2022

Assess, Address, Progress: An Online Approach to Evaluate and Develop Teacher Education Students’ Numeracy Capability

Kate Hartup
The University of Notre Dame Australia

Publication Details
Hartup, K. (2022). Assess, Address, Progress: An Online Approach to Evaluate and Develop Teacher Education Students’ Numeracy Capability [Doctor of Philosophy (College of Education)}]. The University of Notre Dame Australia. https://researchonline.nd.edu.au/theses/362

ASSESS, ADDRESS, PROGRESS:

AN ONLINE APPROACH TO EVALUATE AND DEVELOP TEACHER EDUCATION STUDENTS' NUMERACY CAPABILITY

Kate Hartup

BPhysEd; GradCert (Mathematics); MEd (Mathematics)

Submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy

SCHOOL OF EDUCATION, SYDNEY CAMPUS

THE UNIVERSITY OF NOTRE DAME AUSTRALIA

March 2022

Supervisors:

Dr Thuan Thai (UNDA)

Dr Amanda Yeung (UNSW)

Declaration of Authorship

I, Kate Hartup, declare that the work in this thesis, submitted in fulfillment of the requirements of the award of Doctor of Philosophy, in the School of Education at the University of Notre Dame, Australia, is my own work.

To the best of my knowledge, this thesis contains no material previously published by another person, except where due acknowledgement has been made.

Human Ethics: The research presented in this thesis was conducted in accordance with the National Health and Medical Research Council National Statement on Ethical Conduct in Human Research (2007; updated 2018). The research was proposed to both institutions involved in the study and the Research Ethics Committees reviewed and approved this research at Institution A (Reference Number RH12524) and Institution B (Reference Number 017040S).

Kate Hartup
16 February 2022

Abstract

For as long as numeracy has been assessed as a major domain in Australia (since 2003), Australian school students' numeracy levels have been declining (Thomson et al., 2019). This decline is evident despite it being a national requirement for teachers to teach numeracy skills across all subject areas at all year levels (Australian Curriculum and Reporting Authority (ACARA), n.d.).

Although there are several definitions of numeracy, fundamentally it is the application of mathematics and the ability to interpret and utilise mathematical information (Australian Council for Educational Research (ACER), 2017). Given that teachers' mathematical knowledge has been shown to affect their students' performance in the classroom (Shirvani, 2015; Tchoshanov et al., 2017), it is possible that teachers' personal numeracy capabilities also have an impact on their students' numeracy skills. To that end, it is essential that initial teacher education providers have mechanisms to evaluate teacher education students' (TES) numeracy capabilities and provide support for their development. To date, limited research has been conducted to evaluate TES' numeracy capabilities during their tertiary education (Callingham et al., 2015; Forgasz & Hall, 2019; Sellings et al., 2018). As such, this research addressed this need by exploring the following research question: To what extent is the use of an online Diagnostic Test associated with improvement in TES' numerical skills?

This study is positioned within the positivist paradigm, which is guided by the belief that real life is scientific and phenomena are measurable (Brown & Baker, 2007), and explored numeric measures of the numeracy skills of TES. The study adopts overlaps identified between commonly used online learning theoretical frameworks (the Technology Acceptance Model (TAM), the Analysis, Design, Development, Implementation, and Evaluation (ADDIE) Model, and the e-learning systems framework of Aparicio et al. (2016)) and the Assessment *for* Learning (A*f*L) theory of Black and Wiliam (1998) to embed into the design of the research. Furthermore, the elements described in the more recently proposed triangulated A*f*L framework by Tan (2013) were used to interpret the findings and determine whether numeracy skills can be improved through the use of an online practice test.

Test and Assessment was chosen as the methodology as it is commonly used to measure achievement and potential (Cohen et al., 2011). This methodology guided the development of the online test as the data collection instrument. Initially, a Pilot Test was conducted consisting of 40 numeracy items. Based on the Pilot Test results, a total of 272 questions were developed for the main Diagnostic Test. For each attempt in the main Diagnostic Test, 40 questions were drawn from the larger pool of items which were categorised into mathematical strands (Number and Algebra (NA), Measurement and Geometry (MG), Statistics and Probability (SP), and Non-Calculator (NC)), and then sub-divided into mathematical content areas following the Australian Curriculum (ACARA, n.d.). Items were further categorised into item type (Fill in the Blank, Multiple Choice, and True/False), their context domain (Personal and Community, Workplace and Employment, and Education and Training) according to the ACER Literacy and Numeracy Test for Initial Teacher Education (LANTITE) Assessment Framework (ACER, 2017), and their level of difficulty (Level 2-5) according to the Australian Core Skills Framework (ACSF) levels (Department of Employment, 2015). A specific number of items was drawn from each pool of questions (strand and content) to ensure that an even spread of mathematical areas were presented on each test attempt. Furthermore, worked solutions were developed for each item and displayed at the completion of each test attempt for the items that had been answered incorrectly on that attempt.

Quantitative data were collected from two Australian universities (Institution A and Institution B) between March 2018 and March 2019 using the testing module through the Blackboard Learning Management System. Overall, there were 1283 attempts made (n=878 for Institution A, and n=405 for Institution B). The data were analysed using both raw score analysis and also through the application of the Rasch Measurement Model to ensure valid comparisons could be made between test attempts.

Findings from this study describe the extent to which TES' numeracy skills can be evaluated and developed using an online practice test. It is expected that knowledge gained from this research will provide a model for initial teacher education providers to understand the skills TES are taking to the classroom and also the skills in which they require further development. The results will allow for further, more targeted, teaching and learning strategies to be implemented into university education programmes. Finally, this research provides a method of assessing, tracking and developing TES' numeracy skills. In the long term, this research will benefit schools by having increasingly more numeracy-competent teachers educating Australian students.

Acknowledgements

My appreciation goes to those who have supported me through the development of this thesis and encouraged me to reach the finish line. My gratitude also goes to the Australian Government – for the scholarship provided through the Research Training Program (RTP) scheme. Without this support, I would not have been able to undertake this study.

I would like to thank my principal supervisor, Doctor Thuan Thai, for your patience, ongoing support, and consistent encouragement. You ensured that I remained focused throughout the entire process, guided me to improve my work, and provided years of mentorship beyond my expectations. Your positivity and belief in me were the greatest motivations. For everything you did, I am so very grateful.

When I reflect upon my journey, I often think of how it all began and my appreciation goes to those who helped me in the initial phase. Firstly, I would like to give thanks to the late Doctor Frank Moisiadis. Frank inspired me to commence a research degree, and I would not be where I am today without Frank's encouragement. Although we did not see a research project through together, I know he would have been very proud of what I have achieved. Secondly, I would like to thank Professor Boris Handal for the guidance provided to me in the early phase of my degree. Thank you, especially, for recommending that I be involved in the initial development of a numeracy practice test. This is where I found my passion, and I am truly appreciative that you gave me this opportunity.

