding the feasibility of collecting contextualised stress data over time using wearable devices. As was discussed in Chapter 8 this allows the exploration of stress phenomena such as the impact of sleep and routine on stress, the notion of persistence of distress in clinical cases, and may provide a means of resolving the definitional debates around the meanings of the words stress and anxiety.

This is not the first investigation into stress using wearable devices, but the review of the literature suggests that it was one of relatively few that could be found at this time, and so doubtless mistakes were made that will be the source of criticism from later researchers. However, it is by making these mistakes that the state of knowledge progresses, and so the next section will discuss some of these mistakes and the implications of using wearable devices in research, so that hopefully these mistakes will not be repeated by future researchers.

9.3. The Use of Wearable Devices

In addition to contributing proof of concept regarding the importance of contextualising stress data, this research also explored some of the potential strengths and weaknesses of wearable devices as research tools.

As was discussed in Chapter 3, when this research started in 2018 the researcher attempted to find similar studies, but only able to locate a handful of studies into stress using wearable devices, and none could be found in the area of education and stress. This was worrying on multiple levels. The researcher was unsure whether this genuinely represented this study being novel, or whether the researcher was using the incorrect search terms and missing similar research. More recent meta-analyses, such as Pang, et al. (2019), show that while wearable devices are gaining acceptance as research devices, there are still very few studies using wearable devices, with Pang, et al. finding only nine qualifying studies for their meta-analysis. As of November 2021, no studies could be found using wearable devices to investigate stress in an educational environment. This makes this study novel and this is in and of itself a potential contribution to the body of knowledge.

Some of the mistakes made in this research seem obvious in retrospect. The problems with the *Garmin Vivosmart 3/4s* fitting too loosely to collect accurate readings on participants with smaller wrists was a result of buying only the regular sized devices, but affected data collection as can be seen in the incomplete data in Case Study 3A and possibly in Case Study 3B. In the case of the *Garmin Vivosmart 3/4* the issue seemed to be the size of the watch face. While the wristband fastened, the watch face did not sit snugly against thinner wrists, and this is the suspected source of the loss of a significant amount of data. The problem in addressing this issue is that the identity of participants may not be known, and budget limitations may mitigate against buying a large number of devices of various sizes, when some might not be used.

Issues of length of observation are a trickier area to navigate. The data in this study seems to suggest that the frequency of the phenomenon to be observed is important. If it is something that happens every day, then it may be that a week of data is sufficient for valid comparisons. However, if it is less frequent then a longer period is required so that sufficient instances of the phenomenon can be recorded to capture any potential variation. In this research there seemed to be merit in acquiring at least two weeks of data, as it allowed the researcher to see patterns in activity between the same day on different weeks. Again, this seems obvious in retrospect, with many people having routines like days for the gym or doing other regular exercise routines that would look like aberrations or outliers in the data without sufficient context. An error in this research was forgetting about public holidays, which may disrupt routines and regular patterns of behaviour for several days before or after.

While longer observation periods may yield more data, this is not a certainty as the literature suggests that longer periods of observation using wearable devices may result in lower response rates. The literature suggests that in long-term studies of over 6 months, there is a high degree of attrition in participation (Strath and Rowley, 2018). While this can theoretically be planned for by increasing the sample size, this may not be practical because of budgetary reasons or because of participant availability and motivation.

It has been theorised in this study that part of the reason for participant attrition may be because of concerns over privacy. Wearable devices may be considered unobtrusive in the sense of not impeding or interfering with research participants' normal daily activities, and not requiring much special effort from participants. However, as shown in the case studies, the level of detail in the data amounts to a narrative of the participant's life for the period of the study. In this sense wearable devices may be viewed as highly intrusive, and it would be understandable if participants were reticent to disclose this level of detail about their lives.

Awareness of this level of intrusiveness was why this study opted for the assessor single-blind approach, where the research was completely unable to identify individual participants. However, this also imposes limitations on the use of wearable devices. This created certain ethical issues where the researcher wanted to suggest that some participants should seek professional advice about their health (based on their GHQ-28 scores), but this was impossible because participants could not be identified. Less importantly, but still an issue, was that it was impossible to follow up on the non-return of wearable devices.

Wearable devices are a new technology, and while there is tremendous promise in this technology for researchers, there are also serious concerns about how this technology should be used. This is true of any new technology and should not be viewed as an absolute barrier to their use, however caution should be taken to step carefully in using this technology to avoid errors in so far as possible.

9.4. Closing Comments

In closing, this study has demonstrated the utility of wearable devices as tools for delivering contextualised longitudinal data about complex phenomena in a manner that was previously impossible. The exploration of LLA resulted in data that suggested that one of the foundational assumptions of LLA was not evident in the data. Without the existence of a situation-specific stressor that makes language learning environments more stressful than other environments, the core assumption on which LLA is built falls away. Evidence was also gathered that suggested further potential methodological weaknesses that should be addressed in future research into language learning, such as conflating learning and performance. The importance of separating out prior learning was demonstrated.

While the data was insufficient to produce a new model of language learning and stress, the detailed stress measurements from the wearable devices, when combined with the other research measures used, provided sufficient evidence to suggest that the inverted U model of stress proposed by theorists such as Hebb (1955) and Sapolsky (2015) probably holds for language learning.

In addition to the core findings this research suggests many potential uses for wearable devices in educational and psychological research, some of which have been suggested in this chapter. As always further research is required and encouraged, and this new technology opens up exciting avenues for research that may help to resolve questions that have previously defied analysis.

References

Akbar, F., Mark, G., Prausnitz, S., Warton, E. M., East, J. A., Moeller, M. F., ... & Lieu, T. A. (2021). Physician Stress During Electronic Health Record Inbox Work: In Situ Measurement With Wearable Sensors. *JMIR Medical Informatics*, 9(4), e24014. https://doi.org/10.2196/24014

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (5th ed.)*. https://doi.org/10.1176/appi.books.9780890425596

Anaya, L. S., Alsadoon, A., Costadopoulos, N., & Prasad, P. W. C. (2018). Ethical implications of user perceptions of wearable devices. *Science and Engineering Ethics*, 24(1), 1-28. https://doi.org/10.1007/s11948-017-9872-8

Andersen, T. O., Langstrup, H., & Lomborg, S. (2020). Experiences with wearable activity data during self-care by chronic heart patients: qualitative study. *Journal of Medical Internet Research*, 22(7), e15873. https://doi.org/10.2196/15873

Ando, S., Yamaguchi, S., Aoki, Y., & Thornicroft, G. (2013). Review of mental-health related stigma in Japan. *Psychiatry and Clinical Neurosciences*, 67(7), 471-482. https://doi.org/10.1111/pcn.12086

Andrade, C. (2019). The P value and statistical significance: misunderstandings, explanations, challenges, and alternatives. *Indian Journal of Psychological Medicine*, 41(3), 210-215.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1), 17-21.

Bados, A., Gómez-Benito, J., & Balaguer, G. (2010). The state-trait anxiety inventory, trait version: does it really measure anxiety?. *Journal of Personality Assessment*, 92(6), 560-567. https://doi.org/10.1080/00223891.2010.513295

Bailey, P., Daley, C. E., & Onwuegbuzie, A. J. (1999). Foreign language anxiety and learning style. *Foreign Language Annals*, 32(1), 63-76.

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology, 8*, 456. https://doi.org/10.3389/fpsyg.2017.00456

Barnes, L. L., Harp, D., & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger state-trait anxiety inventory. *Educational and Psychological Measurement, 62*(4), 603-618.