For their significant contributions to the project, I would like to thank Doctor Amanda Yeung and Mrs. Adelle Colbourn. Thank you to Amanda for guiding me through the data analysis process and being a supportive and knowledgeable co-supervisor. To Adelle, thank you for your assistance in the development of the testing instrument. Your insights were so helpful, and I also appreciate the time you gave up to present at conferences with me over the past few years.

To the people in my life who have inspired me, believed in me, listened to me, and cared for me, thank you so much. I truly appreciate the love and support you have shown, especially towards to end of this journey. A part of this is dedicated to you.

In particular, thank you to my wonderful family, who are a never-ending source of love and support. To my mum and dad, a very special thank you for the babysitting, the meals, the washing, and the belief you have always had in me to achieve anything I want. To my brothers, Daniel and Matthew, thank you for always being there for me. You have always supported and encouraged me in everything I do, and I appreciate this so much.

Finally, last but certainly not least, thank you to my husband, Ryan, and my three beautiful children, Jayde, Zayn, and Liana. Ryan, thank you for your patience, always being there to listen, and taking on so many of the 'mum' duties when I could not. You never let me give up, always found a way to balance my stress, and you helped me achieve this more than you realise. To Jayde, Zayn, and Liana, you have all been so patient with me, and even though you may not have entirely understood what I was doing, you knew that it was important to me. I hope I have shown you all that you can achieve any goals you set in life. My beautiful family, I did this for you!

List of Publications and Presentations

Publications:

Thai, T., Hartup, K., Colbourn, A., & Yeung, A. (2021). Using an online Numeracy practice test to support education students for the numeracy component of the LANTITE. *Australian Journal of Teacher Education.*

Presentations:

Hartup, K., Thai, T., Colbourn, A., & Yeung, A. (2019). An online approach to track and evaluate teacher education students numeracy capability. *Education Scholar Conference: Assessment in Times of Change*, Paper presented at the Annual University of Notre Dame Australia Education Scholar Conference (Sydney, NSW, 2019).

Thai, T., Hartup, K., Colbourn, A., & Yeung, A. (2019). Learning from their mistakes: An online approach to evaluate teacher education students' numeracy capability. *In G. Hines, S. Blackley, & A. Cooke (Eds.), Mathematics education research: impacting practice: Proceedings of the 42nd annual conference of the Mathematics Education Research Group of Australasia* (pp. 707-714). WA: MERGA.

Thai, T., Hartup, K., Colbourn, A., & Yeung, A. (2018). A test to test test-readiness – improving teacher education students' numeracy skills development. *In J. Hunter, P. Perger, & L. Darragh (Eds.). Making waves, opening spaces: Proceedings of the 41st annual conference of the Mathematics Education Research Group of Australasia* (pp. 744). Auckland: MERGA.

Thai, T., Hartup, K., Colbourn, A., & Yeung, A (2018). An online approach to improving teacher education students numeracy skill development. *Education Scholar Conference*, Paper presented at the Annual University of Notre Dame Australia Scholar Conference (Sydney, NSW, Oct 19, 2018).

Thai, T., Hartup, K., Colbourn, A., & Yeung, A. (2017). Is it too late to master your numeracy when the days in your masters are numbered? *First year in Maths NSW*, Paper presented at the First Year in Maths (FYi Maths) Workshop (Sydney, NSW, Dec 8, 2017)

Table of Contents

List of Figures

List of Tables

Definitions of Terms

A*a*L	Assessment as Learning
ACARA	Australian Curriculum and Reporting Authority
ACER	Australian Council for Educational Research
ADDIE	Analysis, Design, Development, Implementation, and Evaluation Model
A*f*L	Assessment for Learning
AITSL	Australian Institute for Teaching and School Leadership
A*o*L	Assessment of Learning
CBEST	California Basic Educational Skills Test
ICT	Information Computer Technology
LANTITE	Literacy and Numeracy Test for Initial Teacher Education
LMS	Learning Management System
OECD	Organisation for Economic Co-operation and Development
MG	Measurement and Geometry
NA	Number and Algebra
NAPLAN	The National Assessment Program – Literacy and Numeracy
NC	Non-Calculator
NESA	NSW Education Standards Authority
NSW	New South Wales

PIAAC	The Programme for the International Assessment of Adult Competencies
PISA	The Programme for International Student Assessment
SP	Statistics and Probability
TES	Teacher Education Students
TAM	Technology Acceptance Model
TIMSS	Trends in International Mathematics and Science Study
UK	United Kingdom
US	United States of America

Chapter One: Introduction

1.1 Overview

Individuals need to possess adequate numeracy skills considering the many benefits that exist (Bynner et al., 2001). For example, research suggests that the application of numeracy skills allows for well-informed and calculated decisions that will better equip a person to make the best choices in life (Reyna et al., 2009). Specifically, in Australia, it has been reported that adults require high numeracy levels to self-manage certain areas of their lives (Hartley & Horne, 2006). However, despite the need for Australians to possess adequate numeracy skills, it has been identified that many Australian adults display very poor numeracy skills (Organisation for Economic Co-operation and Development, 2013). Furthermore, the numeracy skills of Australian students are also increasingly becoming a concern with results from international standardised tests confirming that skills of Australian students are declining in comparison to many other countries (Australian Council for Educational Research (ACER), 2018a; ACER, 2018b; Baker, 2019; Thomson et al., 2019). Fortunately, these issues have been recognised in the Australian Curriculum, and it is now a national requirement for teachers to teach numeracy skills across all subject areas at all year levels (Australian Curriculum and Reporting Authority (ACARA), n.d.). However, since this requirement has been introduced, Australian students are still underperforming in numeracy internationally, and only minor improvements have been evident in The National Assessment Program – Literacy and Numeracy (NAPLAN) (McGaw et al., 2020). It is crucial to address the issue of Australian students' declining worldwide numeracy performance and aim for further, more significant, improvements in NAPLAN.

Given that teacher knowledge is an essential element that informs preparation and teaching (Shulman, 1987), it is reasonable to expect that all teachers have an understanding of numeracy concepts in sufficient depth to be able to teach these skills. Furthermore, evidence suggests that teachers' mathematical content knowledge affects their students' performance (Shirvani, 2015; Tchoshanov et al., 2017). Therefore, it is possible that teachers' personal numeracy capabilities also impact their students' numeracy skills. To that end, it is crucial for initial teacher education providers to gain a thorough understanding of the numeracy skills that

Teacher Education Students (TES) possess during their tertiary studies and as they prepare to enter into teaching careers. This is especially important to ensure that TES are graduating with adequate numeracy capabilities that will allow them to teach these skills to their students successfully.

Although there are many definitions of numeracy, fundamentally, it is the application of academic mathematics skills and the ability to interpret and utilise mathematical information (ACER, 2017). It has been suggested that the terms numeracy and mathematics should not be used interchangeably (Zevenbergen, 2004); however, there are overlaps between the two. For example, numeracy has been described as the ability to use mathematical skills to cope confidently with the practical demands of everyday life (Cockcroft, 1982). It has also been defined in terms of the mathematical skills, concepts, and procedures that one must know and understand (Willis, 1998). More recently, numeracy has been described as a bridge connecting mathematics at school and the real world (Park, 2010). Therefore, it is important to acknowledge that being numerate is not solely an academic pursuit and is also a skill required in broad areas of our lives, depending on the situation, context, and purpose.

Considering the importance for individuals to possess adequate numeracy skills that will allow them to successfully carry out activities relevant to their life, for this research, which explored numeracy skills of TES, we adopted ACER's (2017) definition of numeracy: "interpreting and communicating important non-technical mathematical information and using such information to solve relevant real-world problems to participate in an education community, to achieve one's goals, and to develop one's knowledge and potential as a teacher" (p. 21). This definition describes the overall capabilities required for teachers and TES to be considered numerate.

Little is known about TES' numeracy skills during their tertiary education. With the few studies that have been conducted (Afamasaga-Fuata'i et al., 2007; Linsell & Anakin, 2012), it is evident that not all TES possess the requisite numeracy capability, which has implications on their future students' learning and on their ability to address numeracy demands in the curriculum. Furthermore, studies that have investigated TES' numeracy skills throughout their program of study have shown only minor improvements have occurred and overall, the coursework and program of study does not address this problem (Callingham et al., 2015; Sellings et al., 2018). Therefore, it is worth exploring other ways to assess and support TES' numeracy development.

Given that formative assessments have been reported to enhance learning in many contexts (Black & Wiliam, 1998; Boud, 2000; Clark, 2012), and that online resources has been reported to improve mathematics learning (Alcoholado et al., 2016), this study explored the use of online formative assessments as a way to evaluate TES' numeracy capabilities and support their development of these skills. In developing this online formative assessment, this study adopted the theory of Assessment *for* Learning (A*f*L) and incorporated the elements that have been reported in the literature to improve learning.

The research described in this thesis considered the declining numeracy skills of students in Australia, the benefits of being numerate, the link between teachers' knowledge and student performance, and the existing literature that suggests TES' numeracy capabilities may be weak. Therefore, the research arose from concerns about teachers' numeracy capabilities in Australia that could impact their students' numeracy performance; thus, it was of particular interest to examine TES' numeracy capabilities to determine the current state of their skills before they become teachers. Additionally, it was of interest to establish a method of assessing, tracking, and improving TES' numeracy skills and, more specifically, determine whether TES' numeracy skills could be improved through the use of an online test. To achieve this, the research identified and evaluated TES' numeracy skills from two Australian universities using learning analytics captured through a numeracy skills practice test on the Blackboard Learning Management System (LMS). This thesis describes this exploration, reports on significant findings, and provides conclusions and recommendations that will provide initial teacher education providers with a better understanding of their TES' numeracy capabilities and approaches to develop these skills.

1.2 Purpose

The purpose of this study is to explore TES' numeracy capabilities and the extent to which an online test can be used to support TES to develop these numeracy skills, which is currently lacking in the research literature. In fact, although internationally there have been a few studies that have investigated TES' numeracy capabilities (Afamasaga-Fuata'i et al., 2007;

Linsell & Anakin, 2012), there has not been any research conducted in this area in Australia. Given that the requirement for all teachers to teach numeracy is a recent initiative in Australia, the few current studies that exist have only explored TES' numeracy skills through these new initiatives (Callingham et al., 2015; Forgasz & Hall, 2019; Sellings et al., 2018).

The majority of existing research focuses on primary and secondary school students' numeracy capabilities (ACER, 2018a; ACER 2018b; Baker 2019) or pedagogical practices that enhance students' understanding (Stronge et al., 2011). Studies that have evaluated TES' mathematical and/or numeracy skills have typically focused on their coursework achievements (Norton, 2019), which have reported mixed results in terms of improving TES' numeracy capabilities. For example, benefits were found by Forgasz and Hall (2019) who explored the implementation of a compulsory unit at Monash University. Their findings showed that TES' understanding and confidence to implement numeracy into their teaching improved after completing the unit. On the other hand, Callingham et al. (2015) explored TES' numeracy capabilities through a unit conducted in the first semester of a teaching degree and found that not all TES recognised the nature of numeracy and the prospects of its development. Similarly, Sellings et al. (2018) evaluated an institution initiative developed to improve TES' numeracy skills and found that not all TES improved their test results due to interventions. Sellings et al. (2018) also reported that after the intervention, some students performed lower on the subsequent test.

Given the importance of numeracy in everyday life, the requirement of all teachers to teach numeracy skills as part of the Australian Curriculum, and the impact that teachers' mathematical knowledge has on student achievement, it is necessary to understand the current state of TES' numeracy capabilities so that measures can be put in place to support the development of their skills before they enter their teaching careers. Importantly, these measures should be sustainable and go beyond individual units of study, such that TES may continue to refine and develop their numeracy capabilities throughout their tertiary education.

This study will evaluate the numeracy capabilities of TES at two Australian universities, enrolled in an initial teacher education program. Participants include students in undergraduate and postgraduate programs.

1.3 Potential Significance

It is important to consider how the research could advance the field of TES' numeracy capabilities and development. To that end, the significance of the research was thoroughly considered, and it is expected that the knowledge gained from the research described in this thesis will make numerous contributions to the literature of TES numeracy capabilities. Furthermore, it is anticipated that the research will impact and benefit Australian TES, initial teacher education providers, schools, and ultimately primary and secondary school students.

Firstly, it is expected that TES will improve their numeracy skills by participating in the online practice test designed through the adoption of elements identified in online learning theoretical frameworks and the A*f*L theory. Through the repeated use of the online test, this study aims to support TES to improve their numeracy capabilities to ensure their skills are adequate before they graduate and become teachers. Furthermore, it is now a regulatory requirement that all TES in Australia must pass the Literacy and Numeracy Test for Initial Teacher Education students (LANTITE) to demonstrate a satisfactory level of literacy and numeracy skills prior to graduation (ACER, 2018c). To that end, it is expected that TES will use the test in this study to determine their readiness to attempt the numeracy component of the LANTITE. Through use of the test, it is anticipated that TES will gain an understanding of their current skills and make appropriate decisions about whether their skills are developed enough to achieve the required standard.

Secondly, it is expected that knowledge gained from the research described in this thesis will enable initial teacher education providers to understand the skills TES are taking to the classroom and the skills in which they require further development. This will allow further, more targeted teaching and learning strategies to be implemented into tertiary education programs. Gaining a thorough understanding of TES' numeracy capabilities during their initial teacher training is especially important so appropriate measures can be put in place to develop their skills before becoming teachers.

Since teachers' personal numeracy capabilities can impact their students' numeracy skills, teachers must have adequate numeracy skills. Considering this research aims to provide a method of developing TES skills before they graduate, this research could support the development of these skills. Therefore, thirdly, in the long term, the knowledge gained from this research could assist in improving TES numeracy skills that will ensure increasingly more numeracy competent teachers are educating Australian students.

1.4 Research Questions, Aims and Hypothesis

As outlined previously, the research described in this thesis aimed to evaluate the numeracy capabilities of TES at two Australian universities and explore whether an online test can be used to enhance learning. Therefore, the following research question guided the study: To what extent is the use of an online Diagnostic Test associated with improvement in TES' numerical skills?