Bessette, A. (2007). TOEIC: Uses and Misuses. *Pool Gakuin Daigaku Kenkyu Kiyo*, 47, 35-45.

Black, D. W., & Grant, J. E. (2014). *DSM-5® guidebook: The essential companion to the diagnostic and statistical manual of mental disorders*. American Psychiatric Publishing.

Borovoy, A. (2008). Japan's hidden youths: Mainstreaming the emotionally distressed in Japan. *Culture, Medicine, and Psychiatry*, 32(4), 552-576. https://doi.org/10.1007/s11013-008-9106-2

Bravata, D. M., Smith-Spangler, C., Sundaram, V., Gienger, A. L., Lin, N., Lewis, R., Stave, C.D., Olkin, I., & Sirard, J. R. (2007). Using pedometers to increase physical activity and improve health: a systematic review. *JAMA*, 298(19), 2296-2304. https://doi.org/10.1001/jama.298.19.2296

Brown, H. D. (2000). *Principles of language learning and teaching (Vol. 4)*. New York: Longman.

Brown, J. D., Robson, G., & Rosenkjar, P. (2001). Personality, motivation, anxiety, strategies, and language proficiency of Japanese students. In Dornyei, Z. & Schmidt, R. (Eds.) *Motivation and second language acquisition* (Technical Report #23, p. 361-398). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Cadmus-Bertram, L. A., Marcus, B. H., Patterson, R. E., Parker, B. A., & Morey, B. L. (2015). Randomized trial of a Fitbit-based physical activity intervention for women. *American Journal of Preventive Medicine*, 49(3), 414-418.

Carmody, J., & Baer, R. A. (2008). Relationships between mindfulness practice and levels of mindfulness, medical and psychological symptoms and well-being in a mindfulness-based stress reduction program. *Journal of Behavioral Medicine*, 31(1), 23-33. https://doi.org/10.1007/s10865-007-9130-7

Casado, M. A., & Dereshiwsky, M. I. (2001). Foreign Language Anxiety of University Students. *College Student Journal*, 35(4).

Cassady, J. C. (2010). *Anxiety in schools: The causes, consequences, and solutions for academic anxieties (Vol. 2)*. Peter Lang.

Chastain, K. (1975). Affective and ability factors in second-language acquisition. *Language learning, 25*(1), 153-161. https://doi.org/10.1111/j.1467-1770.1975.tb00115.x

Cheff, R. (2018). *Compensating research participants: A survey of current practices in Toronto.* Wellesley Institute.

Chen, J., Choi, Y. J., & Sawada, Y. (2009). How is suicide different in Japan? *Japan and the world economy*, *21*(2), 140-150.

Chiang, J. J., Cole, S. W., Bower, J. E., Irwin, M. R., Taylor, S. E., Arevalo, J., & Fuligni, A. J. (2019). Daily interpersonal stress, sleep duration, and gene regulation during late

adolescence. *Psychoneuroendocrinology*, 103, 147-155. https://doi.org/10.1016/j.psyneuen.2018.11.026

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Cohen, S., Kessler, R. C., & Gordon, L. U. (1995). Strategies for measuring stress in studies of psychiatric and physical disorders. *Measuring stress: A guide for health and social scientists.* Oxford University Press.

Contrada, R., & Baum, A. (Eds.). (2011). *The handbook of stress science: Biology, psychology, and health*. Springer Publishing Company.

Coryell, J. E., & Clark, M. C. (2009). One right way, intercultural participation, and language learning anxiety: A qualitative analysis of adult online heritage and nonheritage language learners. *Foreign Language Annals*, 42(3), 483-504.

Costa-Font, J., Fleche, S., & Pagan, R. (2021). *Welfare Effects of Time Reallocation: Would Ending Daylight Saving Time Affect Wellbeing?*. IZA Discussion Paper No. 14570. https://ssrn.com/abstract=3892598

Creswell, J. W. (2014). *Qualitative, quantitative and mixed methods approaches*. Sage.

Crowder, M. J., & Hand, D. J. (2017). *Analysis of repeated measures*. Routledge.

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety: The experience of play in work and leisure*. San Francisco, CA: Jossey-Bass.

Csikzentmihalyi, M. (1997). *Finding flow*. Ziff-Davis Publishing Company.

Dawes, M. R. (1991). Giving up Cherished Ideas: The Rorschach Ink Blot Test. Institute for *Psychological Therapies Journal*, 3(4).

Day, S. J., & Altman, D.G. (2000). Statistics notes: blinding in clinical trials and other studies. *BMJ*. Aug 19-26;321(7259):504. https://10.1136/bmj.321.7259.504

De Witte, N. A., Buyck, I., & Van Daele, T. (2019). Combining biofeedback with stress management interventions: A systematic review of physiological and psychological effects. *Applied Psychophysiology and Biofeedback*, 44(2), 71-82.

Diamond, D. M., Campbell, A. M., Park, C. R., Halonen, J., & Zoladz, P. R. (2007). The temporal dynamics model of emotional memory processing: a synthesis on the neurobiological basis of stress-induced amnesia, flashbulb and traumatic memories, and the Yerkes-Dodson law. *Neural Plasticity*, 2007. https://doi.org/10.1155/2007/60803

DiVincenzo, A. (Ed.). (2014). *Find your purpose: The path to a successful doctoral experience.* Available from http://lc.gcumedia.com/res811/find-your-purpose-the-path-to-a-successful-doctoral-experience/v1.1/.

Eller, J. D. (2020). *Cultural anthropology: Global forces, local lives*. Routledge.

Evans, D. L., Foa, E. B., Gur, R. E., Hendin, H., O'Brien, C. P., Seligman, M. E., & Walsh, B. T. (Eds.). (2005). *Treating and preventing adolescent mental health disorders: What we know and what we don't know.* Oxford University Press.

Friedenberg, J., & Silverman, G. (2012). *Cognitive science: An introduction to the study of mind (2nd ed.)*. Sage Publications, Inc.

Furr, R. M. (2017). *Psychometrics: An introduction*. SAGE publications.

George, D., & Mallery, P. (2019). *IBM SPSS statistics 26 step by step: A simple guide and reference*. Routledge.

Gibbs, P., & Hiroshi, S. (1996). What is Occam's razor. *Available at: иид ЛЛллпКл йж К гбЛд нз зЛг бК иба*. https://math.ucr.edu/home/baez/physics/General/occam.html

Goldberg, D. (2013). *The General Health Questionnaire: General Health Questionnaire Manual (Enhanced Edition)*. Nihon Bunka Kagakusha.

Goldberg, D. P., & Hillier, V. F. (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine*, *9*(1), 139-145.

Grant-Smith, D., & McDonald, P. (2018). Ubiquitous yet ambiguous: An integrative review of unpaid work. *International Journal of Management Reviews*, 20(2), 559-578. https://doi.org/10.1111/ijmr.12153

Gregersen, T., MacIntyre, P. D., & Meza, M. D. (2014). The motion of emotion: Idiodynamic case studies of learners' foreign language anxiety. *The Modern Language Journal*, 98(2), 574-588.

Gruber, R., & Cassoff, J. (2014). The interplay between sleep and emotion regulation: conceptual framework empirical evidence and future directions. *Current Psychiatry Reports*, 16(11), 500.

Hagberg, L. A., & Lindholm, L. (2010). Measuring the time costs of exercise: a proposed measuring method and a pilot study. *Cost Effectiveness and Resource Allocation*, 8(1), 1-7. https://doi.org/10.1186/1478-7547-8-9

Harford, T. (2020). *How to Make the World Add Up: Ten Rules for Thinking Differently About Numbers*. Hachette UK.

Hebb, D. O. (1955). Drives and the CNS (conceptual nervous system). *Psychological Review, 62*(4), 243.

Hernández-Orallo, E., Manzoni, P., Calafate, C. T., & Cano, J. C. (2020). Evaluating how smartphone contact tracing technology can reduce the spread of infectious diseases: the case of COVID-19. *Ieee Access, 8*, 99083-99097. https://doi.org/10.1109/ACCESS.2020.2998042

Hickey, B. A., Chalmers, T., Newton, P., Lin, C. T., Sibbritt, D., McLachlan, C. S., ... & Lal, S. (2021). Smart Devices and Wearable Technologies to Detect and Monitor Mental Health Conditions and Stress: A Systematic Review. *Sensors*, 21(10), 3461. https://doi.org/10.3390/s21103461

Hickman, M. J., Fricas, J., Strom, K. J., & Pope, M. W. (2011). Mapping police stress. *Police Quarterly*, 14(3), 227-250. https://doi.org/10.1177/1098611111413991

Hirten, R. P., Stanley, S., Danieletto, M., Borman, Z., Grinspan, A., Rao, P., ... & Sands, B. E. (2021). Wearable devices are well accepted by patients in the study and management of inflammatory bowel disease: a survey study. *Digestive Diseases and Sciences, 66*(6), 1836-1844. https://doi.org/10.1007/s10620-020-06493-y

Hubbard, R., & Armstrong, J. S. (1997). Publication bias against null results. *Psychological Reports*, 80(1), 337-338. https://doi.org/10.2466/pr0.1997.80.1.337

Im, G. H., & Cheng, L. (2019). The test of English for international communication (TOEIC®). *Language Testing*, 36(2), 315-324. https://doi.org/10.1177/0265532219828252

Iwata, N., & Saito, K. (1992). The factor structure of the 28-item General Health Questionnaire when used in Japanese early adolescents and adult employees: age-and cross-cultural comparisons. *European Archives of Psychiatry and Clinical Neuroscience*, 242(2), 172-178.

Jauho, A. M., Pyky, R., Ahola, R., Kangas, M., Virtanen, P., Korpelainen, R., & Jämsä, T. (2015). Effect of wrist-worn activity monitor feedback on physical activity behavior: a randomized controlled trial in Finnish young men. *Preventive Medicine Reports*, 2, 628-634. https://doi.org/10.1016/j.pmedr.2015.07.005

Johnston, O., & Wildy, H. (2016). The effects of streaming in the secondary school on learning outcomes for Australian students–A review of the international literature. *Australian Journal of Education, 60*(1), 42-59. https://doi.org/10.1177/0004944115626522

Jovanov, E., Frith, K., Anderson, F., Milosevic, M., & Shrove, M. T. (2011, August). Real-time monitoring of occupational stress of nurses. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 3640-3643). IEEE. https://doi.org/10.1109/IEMBS.2011.6090612

Kay, R., & Jovanovic, P. (2021, March). Examining practical issues associated with the use of wearable teachnology in K-12 classrooms: A review of the literature. In *Proceedings of INTED2021 Conference* (Vol. 8, p. 9th). https://doi.org/10.21125/inted.2021

Kamper, S.J. (2018). Blinding: Linking Evidence to Practice. *J Orthop Sports Phys Ther.* Oct; 48(10):825-826. https://doi.org/10.2519/jospt.2018.0705

Kim, H. G., Cheon, E. J., Bai, D. S., Lee, Y. H., & Koo, B. H. (2018). Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry Investigation*, 15(3), 235. https://doi.org/10.30773/pi.2017.08.17

Kirk, R. (2007). *Statistics: An introduction*. Nelson Education.

Klasnja, P., & Pratt, W. (2012). Healthcare in the pocket: mapping the space of mobile-phone health interventions. *Journal of Biomedical Informatics*, 45(1), 184-198. https://doi.org/10.1016/j.jbi.2011.08.017

Klein, E. (1971). *A Comprehensive Etymological Dictionary of the English Language*. Amsterdam: Elsevier Scientific Publishing.

Klengel, T., & Binder, E. B. (2015). Epigenetics of stress-related psychiatric disorders and gene× environment interactions. *Neuron*, 86(6), 1343-1357. https://doi.org/10.1016/j.neuron.2015.05.036

Kondo, D. S., & Ying-Ling, Y. (2004). Strategies for coping with language anxiety: The case of students of English in Japan. *ELT Journal*, 58(3), 258-265. https://doi.org/10.1093/elt/58.3.258

Kostić-Bobanović, M. (2009). Foreign language anxiety of university students. *Economic Research-Ekonomska istraživanja*, 22(3), 47-55.

Li, S., Cullen, W. K., Anwyl, R., & Rowan, M. J. (2003). Dopamine-dependent facilitation of LTP induction in hippocampal CA1 by exposure to spatial novelty. *Nature Neuroscience, 6*(5), 526-531. https://doi.org/10.1038/nn1049

Li, W. C., Nirei, M., & Yamana, K. (2019). Value of data: there's no such thing as a free lunch in the digital economy. *RIETI*.

Lieber, M. (2017). Assessing the mental health impact of the 2011 great Japan earthquake, tsunami, and radiation disaster on elementary and middle school children in the Fukushima prefecture of Japan. *PLoS One*, 12(1), e0170402. https://doi.org/10.1371/journal.pone.0170402

Liu, M. (2006). Anxiety in EFL classrooms: Causes and consequences. *TESL Reporter*, 39, 20-20.

Liu, M. (2021). Foreign Language Classroom Anxiety, Gender, Discipline, and English Test Performance: A Cross-lagged Regression Study. *The Asia-Pacific Education Researcher*, 1-11. https://doi.org/10.1007/s40299-020-00550-w

Lupien, S. J., Ouellet-Morin, I., Hupbach, A., Tu, M. T., Buss, C., Walker, D., Pruessner, J., & McEwen, B. S. (2006). Beyond the stress concept: Allostatic load--a developmental biological and cognitive perspective. In D. Cicchetti & D. J. Cohen (Eds.), *Developmental psychopathology: Developmental neuroscience* (pp. 578–628). John Wiley & Sons Inc.

MacIntyre, P. D. (1995). How does anxiety affect second language learning? A reply to Sparks and Ganschow. *The Modern Language Journal*, 79(1), 90-99.

MacIntyre, P. D. (2007). Willingness to communicate in the second language: Understanding the decision to speak as a volitional process. *The Modern Language Journal*, 91(4), 564-576.

Markov, D. & Goldman, M. (2006). Normal sleep and circadian rhythms: Neurobiologic mechanisms underlying sleep and wakefulness. *Psychiatric Clinics of North America, 29*(4), 841-853.

McEwen, B. S., & Stellar, E. (1993). Stress and the individual: Mechanisms leading to disease. *Archives of internal Medicine*, 153(18), 2093-2101.

McNamara, T., Knoch, U., Fan, J., & Rossner, R. (2019). *Fairness, justice & language assessment.* Oxford University Press.

Milojevich, H. M., & Lukowski, A. F. (2016). Sleep and mental health in undergraduate students with generally healthy sleep habits. *PloS One*, 11(6), e0156372.

Mitchell, T. R., & Daniels, D. (2003). Motivation. WC Borman, RJ, Klimoski,(Eds.) *Handbook of Psychology, Vol. 12: Industrial Organizational Psychology*, Borman, RJ, Klimoski (Eds.) (pp. 225-254).