To answer the above research question, the study explored the following sub-questions:

1. To what extent can an online test be used to diagnose TES' numeracy capabilities?
2. What are TES' numeracy strengths and weaknesses?
3. To what extent can an online diagnostic test be used to improve TES' numeracy skills?

The research questions are now discussed in more details.

1.4.1 To what Extent can an Online Test be used to Diagnose TES' Numeracy Capabilities?

To address research question one, the accurate diagnosis of TES numeracy capabilities were explored through an online practice test, not associated with any coursework unit. This research question also prompted the assessment of whether TES' capabilities can be examined by test sections (Calculator-allowed and Non-Calculator (NC)), mathematics strands, topics, item types, context domains, and the Australian Core Skills Framework (ACSF) levels. Additionally, this research question was addressed by exploring whether capabilities can be examined and diagnosed using TES' first attempt, subsequent attempt, and final attempts of the online Diagnostic Test.

1.4.2 What are TES Numeracy Strengths and Weaknesses?

To address this research question, specific numeracy strengths and weaknesses of TES were evaluated and trends in capabilities were explored. Capabilities were explored in greater detail through examining TES’ strengths and weaknesses in the three mathematical strands and specific content areas. This research question also prompted the assessment of whether there are specific areas of strength and weakness in the types of numeracy questions TES are presented with, for example, in item types (Fill-in-the-Blank, Multiple-Choice, and True/False items), in the context domains of the questions, and the ACSF level of the questions.

1.4.3 To What Extent can an Online Diagnostic Test be used to Develop TES’ Numeracy Skills?

Research Question Three acknowledges that online tests often used repeated times and can be designed specifically for this purpose. To address this research question, we explored whether improvements can be made through repeated use of the online test and compared performance and ability on students’ first attempt and subsequent attempts of the Diagnostic Test. Changes in estimated student ability on the questions were also explored to determine whether improvements can be made for questions in all areas, such as test categories, topics, item-types, context domains, or ACSF levels. Here, an investigation of the effectiveness of an online Diagnostic Test as a formative A*f*L tool for improving TES’ numeracy capabilities was also undertaken.

1.4.4 Hypothesis

Given that online formative assessments have been shown to improve self-paced learning and improvement, it was hypothesised that the online Diagnostic Test would improve TES’ numeracy skills development and provide a rich source of learning analytics data.

1.5 Outline of Chapters

The thesis is comprised of eight chapters:

Chapter One details an overview of the research. This chapter describes the purpose and significance of the study, outlines other existing initiatives that have been implemented to improve TES numeracy capabilities, and the gap in the research literature. Additionally, Chapter One introduces the research questions and discusses them in detail, provides the research hypothesis, and an outline of the thesis' chapters.

Chapter Two provides a review of the significant literature that informs the scope of the study. The literature reviewed in this chapter includes definitions of numeracy, the benefits of being numerate, the current state of numeracy capabilities in Australia, and numeracy requirements in the Australian Curriculum. This chapter also explores TES' numeracy skills and the mathematical skills they possess that suggest numeracy competency. Furthermore, the chapter discusses existing approaches to assess TES' numeracy competencies and initiatives that have been implemented to improve capabilities. Finally, Chapter Two concludes with an examination of online diagnostics tests and their potential to improve and track learning.

Chapter Three considers that the main elements of the research and acknowledges that an appropriate framework will need to encapsulate the development of TES' numeracy skills as well as self-paced online learning through assessment. This chapter initially provides a systemic review of considered theoretical frameworks in the study of online learning and separately examines theories of assessment to discern overlaps. The roles of assessment, types of assessments, developments of A*f*L, and Black and Wiliam's (1998) development of the A*f*L theory are discussed. The chapter concludes to identify overlaps between models of online learning and Black and Wiliam's (1998) A*f*L theory. Finally, a discussion of how the overlapping elements are specifically adopted in this research is provided and Tan's (2013) extension to the A*f*L model used to interpret findings in this thesis is discussed.

Chapter Four describes the methodology used for this study and explains the choice of techniques used for the methods adopted. This chapter explains how the research is situated in the positivist paradigm and describes the design of the research. A detailed description of the development of the Pilot Test is provided and the main Diagnostic Test used to collect the data, including an outline of how the common elements of the A*f*L theory and the online learning models are incorporated in the test design. Additionally, the techniques of analysis used are explained in detail. The chapter concludes with the limitations to the study's design and the ethical considerations.

Chapter Five outlines the results of the Pilot Test. This chapter presents the raw score results of the test and discusses the findings to assess the functionality of the testing system. Chapter Five also discusses the results of the open-ended questions given to students to gather feedback about the usability of the test. This chapter concludes with the overall findings from the Pilot Test and insights that informed the development of the main Diagnostic Test.

Chapter Six presents the findings from analysing the raw scores from the main Diagnostic Test. This chapter presents the number of attempts made by students at the two institutions and documents the result of the overall test performance at each institution. This chapter also reports the findings of TES' capabilities in specific content areas and the trends observed from the data. Chapter Six concludes with a discussion on the findings and compares the results obtained from this research and findings from the existing literature.

Chapter Seven presents findings after applying Rasch analysis to the data. This chapter presents an initial attempt and item overview, details the analyses and methods used to improve the data's Rasch Model fit, and presents the results once item anchors have been generated and applied to the data. Chapter Seven also presents the findings of comparisons between test attempts and identifies areas of improvements (or regression) in specific test categories. This chapter concludes with a discussion on the changes in TES' numeracy capabilities that were evident in the findings.

Chapter Eight provides a final discussion and conclusions to the research. The chapter discusses the research questions with respect to the findings produced and provides possible explanations for the results observed. Implications of our research findings are discussed, highlighting the utility of an online diagnostic test as a form of A*f*L to support TES' numeracy development. Finally, recommendations are provided, including the proposal of a potential extension of Tan’s (2013) A*f*L Model, with suggestions for future initiatives and research to ensure the successful development of TES’ numeracy skills.

Chapter Two: Literature Review

2.1 Introduction

This chapter presents the literature review that informs the scope of the study, points to the significance of the issue while also highlighting the gap in the literature and the need for the study. Central ideas reviewed in the literature include definitions of numeracy, the benefits of being numerate, the current state of numeracy capabilities in Australia, and the numeracy requirements in the Australian Curriculum. This chapter also explores TES' numeracy skills and the mathematical skills they possess that suggest numeracy potential. Furthermore, the chapter discusses tests that exist to assess TES' numeracy competencies, initiatives that have been implemented to improve TES' numeracy capabilities, and examines online diagnostics tests and their potential to improve and track learning. The chapter concludes with a summary of the reviewed literature arguing for the need to explore online diagnostic tests as a strategy to evaluate, track, and improve TES' numeracy skills.

2.2 Numeracy

The term numeracy, initially introduced in the United Kingdom (UK) Crowther Report published in 1959 (Cockcroft, 1982), is often referred to as mathematical literacy or quantitative literacy. Many scholars consider that to be numerate, there are various important requirements. For example, Cockcroft (1982) suggested that being numerate implies being comfortable to use numbers, making use of mathematical skills to cope with the demands of life, and having the ability to understand and appreciate information that are presented mathematically. In particular, Cockcroft (1982) highlighted that to build numeracy skills, consideration should be given to these broader aspects of numeracy that are not simply confined to the development of mathematical computation skills. Other requirements of being numerate were more recently outlined by Perso (2006) who stated that knowledge and understanding of mathematics are necessary to be numerate; however, possessing mathematical skills is not a sufficient determinant. Instead, Perso (2016) described that in addition to being comfortable using numbers, being numerate involves having the confidence to choose processes and apply mathematics appropriately.

It is important for numeracy and mathematics to be recognised as different skills. That is, numeracy is the application of mathematics and the ability to interpret and utilise mathematical information (Australian Council for Educational Research (ACER), 2017), whereas, mathematics is the science of numbers and their operations, quantity, and space (Vinner, 1991). Although it has been suggested that the terms should not be used interchangeably (Zevenbergen, 2004), many definitions of numeracy highlight that there are obvious overlaps between the two. For example, the connection was evident even in the early definition provided by Cockcroft (1982), who described numeracy as the ability to use mathematical skills to cope confidently with the practical demands of everyday life. In later definitions, this connection remained apparent. For example, Willis (1998) outlined that numeracy can be described in terms of the mathematical skills, concepts, and procedures that one must know and understand. Similarly, in more recently definitions, it has been stated that some mathematics must be known for one to be numerate (Perso, 2006). Since then, the definition of numeracy has evolved to more generally, the ability to use mathematical skills efficiently to deal with the quantitative aspects of life (Park, 2010). Park (2010) argued that numeracy is the "bridge which connects school mathematics and the real world that lies behind and beyond school" (p. 18). In fact, Park (2010) suggested that a traditional mathematics curriculum could actually limit potential, whereas numeracy increases awareness of how and why the skills can be applied in everyday life. Hence, Park's (2010) definition highlighted that numeracy skills are required to cope with certain areas of life.

Various numeracy definitions acknowledge the requirements of applying mathematics in life. Therefore, it is essential to acknowledge that being numerate is not solely an academic pursuit and is a skill required in broad areas of our lives, depending on the situation, context, and purpose. For this reason, to date, the definition of numeracy continues to evolve. In light of the technology-driven world that now exists and its importance in everyday function, Kus (2018) argued that the definition of numeracy has shifted from applying mathematics in other subject areas to focusing on understanding our data-rich society. In other words, numeracy is no longer just about the ability to perform mathematical calculations in everyday situations but also includes understanding how data are captured, processed or analysed, and involves the interpretation of results allowing for adequate problem solving.

There are many benefits of being numerate; for example, the application of numeracy skills allows for better-informed, calculated decisions, and the development of such skills will better equip the person to make the best choices in life (Peters et al., 2006). Further to civic activities, the benefit of being numerate also applies to post-secondary school education and work. For example, Galligan and Hobohm (2015) stated that many university courses require certain numeracy skills or a certain level of numeracy skills. However, many students struggle to meet this standard because they do not have the required capabilities. Galligan and Hobohm (2015) suggested that university students who do not have sufficient numeracy skills are less likely to be successful in their course. Evidence also suggests that highly educated individuals with more developed numeracy skills earn higher wages and, therefore, have more financially successful careers (Shomos, 2010). More specifically, Shomos (2010) used models to estimate the effect of improved numeracy skills on the probability of labour force participation and income to show that improving numeracy skills positively affected labour market outcomes.

Importantly, being numerate also contributes to understanding and interpreting health information that allows critical life decision-making (Reyna et al., 2009). For example, low numeracy capability has been reported to hinder people's understanding of the risks of breast cancer and the benefits of mammography (Schwartz et al., 1997). In this study, Schwartz et al. (1997) assessed older women's understanding of basic probability and numerical concepts and compared these skills to their ability to adjust their perceived risks after being presented with risk reduction (numeracy-related) data. Only 16% of the respondents answered all numeracy problems they were presented with correctly, 46% could not answer a basic probability question correctly, less than half could not convert a percentage to proportion, and 80% could not convert proportion to a percentage. Additionally, participants who displayed low accuracy in applying risk reduction information were associated with low numeracy capabilities. The findings from this study suggested that critical decisions might be made by people who do not adequately understand the information they were presented with. Being able to make informed and critical life decisions involving an understanding of data, highlights the importance of good numeracy skills to participate in all areas of life.

Additionally, there are many other reasons for individuals to possess a good level of numeracy. For example, Bynner et al. (2001) reported that individuals with improved basic skills, improved their chances in the labour market, suffered less from poor physical health, were less likely to have children experiencing difficulty at school, were more likely to be active in society, voted, and expressed interest in politics, and had less discriminatory attitudes. It has also been reported that people are increasingly needing to self-manage areas of their lives that require high numeracy levels (Hartley & Horne, 2006). This includes being able to negotiate contracts, make retirement decisions, and manage health conditions. Other tasks that require applying mathematical skills include deciding on a mortgage, renovating and decorating a room, and budgeting (e.g., planning a holiday or going shopping). Given the role that numeracy plays in these financial, legal, health and life decisions, it is important to examine the current state of numeracy competencies in Australia.

2.3 Australian Adults' and Students' Numeracy Competencies

Although the importance for children and adults to possess adequate basic numeracy skills is widely recognised, performance in numeracy is a concern internationally. For example, when analysing data of 16-19 year olds from the Survey of Adult Skills (2012), the OECD found that almost 40% possessed low numeracy levels in the US, approximately 30% possessed low numeracy skills in England, and between 20-30% of participants possessed low numeracy skills in Ireland, Italy, Spain, France, and Canada. The best performing countries were Korea and Netherlands, both displaying less than 10% of participants possessing low numeracy skills (World Economic Forum, 2016).

In Australia, performance in numeracy has continued to decline compared to many other countries (ACER, 2014; Thomson et al., 2019). However, the issues lie beyond school students' numeracy capabilities that are frequently discussed in the media, and the numeracy skills of adults in Australia are also a major concern. For example, Australian adults' numeracy capabilities have been reported to be lower than many other countries (OECD, 2013). This was evident in the 2011-2012 Programme for International Assessment of Adult Competencies (PIAAC), where the Survey of Adult Skills was administered to 7430 Australians aged 16-65 to gather information about their ability to use numerical concepts. In this study, numeracy and mathematical concepts were assessed, and proficiencies were described in terms of a scale divided into six levels: Below Level 1, Level 1, Level 2, Level 3, Level 4, and Level 5. For example, an adult below Level 1 is able to carry out simple processes such as counting and performing operations with whole numbers. At Level 1, an adult can perform basic mathematical processes, basic arithmetic operations, and understand simple percentages. An adult at Level 2 can apply two or more steps involving whole numbers and partial numbers and understand simple measurement concepts and data representation. At Level 3, an adult has a good sense of number and space, can recognise mathematical relationships, patterns, proportions, can interpret and perform basic data analyses. An adult at Level 4 has a broad range of mathematical information understanding that may be complex or in unfamiliar contexts. At the highest level, at Level 5, an adult can integrate multiple types of mathematical information and draw inferences to develop mathematical arguments (OECD, 2013).