Mitchell, T. R., Daniels, D. (2003). "Motivation". In Borman, W. C., Ilgen, D. R., Klimoski, R.J. (eds.). *Handbook of Psychology (volume 12)*. John Wiley & Sons, Inc.

Montague, E. K. (1953). The role of anxiety in serial rote learning. *Journal of Experimental Psychology, 45(2*), 91–96. https://doi.org/10.1037/h0062644

Morse, J. M. (2005). The paid/unpaid work of participants. *Qualitative Health Research*, 15 (6), p. 727-728. Sage Publications.

Ometov, A., Shubina, V., Klus, L., Skibińska, J., Saafi, S., Pascacio, P., ... & Lohan, E. S. (2021). A survey on wearable technology: History, state-of-the-art and current challenges. *Computer Networks*, 193, 108074. https://doi.org/10.1016/j.comnet.2021.108074

Pakhomov, S.V.S., Thuras, P.D., Finzel, R., Eppel, J., & Kotlyar, M. (2020) Using consumer-wearable technology for remote assessment of physiological response to stress in the naturalistic environment. *PLoS One* 15(3): e0229942. https://doi.org/10.1371/journal.pone.0229942

Pang, I., Okubo, Y., Sturnieks, D., Lord, S. R., & Brodie, M. A. (2019). Detection of near falls using wearable devices: a systematic review. *Journal of Geriatric Physical Therapy*, 42(1), 48-56. https://doi.org/10.1519/JPT.0000000000000181

Papadimitriou, G. N., & Linkowski, P. (2005). Sleep disturbance in anxiety disorders. *International Review of Psychiatry, 17*(4), 229-236. https://doi.org/10.1080/09540260500104524

Poirier, J., Bennett, W. L., Jerome, G. J., Shah, N. G., Lazo, M., Yeh, H. C., Clark, J. M., & Cobb, N. K. (2016). Effectiveness of an activity tracker-and internet-based adaptive walking program for adults: a randomized controlled trial. *Journal of Medical Internet Research*, 18(2), e5295. https://doi.org/10.2196/jmir.5295

Polit, D. F., & Beck, C. T. (2010). Generalization in quantitative and qualitative research: Myths and strategies. *International Journal of Nursing Studies, 47*(11), 1451-1458.

Polman, R., & Borkoles, E. (2011). The fallacy of directional anxiety. *International Journal of Sport Psychology*, 42(3), 303-306.

Pritschet, L., Powell, D., & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*, *27*(7), 1036-1042.

Public Health England (2021, September 30). *Common Mental Health Disorders – Glossary.* https://fingertips.phe.org.uk/profile/common-mental-disorders/supporting-information/Glossary

Reite, M., Weissberg, M. P. & Ruddy, J. (2008). *Clinical manual for evaluation and treatment of sleep disorders*. Washington, DC: American Psychiatric Publishing.

Riederer, C., Erramilli, V., Chaintreau, A., Krishnamurthy, B., & Rodriguez, P. (2011, November). For sale: your data: by: you. In *Proceedings of the 10th ACM Workshop on Hot Topics in Networks* (pp. 1-6).

Ripley, E., Macrina, F., Markowitz, M., & Gennings, C. (2010). Why do we pay? A national survey of investigators and IRB chairpersons. *Journal of Empirical Research on Human Research Ethics*, *5*(3), 43-56. https://doi.org/10.1525/jer.2010.5.3.43

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review*, 3(1), 16-20.

Salmon, P. (2001). Effects of physical exercise on anxiety, depression, and sensitivity to stress: a unifying theory. *Clinical Psychology Review*, 21(1), 33-61. https://doi.org/10.1016/S0272-7358(99)00032-X

Sapolsky, R. M. (2015). Stress and the brain: individual variability and the inverted-U. *Nature Neuroscience, 18*(10), 1344-1346.

Sasaki, K., & Maruyama, R. (2014). Consciously controlled breathing decreases the high-frequency component of heart rate variability by inhibiting cardiac parasympathetic nerve activity. *The Tohoku Journal of Experimental Medicine*, 233(3), 155-163.

Scott, B. G., & Weems, C. F. (2014). Resting vagal tone and vagal response to stress: associations with anxiety, aggression, and perceived anxiety control among youths. *Psychophysiology*, 51(8), 718-727.

Smith, E. N., Santoro, E., Moraveji, N., Susi, M., & Crum, A. J. (2020). Integrating wearables in stress management interventions: Promising evidence from a randomized trial. *International Journal of Stress Management*, 27(2), 172. https://doi.org/10.1037/str0000137

Smoller, J. W. (2016). The genetics of stress-related disorders: PTSD, depression, and anxiety disorders. *Neuropsychopharmacology*, 41(1), 297-319. https://doi.org/10.1038/npp.2015.266

Sparks, R. L., & Ganschow, L. (2007). Is the foreign language classroom anxiety scale measuring anxiety or language skills?. *Foreign Language Annals*, 40(2), 260-287.

Spielberger, C. D. (1983). *State-Trait Anxiety Inventory for Adults (STAI-AD)* [Database record]. APA PsycTests. https://doi.org/10.1037/t06496-000 https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft06496-000

Stangroom, J. (2021). *P Value from Pearson (R) Calculator*. Social Science Statistics. https://www.socscistatistics.com/pvalues/pearsondistribution.aspx

Stephens, T. (1988). Physical activity and mental health in the United States and Canada: evidence from four population surveys. *Preventive Medicine, 17*(1), 35-47. https://doi.org/10.1016/0091-7435(88)90070-9

Steptoe, A., Wardle, J., Fuller, R., Holte, A., Justo, J., Sanderman, R., & Wichstrøm, L. (1997). Leisure-time physical exercise: prevalence, attitudinal correlates, and behavioral correlates among young Europeans from 21 countries. *Preventive Medicine*, 26(6), 845-854. https://doi.org/10.1006/pmed.1997.0224

Sterling, M. (2011). General health questionnaire–28 (GHQ-28). *Journal of Physiotherapy*, 57(4), 259. https://doi.org/10.1016/S1836-9553(11)70060-1

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: evolving generalization through selective processing. *Nature Neuroscience*, 16(2), 139-145.

Strack, J., & Esteves, F. (2015). Exams? Why worry? Interpreting anxiety as facilitative and stress appraisals. *Anxiety, Stress, & Coping*, *28*(2), 205-214.

Strath, S. J., & Rowley, T. W. (2018). Wearables for promoting physical activity. *Clinical Chemistry, 64*(1), 53-63. https://doi.org/10.1373/clinchem.2017.272369

Suda, M., Nakayama, K., & Morimoto, K. (2007). Relationship between behavioral lifestyle and mental health status evaluated using the GHQ-28 and SDS questionnaires in Japanese factory workers. *Industrial Health*, 45(3), 467-473. https://doi.org/10.2486/indhealth.45.467

Taelman, J., Vandeput, S., Spaepen, A., & Van Huffel, S. (2009). Influence of mental stress on heart rate and heart rate variability. In *4th European conference of the international federation for medical and biological engineering* (pp. 1366-1369). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-89208-3_324

Tallon, M. (2009). Foreign language anxiety and heritage students of Spanish: A quantitative study. *Foreign Language Annals*, 42(1), 112-137.