Results from this study found that 13.3% of Australian adults attained Level 4 or Level 5, which was similar to the average of 12.4% across all participating countries. Level 3 was attained by 32.6% of Australian adults, which was slightly below the average in all participating countries (34.4%). However, most concerningly, a substantial proportion of the Australian population (20.1%) were found to have very low levels of numeracy skills in the below Level 1 range. This was slightly higher than the average of 19%. These results suggested that a large proportion of young Australians only have the most basic numeracy and mathematical capabilities. Furthermore, when the percentage of Australian adults scoring Level 3 and Level 4/5 were combined and ranked against the other participating countries, Australia was ranked below average. Overall, Australian adults were found to have a lower proportion of Level 3, 4, and 5 proficiencies than those from other countries such as Finland, Japan, Sweden, and the Netherlands. However, proficiencies were found to be similar to adults in Canada, Germany, and Poland, and higher than those in Spain, Italy, and the United States (OECD, 2013). Since, overall, a poor level of numeracy skills was identified amongst the Australian adult population, it is worth reviewing Australian students' numeracy competencies.

Internationally, students' mathematical and numerical knowledge is assessed on a very large scale through standardised tests administered to thousands of students. One example is the Programme for International Student Assessment (PISA) administered every three years to students aged 15 years, as they near completion of their compulsory schooling. ACER (2018b) outlined the major benefit for Australia's participation in this test being the opportunity to compare student performance on a global scale to provide insight to help improve our education system. Similarly, Starr (2014) outlined the many benefits for Australia's involvement in this test; however, also pointed out that a common criticism of the test is that teachers already know what students can and cannot do, and therefore the test is a waste of time. Despite the controversy that exists with participation in global standardised tests, they are useful in providing a snapshot of students' knowledge, track the performance of participating nations over time, and to compare results between participating nations.

Overall, the numeracy skills of Australian students are increasingly a concern. In particular, results from these international standardised tests have confirmed that Australian students' skills are declining compared to many other countries. For example, the Trends in International Mathematics and Science Study (TIMSS) identified that achievement of Year 4 and Year 8 students has not improved in 20 years (ACER, 2018a). Displaying even more concern are the results from the PISA, which showed a consistent decline in student numeracy competency between 2003 and 2018 (ACER, 2018b). In addition, mathematics results from the 2018 PISA indicated that students did not exceed the average for the first time. Australian students' results displayed one of the largest declines of all countries, and compared to Singapore, students were considered to be three years behind (Baker, 2019). In summary, Australian achievement trends are decreasing, the number of Australian students achieving national proficient standards is declining, and Australia's position in the worldwide ranking is falling (ACER, 2018b).

Exclusively in Australia, primary and secondary school students' numeracy abilities are measured through NAPLAN. Both the government and community have given significant attention to this test, however, like other major standardised tests, there are many criticisms for NAPLAN. For example, some disapprove of the significant funding for the development and implementation of the test (Conyers & Scott, 2012) while others dispute the relevance and appropriateness for all students involved (Perry et al., 2012). To justify the relevance of NAPLAN, ACARA (2012) argued that processes have been put in place to ensure that it is a valid and reliable measurement of students' numeracy skills and noted that it should be acknowledged that the tests allow for an accurate assessment of knowledge and monitoring of development and progress that was previously not able to be performed.

Despite the decline in Australian students' performance on international tests, small improvements in NAPLAN results have been evident in the last decade (McGaw et al., 2020). For example, when comparisons of NAPLAN performance between 2008 and 2019 were made to determine improvements, estimates of the statistical significance of differences and effect sizes were provided. For example, when differences were statistically significant and effect sizes were between 0.2 and 0.5, the change in performance was described as a moderate increase. When the effect size was greater than 0.5, changes were described as substantial. From this analysis, moderate increases in mean scores and national mean standards were determined in Year 5, and moderate increases were determined in the national mean standards in Year 9. Although these results suggested recent improvements in some NAPLAN results, it is important to acknowledge that the improvements were small. In fact, there was no substantial increase determined at any level. Even more concerning, no significant changes in performance were evident at all in Year 3 and Year 7 numeracy achievement; although, Year 7 results were seen to display an increase in the proportion of students at or above the national mean standard. Since the improvements in numeracy achievement between 2008 and 2019 were only small, and not evident across all levels, overall, the concern of Australian students' numeracy capabilities remains and greater improvements at all levels are necessary.

Considering that low numeracy capability is an issue for Australian students and adults, and that being numerate is necessary to adequately participate in the everyday demands of adult life, addressing Australian students' declining numeracy performance should be major concern and required more attention. Therefore, it is important to consider how the Australian Curriculum support students' numeracy skills development.

2.4 Numeracy within the Australian Curriculum

The importance for students to be numerate is recognised in the Australian Curriculum. It is intended that numeracy is developed in all subject areas and is recognised as one of the seven general capabilities described as playing an important role in preparing students to participate successfully in life in the twenty-first century (ACARA, n.d.). As such, the Australian Curriculum mandates a cross-curricula teaching and learning approach for numeracy, which is necessary for all areas of primary and secondary schooling. In other words, the development of students' numeracy skills is the responsibility of all teachers and not solely the responsibility of the mathematics teachers (Handal et al., 2014). For this reason, all teachers must understand numeracy concepts in sufficient depth to be able to teach these skills adequately.

In addition to the required implementation of numeracy as a general capability, there are also specific numeracy demands in each curriculum. For example, the mathematics curriculum has obvious numeracy demands that require the use of these skills to solve mathematical problems. In particular, the F-10 Mathematics Curriculum specifies that the proficiency strands of understanding, fluency, problem-solving, and reasoning are essential to the mathematics content and describe how the content is to be explored (ACARA, n.d.). These proficiencies highlight that the skills and knowledge covered in the Mathematics Curriculum are to be transferred to allow for the application of skills. As an example, in the Year 9 Level Description, it is outlined that a required reasoning proficiency is being able to use statistical knowledge to clarify situations (ACARA, n.d. This indicates that statistical analyses skills taught in Mathematics are also applicable to a variety of context, situations, and subject areas.

Science is another subject area with clear numeracy demands in the curriculum. As part of the Science inquiry strand in the F-10 Science Curriculum, there are many numeracy demands evident. In particular, in the content descriptions of Processing and Analysing Data and Information, and Evaluating, many numeracy applications are presented. For example, the Year 10 Curriculum outlines that students analyse patterns and trends in data, including describing relationships between variables and identifying inconsistencies (ACARA, n.d.). Furthermore, the elaborations outline that students are specifically required to present data in table and graph form, carry out mathematical analyses, describe sample properties (including mean, median, range), and explore relationships between variables. These requirements highlight that the application of mathematical skills is expected in this curriculum, confirming significant numeracy demands in the Science Curriculum.