Taylor, J. A., & Chapman, J. P. (1955). Anxiety and the learning of paired associates. *The American Journal of Psychology, 68*, 671. https://doi.org/10.2307/1418800

Thayer, J. F., Åhs, F., Fredrikson, M., Sollers III, J. J., & Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, 36(2), 747-756. https://doi.org/10.1016/j.neubiorev.2011.11.009

Tian, H., Han, D., & Wang, B. (2021, May). A Summary of wearable textiles power generation. In *IOP Conference Series: Earth and Environmental Science* (Vol. 772, No. 1, p. 012037). IOP Publishing.

Tran, T. T. T. (2012). A Review of Horwitz, Horwitz and Cope's Theory of Foreign Language Anxiety and the Challenges to the Theory. *English Language Teaching*, 5(1), 69-75.

Tweedie, S. (2015). The world's first smartphone, Simon, was created 15 years before the iPhone. *Business Insider, 14.*

U. S. Food and Drug Administration. (2021, June 5). *Digital Health Software Precertification (Pre-Cert) Program*. Author. https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-program

U.S. Census Bureau (1994). 1990 Census: Language Spoken at Home and Ability to Speak English for United States, Regions and States: 1990. https://www.census.gov/data/tables/time-series/dec/cph-series/cph-l/cph-l-133.html

Van den Bos, G. R. (2007). *APA dictionary of psychology*. American Psychological Association.

Volpato, L., del Río Carral, M., Senn, N., & Delefosse, M. S. (2021). General Practitioners' Perceptions of the Use of Wearable Electronic Health Monitoring Devices: Qualitative Analysis of Risks and Benefits. *JMIR mHealth and uHealth*, 9(8), e23896.c https://doi.org/10.2196/23896

Von Worde, R. (2003). Students' Perspectives on Foreign Language Anxiety. *Inquiry*, 8(1), n1.

Wallert, J., Held, C., Madison, G., & Olsson, E. M. (2017). Temporal changes in myocardial infarction incidence rates are associated with periods of perceived psychosocial stress: A SWEDEHEART national registry study. *American Heart Journal, 191*, 12-20.

Webster, M. (2016). *Merriam-Webster's Medical Dictionary*. Springfield, MA: Merriam-Webster.

Weston, M. (2015). Wearable surveillance–a step too far?. *Strategic HR Review,* 14 (6), 214-219. http://dx.doi.org/10.1108/SHR-09-2015-0072

Wong, M. L., Lau, E. Y. Y., Wan, J. H. Y., Cheung, S. F., Hui, C. H., & Mok, D. S. Y. (2013). The interplay between sleep and mood in predicting academic functioning, physical health and psychological health: A longitudinal study. *Journal of Psychosomatic Research*, 74(4), 271-277.

Yamashiro, A., & McLaughlin, J. (2001). Relationships among attitudes, motivation, anxiety, and English language proficiency in Japanese college students. *Second Language Acquisition Research in Japan*, 113-127.

Yap, Y., Slavish, D. C., Taylor, D. J., Bei, B., & Wiley, J. F. (2020). Bi-directional relations between stress and self-reported and actigraphy-assessed sleep: a daily intensive longitudinal study. *Sleep*, *43*(3), zsz250. https://doi.org/10.1093/sleep/zsz250

Appendix A: Participants' Research Diary and Schedule

Author's Note 1:
This manual and schedule was originally formatted for B5 paper size, so some of the formatting has shifted.

An Investigation of Language Learning Anxiety by Biometric Measurement of Young People in Japan

Diary and Instructions

Please read this carefully and try to follow the instructions as closely as possible. If you have any questions please do not hesitate to contact me anonymously at macdonald@sun.ac.jp

Research Participant Number

Author's Note 2:
The research participant number was handwritten by the researcher, and was purely a means of providing a unique identified to each diary. It was in no way linked to the participant's other documentation.

220

Sunday 日	Monday 月	Tuesday 火	Wednesday 水	Thursday 木	Friday 金	Saturday 土
	12	13	14	15	16	17
	Please fill in the Research checklist (p.6), Questionnaires 1 and 2 (p. 7~9), your Normal Weekly schedule (p. 12) Every day please rate your stress (p. 13) Please do not wear the smartwatch.					
10/18	19	20	21	22	23	24
	Every day please rate your stress (p. 13). You can wear the smartwatch if you want.					
10/25	26	27	28	29	30	31
	Please do Questionnaire 3 (p. 9) Every day please rate your stress (p. 13). Please wear the smartwatch as much as possible.					
11/1	2	3	4	5	6	7
	Every day please rate your stress (p. 13). Please wear the smartwatch as much as possible.					
11/8	9	10	11	12	13	14
	Every day please rate your stress (p. 13). You can wear the smartwatch if you want.					
11/15	16	17	18	19	20	21
	Please do Questionnaire 4 (p. 10) Every day please rate your stress (p. 13) Please do not wear the smart watch. 11/24 Please return the smartwatch and research manual to the box outside of Mac's office (W518).					

Author's Note 3: The dates are MM/DD format. In Japanese it is unambiguous and appears with clear indicators. For example: 10 月(month)18 日(day).

Research Schedule

Week 1 10/12~10/17	**Page 5.** Please do the research checklist **Page 6~8.** Please complete questionnaire 1 and 2. **Page 12**. Please write your Normal Weekly Schedule once. Please do not wear the smartwatch. **Page 13.** Please rate your stress level every day from 1 (very low) to 10 (very high). Please note any unusually stressful events.	
Week 2 10/18~10/24	You can wear the smartwatch. Check the smartwatch is fully charged. Download the smartwatch application (Garmin Connect). Create a log in at https://connect.garmin.com/ and set up your profile. Please do not include any personal information, such as your school email address, photographs of yourself, or your real name. Please create a throw-away email address for this. **Page 13.** Please rate your stress level every day. Please note any unusually stressful events. Please write your unique user name and password below so that I can access your Garmin account and download your smartwatch data. **This is also necessary to protect the privacy of your data.**	 
User name:		@gmail.com
Password:		
Week 3 and Week 4 10/25~11/7	**Page 9.** Please complete questionnaire 3. **Please wear the smartwatch as much as possible. It is quite tough.** You can wear it playing sports, or even swimming. **Page 13.** Please rate your stress level every day. Please note any unusually stressful events.	

Week 5 and 6 (11/8～11/21)		
Week 5 11/8~11/14	**Page 13.** Please rate your stress level every day. Please note any unusually stressful events.	
Week 6 11/15~11/21	Please do not wear the smartwatch. **Page 13.** Please rate your stress level every day. Please note any unusually stressful events. **Page 10**. Please complete questionnaire 4.	
Finished! (11/24)		
11/24	Please return all documents and the smartwatch to the box outside of Mac's office (W518).	

Research Checklist

Sex: Male ☐ Female ☐ Other: ☐

Age: ☐

1. What was your lastest TOEIC test result: ☐ 点
2. What was your highest TOEIC test result? ☐ 点
3. What was your university entrance TOEIC test result? ☐ 点

Author's Note 4: Question 3 seems awkwardly phrased in English, but is much easier in Japanese.

4. Do you have any heart conditions? ☐Yes ☐No
5. Do you have a history of anxiety or depression? ☐Yes ☐No
6. Are you currently suffering from anxiety or depression? ☐Yes ☐No
7. Are you suffering from chronic pain? ☐Yes ☐No
8. Is there any other reason why you think your stress levels may be unusual?

☐

Questionnaire 1

Author Note 5:

The Japanese GHQ-28 went here.

This page has been blanked as the author does not have permission from the copyright holders to reproduce the questionnaire for publication.