The Design and Technologies Curriculum also has obvious numeracy demands. For example, as part of the Processes and Productions Skills content description in the Year 5 and 6 curricula, students are required to "generate, develop and communicate ideas and processes for audiences using appropriate technical terms and graphical representation techniques" (ACARA, n.d.). The elaborations of this content description include using modelling and drawing standards such as scales, symbols and codes in diagrams, and pictorial maps. Therefore, for students to appropriately represent and communicate their design ideas, they are expected to demonstrate numeracy skills that will allow them to perform graphical representation, interpret data displays, interpret maps and diagrams, and visualise shapes.

Numeracy demands also exist in other subject areas, such as Health and Physical Education. In particular, many numeracy requirements are evident in the movement and physical activity content in the Year 7 and 8 Curriculum. For example, it is outlined in the Understanding Movement description that students are required to measure heart rates, breathing rates, and the ability to talk to monitor the body's reaction to a range of physical activities (ACARA, n.d.). This requirement suggests that numeracy skills such as interpreting data displays, estimating and calculating, and working with different units of measurement (including time) are necessary. Furthermore, in the Moving our Body description, it is necessary for students to design and perform movement sequences to create, use and defend space. This suggests an implementation of spatial reasoning skills, and in particular, visualising shapes as well as recognising patterns and relationships.

Overall, the importance of numeracy is well recognised in the Australian Curriculum, and it is an expectation that all teachers must teach numeracy in their discipline areas, such as those outlined in the examples above. However, it has been suggested that opportunities to teach numeracy are often overlooked (Ferme, 2014). It has also been asserted that it is unlikely that all teachers would feel equipped to teach these numeracy concepts (Callingham et al. (2015). For example, it has been identified that a challenge exists for teachers to understand and accept the importance of numeracy and engage with numeracy in their subject areas (Callingham et al., 2015). It has also been suggested that this disengagement prevents teachers from carrying out necessary duties that are part of their teaching roles, for example, administering assessments (Wilson, 2013). In particular, Wilson (2013) suggested that some teachers' lack of confidence in dealing with numbers and data affects their ability to carry out assessments with their classes appropriately. If non-mathematics teachers are not confident to identify, use, apply, and teach numeracy, it is plausible to assume that they may not be adequately teaching these skills to their students. This line of reasoning has often been used to explain Australian students' declining numeracy skills, which has gained much attention from the Australian Government and wider community (Barnes & Cross, 2018).

2.5 TES’ Numeracy Capability

To date, there are only a few studies that have investigated TES’ numeracy skills. One example of this is a longitudinal mathematics project in Samoa, which aimed to monitor TES’ mathematical and numeracy competence through the use of two diagnostic tests (Afamasaga-Fuata’i et al., 2007). This study compared 46 TES’ performance on both tests to identify any developmental trends and to determine the extent of the impact of mathematics content and mathematics education courses on students’ numeracy levels. It was reported that there was a significant difference between the students' performance in the pre- and post-test, which suggested that numeracy skills could be improved over time. However, when analysing all results from the tests, the authors found that no one achieved mastery level, which was defined as achieving more than 80% in the diagnostic test (Afamasaga-Fuata’i et al., 2007). Further, the authors found that TES displayed many areas of misconceptions, such as fractions, decimals, geometry, measurement, and probability, which highlighted some numeracy areas in need of improvement.

Another study investigating TES’ numeracy capabilities was conducted in New Zealand by Linsell and Anakin (2012). In this study, two forms of diagnostic instruments, conventional (written) and adaptive (online), were administered to different cohorts to investigate whether professional numeracy standards were being met. Overall, the results showed that less than half of each cohort demonstrated foundation content knowledge and did not meet the professional standards. More specifically, 41% of TES passed the conventional test in 2010 (where a score of 75% was required to pass), and 43% of TES achieved at or above a pass mark in the adaptive assessment in 2011. This study, and the study conducted by Afamasaga-Fuata’i et al. (2007), display concerning results about TES’ professional standards of numeracy internationally.

More recent, research has explored TES’ numeracy capabilities in Australia. For example, as part of a three-dimension approach to developing teachers’ numeracy knowledge, Callingham et al. (2015) collected qualitative data from TES in Tasmania. Their study consisted of conducting interviews to understand TES’ numeracy knowledge and beliefs before and after undertaking a numeracy course. This study involved exploring three types of knowledge: mathematical, contextual, and strategic to understand the nature of numeracy that had been developed. Although TES were found to have expanded views after the course and developed a deeper appreciation of the numeracy involved in teaching, overall it was found that not all TES recognised the nature of numeracy and the importance of developing these skills. In fact, Callingham et al. (2015) found that many students believed a textbook would progress their numeracy understanding. This study suggested an overall lack of numeracy understanding, even after undertaking a course dedicated on exploring numeracy.

In another Australia study, Sellings et al. (2018) examined TES' literacy and numeracy abilities. The numeracy component was a 30-question multiple-choice test administered to 711 TES, which assessed concepts such as number, algebra, measurement, statistics, and probability. Results from this study showed that 156 (22%) did not meet mastery level in numeracy which was defined as being able to achieve 90% or more in the test. Furthermore, interventions were offered to those who did not meet mastery level and they were then required to re-sit a similar test with the same content but different questions. Despite the additional support and opportunity to retake a similar test, Sellings et al. (2018) reported that 82 TES still did not reach mastery level. Concerning, nine TES performed lower on their second attempt. It is worth noting that this study did not report on the specific areas of strengths or weaknesses; however, the results from this study suggested that a single intervention is not enough to support TES with low numeracy skills to address this problem.

More recently, Forgasz and Hall (2019) explored TES numeracy capabilities through a study that assessed numeracy abilities involving concepts of basic arithmetic, fractions, conversion of units, combinations, and interpreting data. The study's main aim was to evaluate a new numeracy unit implemented into the Masters of Teaching coursework at the University of Melbourne by gauging students' views of numeracy and determining changes in their perspectives before and after the unit. Additionally, the study investigated TES numeracy skills before they commenced the unit through a numeracy skills survey. There were six questions in the survey. Three of the questions were drawn from the 2010 Year 9 NAPLAN test, two from the 2012 PISA, and the sixth question was developed by the researchers. Overall, it was found that TES performed best in interpreting data, followed by basic arithmetic, unit conversions, and then fractions. However, the study also found that a weakness was evident in combinations (Forgasz & Hall, 2019) .

Despite several studies reporting that many TES do not reach numeracy mastery levels, very few studies exist that explore specific capabilities. It is, therefore, necessary for TES' numeracy skills to be more thoroughly examined to gain a clearer view of specific numeracy strengths and weaknesses. Furthermore, it is crucial to explore specific numeracy capabilities to determine common trends of strengths and weaknesses that can be used to inform the design and development of teacher education programmes.

2.5.1 TES' Mathematical Skills

Although Perso (2006) acknowledged that it could not be assumed that someone is numerate simply because they have mathematical knowledge, she also stated that the potential to be numerate could be tested through measuring mathematics knowledge and understanding. Noting this, and the fact that mathematics and numeracy are related (Cockcroft, 1982; Park, 2010; Willis, 1998), it is worth exploring what researchers have found about TES' mathematical skills to understand their numeracy capabilities. This is especially important given the limited research that has evaluated TES' numeracy capabilities. Additionally, it is possible to analyse the results of the studies conducted to examine TES' mathematical knowledge and make connections with appropriate numeracy applications to better assess their numeracy potential.