Author Note 6: This is the Japanese version of the FLCAS

Questionnaire 2

Please mark the answer that best describes your feelings with an X for each question.

質問ごとに、自分の気持ちを最もよく表す答えに X を付けてください。

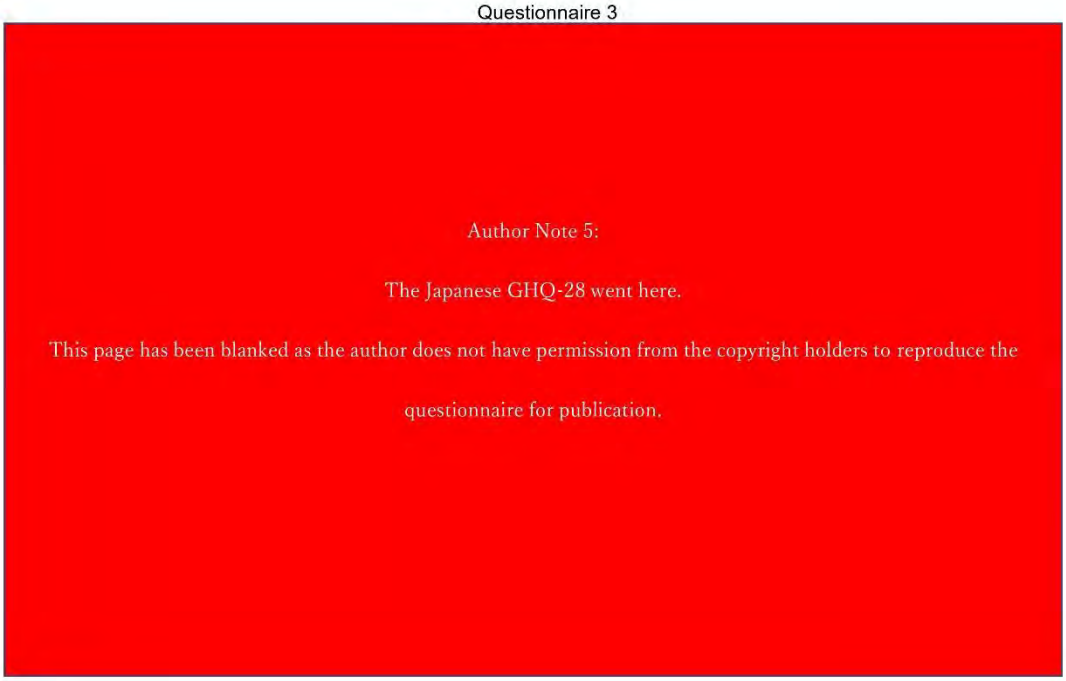
全く 当てはまらない	あまり 当てはまらない	少し 当てはまらない	少し 当てはまる	まあまあ 当てはまる	とてもよく 当てはまる
①	②	③	④	⑤	⑥

1	外国語の授業で話すとき自信がもてない。	① ② ③ ④ ⑤ ⑥
2	外国語の授業で間違うことは気にならない。*	① ② ③ ④ ⑤ ⑥
3	外国語の授業で当てられると思うと体が震える。	① ② ③ ④ ⑤ ⑥
4	外国語の授業で先生の言っていることが理解できないととても不安だ。	① ② ③ ④ ⑤ ⑥
5	もっと外国語の授業があってもよいと思っている。*	① ② ③ ④ ⑤ ⑥
6	外国語の時間授業と関係ないことを考えていることがよくある。	① ② ③ ④ ⑤ ⑥
7	他の生徒の方が自分よりよくできると思っている。	① ② ③ ④ ⑤ ⑥
8	外国語の授業中のテストではだいたい落ち着いている。*	① ② ③ ④ ⑤ ⑥
9	外国語の授業で準備なしに話さないといけない時、パニックになる。	① ② ③ ④ ⑤ ⑥
10	外国語の単位を落としたときの影響が心配だ。	① ② ③ ④ ⑤ ⑥

11	外国語の授業で動揺する人の気持ちがわからない。*	① ② ③ ④ ⑤ ⑥
12	外国語の授業では、緊張のあまり、知ってたことも忘れてしまうときがある。	① ② ③ ④ ⑤ ⑥
13	外国語の授業で自分からすすんで答えるのは恥ずかしい。	① ② ③ ④ ⑤ ⑥
14	外国語をネーティブスピーカーと話すとき緊張しない。*	① ② ③ ④ ⑤ ⑥
15	先生が何を訂正しているのか理解できないとき動揺する。	① ② ③ ④ ⑤ ⑥
16	外国語の授業の予習を十分にしていても心配になる。	① ② ③ ④ ⑤ ⑥
17	よく外国語の授業を休みたくなる。	① ② ③ ④ ⑤ ⑥
18	外国語の授業で話すのに自信がある。*	① ② ③ ④ ⑤ ⑥
19	先生が自分の間違いをいちいち直しそうなので心配だ。	① ② ③ ④ ⑤ ⑥
20	外国語のクラスで当たりそうになると胸がどきどきする。	① ② ③ ④ ⑤ ⑥
21	外国語のテスト勉強をすればするほど、混乱する。	① ② ③ ④ ⑤ ⑥
22	外国語の授業の予習をよくしないといけないというプレッシャーは感じない。*	① ② ③ ④ ⑤ ⑥
23	常に他の学生の方が外国語で話すのが上手だと感じている。	① ② ③ ④ ⑤ ⑥
24	他の学生の前で外国語を話すとき自意識がとても高くなる。	① ② ③ ④ ⑤ ⑥
25	外国語のクラスは進むのが速いのでついていけるかどうか心配である。	① ② ③ ④ ⑤ ⑥
26	他の科目よりも外国語のクラスの方が緊張する。	① ② ③ ④ ⑤ ⑥

27	外国語のクラスで話すとき緊張したり混乱したりする。	① ② ③ ④ ⑤ ⑥
28	外国語のクラスに向かうとき自信をもてるしリラックスしている。*	① ② ③ ④ ⑤ ⑥
29	先生の言うことがすべて理解できないと不安になる。	① ② ③ ④ ⑤ ⑥
30	外国語を話すためにあまりに多くの文法規則を勉強しないといけないので圧倒される。	① ② ③ ④ ⑤ ⑥
31	私が外国語を話すと他の学生が笑うのではないかと思う。	① ② ③ ④ ⑤ ⑥
32	ネーティブスピーカーに会うときおそらくリラックスしていられると思う。*	① ② ③ ④ ⑤ ⑥
33	先生が、前もって準備していなかった質問をすると緊張する。	① ② ③ ④ ⑤ ⑥
(Source: Yashima, Noels, Shizuka, Takeuchi, Yamane, Yoshizawa, 2009)		

Questionnaire 3



9

Questionnaire 4

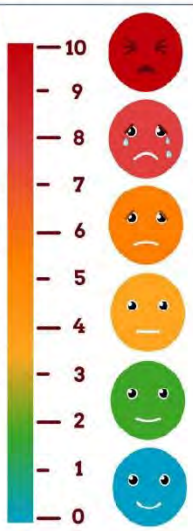
10

例えば

Normal Weekly Schedule 「通常の週間スケジュール」

Instructions: Please write your normal weekly schedule here. Japanese or English is okay. How much stress do you normally feel in each class? Rate it 1 to 10.

時間 Time	月曜日 Mon	火曜日 Tue	hurs	金曜日 Fri	土曜日 Sat	日曜日 Sun
Before Class					Part Time Job（7）１０ｈ００～１８ｈ００	無し （0）
09h00 ~ 10h30	OC3 (8)					
10h40 ~ 12h10		フランス語（3）		モンゴル語（1）		
12h10 ~ 13h00	Lunch!(1)					
13h00 ~ 14h30	国際法 (5)					
14h40 ~ 16h10		ディベート(10)		国際セミナー（5）		
16h20 ~ 17h50						
After Class						



11

Normal Weekly Schedule

Instructions: Please write your normal weekly schedule here. Japanese or English is okay.

時間 Time	Mon	Tue	Wed	Thurs	Fri	Sat	Sun
Before Class							
09h00 ~ 10h30							
10h40 ~ 12h10							
12h10 ~ 13h00							
13h00 ~ 14h30							
14h40 ~ 16h10							
16h20 ~ 17h50							
After Class							

12

Diary

(very low) to 10 (very high). Please note any unusually stressful events.

Sunday 日	Monday 月	Tuesday 火	Wednesday 水	Thursday 木	Friday 金	Saturday 土
10/11	12	13	14	15	16	17
10/18	19	20	21	22	23	24
10/25	26	27	28	29	30	31
11/1	2	3	4	5	6	7
11/8	9	10				
11/15	16	17				

Author Note 7:
The space provided may seem very small, but Japanese tends to be terse and the pictograms allow very dense communication. The sentence, "I had an argument with my boyfriend" is just 7 characters, ("彼氏と口論した").

13

14

Appendix B: Participant Briefing Information

This documentation was provided to participants in Japanese. The original Japanese documents are available on request. It should be noted that while these translations are as close to the Japanese as possible the wording has been changed slightly in some places to ensure readability and to remain true to the nuance and meaning.

The following document was provided to participants at the briefing regarding the research, in addition to the consent forms, withdrawal form, and research diary and schedule (see Appendix A).

Investigating Language Learning Anxiety by Biometric Measures in Japanese Youth

Request for cooperation

This document explains the contents of the research in order to ask you for your cooperation in research on “Investigating Language Learning Anxiety by Biometric Measures in Japanese Youth”. If you understand the content and agreeing to cooperate in the research, please express your consent by signing or affixing your seal to the "Agreement on cooperation for research". If you are under 20 years of age you may ask your parents or guardian to sign or affix their seal in addition to your own if you wish. Bring the signed consent form to the researcher’s office (W518) to receive on of the smart bands for use in this research.

Cooperation in this research is by your own free will, and if you do not agree, you will not suffer any disadvantage for that reason. In addition, even after agreeing you can revoke consent at any time without suffering any disadvantage.

Research Project Name

Investigating Language Learning Anxiety by Biometric Measures in Japanese Youth
(This research was approved by the president after review by the Nagasaki prefectural university general research ethics committee.)

Significance and purpose of research

This research will investigate heart rate variation. It will also continuously observe university students’ sleep, exercise, stress levels, academic results and other physical and mental conditions by the use of a wearable fitness tracker and through questionnaires. By studying

this, we hope to gain insights that will help parents and teachers. Our aim is to promote the health and education of young people who will be the future.

Researcher

William Tait MacDonald [The University of Nagasaki, Siebold Campus]

What is asked of participants:

Students in the research group will wear the fitness bands for two weeks out of the total research period of six weeks. The data collected from this band includes sleep, movement, heart rate, and stress levels. Participants will also fill in a diary at the end of each day where they will record their mood and events of the day. This should take about 10 minutes. Every effort has been made to make filling in the records easy. Participants are also asked to complete a questionnaire at requested intervals, about every two weeks.

On Expected Risks:

Although smart bands are considered safe if you experience any of the following then please remove the smart band immediately:

- Rash or itching, perhaps caused by an allergic reaction to the smart band.
- Heat rash caused by sweating
- The band becoming hot, perhaps as a result of a battery malfunction

Regarding the handling of personal information:

In order to preserve students' anonymity and confidentiality researchers will not meet directly with the students other than to provide them with the research materials and to provide replacements if materials are lost or damaged. Participants will be assigned a number and all data will be organized according to this number. In doing so it will be impossible for researchers to link participants' identities and data.

All physical data will be kept in a locked cabinet. Electronic data will be kept on a laptop protected by fingerprint and password authorization. Some electronic data will be stored on the fitness band manufacturer's website. This data will be stored on a special server for minors and deleted after the research has been completed and the data downloaded by the researcher. After the project is completed data will be kept for 3 years and then shredded or deleted.

Publication of Results

Research results will be included in the researcher's doctoral thesis, but additional research based on the data may be published at conferences and in papers.

Costs of participation

Even if students accidentally damage or destroy a smart band they will not be held responsible. If you contact the researcher then a new one will be issued.

Other:

If you have any questions, please email me anonymously at ***macdonald@sun.ac.jp***. Informed consent is very important, so if you have any questions or concerns please do not hesitate to contact us.

However, if you need assistance urgently during the research then please consult the appropriate individual at the university. The university has a free clinic located on the first floor of the administration block. A public health nurse is available to provide personalized care in response to various concerns and questions related to student health. A mental health support system that includes counselling by a clinical psychologist has also been established for students.

Inquiries about the research:

William MacDonald
Office: W518
Email: ***macdonald@sun.ac.jp***
Physical address:
〒851-2195 西彼杵郡長与町まなび野 1-1-1
長崎県立大学シーボルト校　国際社会学部 W518 研究室
TEL. 095-813-5156

Regarding the Smart Band

For this research, the Garmin Vivosmart 3 or 4 will be used. The box provided to you on handing in the consent form will include the smart band, manual, and charging cord. Please return the box and contents to the box outside of the researcher's office at the end of this study. The following information about the smart band is from the manufacturer's website at ***https://www.garmin.co.jp/products//vivosmart-3-black/***:

Author's notes:

1. The website address included above no longer exists. The Japanese content can currently be found at: ***https://www.garmin.co.jp/products/discontinued/vivosmart-3-black/***
2. The English version offered below is copied and pasted from ***https://www.garmin.com/en-US/p/567813*** as the Japanese version does not differ substantially from the U.S. version in this case.

Smart Fitness Tracker[1] with Wrist-based Heart Rate[1] and Fitness Monitoring Tools

Includes fitness monitoring tools such as VO2 max, fitness age and strength training

Monitor wellness with all-day stress tracking and the relaxation-based breathing timer

24/7 heart rate monitoring with Elevate™ wrist heart rate technology

Tracks steps, floors climbed, calories burned, intensity minutes, sleep and more

Safe for swimming and showering[2]

Battery life: up to 5 days[3]

Daily Fitness Monitoring

Thanks to Elevate wrist-based heart rate technology[1], vívosmart 3 fitness tracker is full of great features that let you get a better idea of your current fitness level. For example, it's our first dedicated fitness tracker to offer VO2 max estimate. It can even estimate your fitness age, which you may be able to decrease over time with hard work and regular exercise.

Wellness Monitoring

Vívosmart 3 tracks heart rate variability (HRV), which is used to calculate and display your stress level. You can look at this graph right on your device anytime you're sitting or at rest.

When the level starts to move up, it indicates that your stress is getting higher. What causes this? Any number of physical or emotional external sources. Vívosmart 3 helps you recognize when this is happening so you can find a way to relieve the stress, for example, by using the relaxation-based breathing timer, built right into the stress widget. Based on the Fourfold breathing technique, it offers a 1 to 5 minute guided breathing exercise.

Automatic Fitness Tracking[1]

Vívosmart 3 fitness tracking capabilities include steps, floors climbed, calories burned, intensity minutes and more. Move IQ® automatically detects exercises such as walking, swimming, cycling and elliptical training. It can also automatically start a timed walk or run activity. So you're free to get up, get moving, and let vívosmart 3 capture your active life. As always, you can review your activities later on ***Garmin Connect***™, our online community.

Author's notes:

3. Participants were also provided with a copy of the manual for the Garmin Vivosmart 3/4 in the box that contained the wearable device.

Appendix C: Garmin Data Handling

The following appendix provides copies of the email correspondence between the author and Garmin regarding the protection of participants' data. It should be noted that only the correspondence relevant to the protection of participants' data is provided, not the entire conversation. In total 28 emails were exchanged with Garmin regarding this matter over a period of a little over two months.

The full correspondence is available on request, however only a short summary will be presented here. The author first emailed the Garmin head office in the U.S., who provided links to the U.S. policies. On reviewing the policies the author noted that the information was specific to the U.S., and enquired if it was the same in Japan. The author was then put in contact with the support staff at ***service.JP@garmin.com*** where initially the contact was in Japanese, but then referred to an English-speaking member of the support staff who after some correspondence clarified the matter as detailed in the email exchange below:

from:	service.JP@garmin.com
reply-to:	service.JP@garmin.com
to:	shalestra@gmail.com
date:	25 Oct 2019, 20:54
subject:	Re: Re: Re: Garmin Vivosmart Data Policy <Q#:1005019> << Reference ID: 10725841K1 >>
mailed-by:	garmin.com
Signed by:	garmin.com
security:	Standard encryption (TLS)

Hi William,

Thank you for contacting Garmin Japan.

This email was transferred to Japan since you live here.

I am checking with the Garmin Japan team on how we can help you. (Or if the US team can help you.)

You want to know what the privacy rules are for Garmin in Japan? You want to know if the age is the same?

Best regards.

Garmin Japan
Product Support
Gabriel Abe

from:	William MacDonald <shalestra@gmail.com>
to:	service.JP@garmin.com
date:	25 Oct 2019, 21:10
subject:	Re: Re: Re: Garmin Vivosmart Data Policy <Q#:1005019> << Reference ID: 10725841K1 >>
mailed-by:	gmail.com

Hi Gabriel,

Thanks for all your help.

Garmin's US policy regarding children states that, "We request individuals under the age of 13 in the U.S. and under the age of 16 in the rest of the world not provide personal data to Garmin. If we learn that we have collected personal data from a child under the age of 13 in the U.S. or under 16 in the rest of the world, we will take steps to delete the information as soon as possible." (***https://www.garmin.com/en-US/privacy/connect/policy/#children***)

However, in Japan anyone under 20 is considered a child. I am doing research using the Garmin Vivosmart 3 where some of the people are under 20. I need to be able to assure them that their data will be kept private, and safe. I'll ask that they not include any personal data, but they might, and I need to know that it will be safe. I also need to know that it will be completely deleted after the study is completed.

Kindest regards,

William MacDonald

from:	service.JP@garmin.com
reply-to:	service.JP@garmin.com
to:	shalestra@gmail.com
date:	29 Oct 2019, 21:05
subject:	Re: Re: Re: Garmin Vivosmart Data Policy <Q#:1005019> << Reference ID: 10725841K1 >>
mailed-by:	garmin.com
Signed by:	garmin.com
security:	Standard encryption (TLS)

Hi William,

Thank you for your clarification sir.

In Japan 20 is the age someone becomes an adult.

I have consulted with my manager.

If you send us the user names of the accounts Garmin Japan will mark all those accounts as childrens data.

Their information will not be shared with anyone.

Child data is on different servers in Japan. It is separate and is not used for any purposes.

After you finish your research you can email Garmin and the accounts will be forever deleted.

Or the account user can delete their data anytime and it is the same.

Best regards.

Garmin Japan
Product Support
Gabriel Abe

Later emails were exchanged where the researcher provided *Garmin* with lists of participant account logins and received confirmation that these accounts and their associated information had been moved to the servers for minors. These emails have not been provided here. Despite the request for participants to create a throw-away email address for this login there is the possibility that some of the participants may not have done so, and so the email addresses will not be published. The researcher manually deleted all of the associated accounts themselves, and then confirmed their deletion in a final email once all accounts had been deleted. Again, as these email exchanges included participant email addresses which may still be in use the researcher has not included them here for publication.

The researcher would like to take this opportunity to thank *Garmin*, and especially Gabriel Abe, for their help and support. During this research more than 200 emails were exchanged between the researcher and Garmin regarding issues ranging from technical enquiries to questions such as these about data security and handling. Without the kind assistance provided this research would not have been possible.

Appendix D: Ethical Clearance

Ethical clearance for this study was obtained from both Rhodes University and The University of Nagasaki, Siebold Campus. The ethical clearance letters are attached here with the relevant tracking numbers. Full documentation of the ethical clearance process is available on request.

D.1. Rhodes Ethical Clearance (Tracking Number: PSY2017/17)

RHODES UNIVERSITY
Where leaders learn

Psychology Department
1 University Road, Grahamstown, 6139, South Africa
PO Box 94, Grahamstown, 6140, South Africa
T: +27 (0) 46 603 8500
T: +27 (0) 46 603 7614
E: psychology@ ru.ac.za

RESEARCH PROJECTS AND ETHICS REVIEW COMMITTEE

12 April 2017

William MacDonald
Department of Psychology
RHODES UNIVERSITY
6140

Dear William

ETHICAL CLEARANCE OF PROJECT PSY2017/17

This letter confirms your research proposal with tracking number PSY2017/17 and title, 'Towards a New Model of Language Learning Anxiety: Investigating Biometric Measures in a Sample of Japanese Youth', served at the Research Projects and Ethics Review Committee (RPERC) of the Psychology Department of Rhodes University on 29 March 2017. The project has been approved.

Please ensure that the RPERC is notified should any substantive change(s) be made, for whatever reason, during the research process. This includes changes in investigators.

Yours sincerely

Mr. Werner Bohmke
CHAIRPERSON: RPERC

D.2. The University of Nagasaki Ethical Clearance (Application 368, Judgement 355)

Attached here is a scanned and annotated version of the ethical clearance document issued by the University of Nagasaki ethics committee. The original version is available on request, as is the full documentation submitted in support of the documents.

様式第２号（第５条関係）
一部改正［平成27年規程第79号］

Form 2 (Related to Article 5) of Regulation 79 of of 2015

判　定　通　知　書

(Research Ethics Committee) Judgement

平成30年11月21日

21st November 2018

Department of International Social Studies, William MacDonald

国際社会学科　特任講師　W. マクドナルド　様

長崎県立大学長　太田　博道　印

University of Nagasaki
President Hiroto Ota
(Stamped with official seal)

受付番号　368
承認番号　355

Application Number 368
Judgement Number 355

課 題

日本の若者の生体測定による言語学習不安の調査

An Investigation of Language Learning Anxiety by Biometric Measurement of Young

研究責任者　所属 国際社会学科　職名 特任講師　氏名 W. マクドナルド

Chief Investigator: Department of International Social Studies, William MacDonald

さきに申請のあった上記課題にかかわる研究計画を、次のとおり判定したので通知する。

Regarding the research plan submitted earlier the judgement of the research ethics committee is as follows

判定	Approved　承認　不承認　Rejected 非該当 Requires amendment
理由等	Comments field regarding above.