Studies that have explored TES' general mathematical understandings and misconceptions have found that their mathematical competency is inadequate for teaching (Ball, 1990; Matthews & Ding, 2011). When examining what TES in America (n=252, from five different institutions) understood about mathematics as they entered teacher education programmes, Ball (1990) found that many TES displayed poor mathematical knowledge. Students who were able to provide the correct mathematical responses demonstrated difficulties in doing so. In addition, many expressed that they were worried about specific mathematical content in problems, for example in fractions (Ball, 1990). These results suggested that TES lack the mathematical knowledge and understanding to teach mathematics beyond a superficial level. Accordingly, where TES' mathematical understandings are inadequate for teaching, there is the potential to transfer these mathematical and numeracy deficiencies to students.

Given the lack of studies that have specifically explored TES' numeracy skills, it is worth examining TES' understandings of mathematical concepts, especially since there are many overlaps between mathematical competencies and numeracy capabilities. In 2008, Glidden conducted a study in the US investigating how well TES (n=381) solved four arithmetic problems that required using order of operations. The problems were presented at a basic level (performing multiplication before addition) to determine whether TES understood that multiplication/division and addition/subtraction had the same priority, and their understanding of exponents. Overall, it was found that fewer than half the TES answered more than two out of four questions correctly, and less than 10% of TES answered all four questions correctly. Interestingly, of those who answered correctly, 30% performed addition before subtraction, and 38% performed multiplication before division.

In the US, the acronym PEMDAS stands for parentheses, exponents, multiplication, division, addition and subtraction. This is used as a strategy to help students remember the order of operation; however, it is also necessary to understand that PEMDAS does not strictly dictate the order of the operations. Multiplication and division may be performed in any order from left to right, and addition and subtraction may be performed in any order from left to right. The results from this study indicated that many of the students used the acronym PEMDAS without fully understanding how it should be applied. Many students followed the acronym in its written order; therefore, it was evident that many did not have a strong understanding of the correct application of the order of operation. Overall, the results from this study concluded that TES have a surface understanding of the order of operation (Glidden, 2008). Given the importance and wide-ranging application of the order of operation in mathematical calculations, it stands to reason that TES who struggles to apply this concept will have other mathematical misconceptions and deficiencies that will limit their numeracy competency. For example, number processes are essential skills that can assist in organising (e.g., calculating whether there is enough of something), budgeting (e.g., income and expenditure), and prescribing medication (e.g., calculating or diluting the concentration of medication). TES who have weak mathematical processing skills are likely to struggle to perform computations in everyday numeracy applications.

In regards to the NA mathematical strand, TES' knowledge about fractions has been widely explored and is worth examining further to determine their numeracy application potential. Specifically, Tobias (2013) studied the way elementary TES developed an understanding of language use, Lo and Luo (2012) explored knowledge of fraction division possessed by Taiwanese elementary TES, and Son and Lee (2016) characterised profiles of teachers education students' competence with fraction multiplication. The results from these studies suggested that common fraction knowledge is quite broad. In fact, TES had difficulties with fraction language and, particularly, distinguishing among certain phrases such as *of a*, *of one*, *of the*, and *of each* (Tobias, 2013). Furthermore, TES' multiplication abilities were found to be problematic, whereby many students were unable to solve a given problem (Son & Lee, 2016). Although TES' fraction division skills were found to be adequate, tasks that required students to represent fraction division through word or diagrams were found to be challenging even for the most proficient TES (Lo & Luo, 2012). Overall, the results from these studies suggested that while TES may demonstrate procedural knowledge for fractions, their knowledge and depth of understanding is limited. In fact, the specific skills used for numeracy applications such as language understanding, thought representation, and problem solving were common weaknesses (Lo & Luo, 2012; Son & Lee, 2016). It is important to acknowledge that a thorough understanding of fractions is necessary for numeracy applications such as sharing, conversions, and problem solving. For example, knowledge of fractions is beneficial when following recipes involving fractional amounts. Therefore, if knowledge and understanding of fraction language are weak, numeracy capabilities and the ability to apply these concepts may be limited.

Since numeracy involves solving problems, and algebra is a mathematical skill that can be applied to solve problems, a good command of algebraic skills is essential. Indeed, many studies have explored TES algebra skills; for example, Guler and Celik (2018) conducted a study in Turkey to assess TES' understanding of the five main algebra concepts (algebra expressions, patterns, equations, inequalities, and functions). Overall, the study reported that 65 out of 101 TES performed below the accepted achievement level. Importantly, TES' performance on algebra expressions, equations, and inequalities (which focused on procedural competency) was much better than performance on functions and their properties (which focused on understanding). In agreement with these results, Isik and Kar (2012) found that TES had difficulties in posing problems about different equations. In this study, Isik and Kar (2012) administered a test including five items to 20 TES in Eastern Turkey and conducted semi-structured interviews with those who experienced difficulties in answering the test questions.

The test items included two equations with one unknown and three pair of equations with two unknowns. Overall, there were many difficulties identified, such as incorrect translations of operation and parentheses, unrealistic values given to unknowns, posing separate problems for each equation in an equation pair, and failure to establish relationships between variables. The difficulties identified in this study suggest that TES' knowledge and ability to use and understand equations are limited. Since a thorough understanding of these concepts is necessary for numeracy applications such as understanding formulae, finding unknown quantities, and solving problems, it is likely that numeracy applications of these skills may also be weak.

Weaknesses have also been identified in the MG strand with specific weaknesses reported in content knowledge relating to geometry and geometrical figures. For example, when early childhood TES (n=11) were examined and observed while undertaking their ten-week practicum in Cyprus, Paparistodemou et al. (2014) found they had a lack of geometry understanding in the tasks they developed. TES' attention to geometrical tasks was observed in this study as they planned and delivered lessons involving geometry concepts. Although it was found that the TES designed lessons that involved rich processes, the findings indicated that understanding of simple geometry properties was poor. It was also found that TES were not aware of the significant role of the geometry activities in developing students' knowledge. Aligning with these findings, weaknesses in geometry have also been reported by Yigit (2014). In this exploration, TES' mental constructions of the concepts of angles and angle measurement were examined through interviews. Concerning, Yigit (2014) identified that TES could not identify an angle from a straight line, and they could not identify angles in circles. The TES provided that it was not possible because they needed two lines or a vertex. These findings suggest a weakness in the understanding of geometry properties.

Weaknesses have also been reported in TES' knowledge of properties of geometrical figures. For example, Fujita and Jones (2006) reported that TES have a poor definition command and a weakness in classifying shapes. In this study, data were collected from a survey (n=158), and a task (n=124) to explore TES' understanding of quadrilaterals in accordance with the properties school students are expected to understand. Results from the survey showed that although most TES could correctly draw a parallelogram (97%), a square (98%), a rectangle (98%), and a trapezium (61%), very few were able to answer questions relating to definitions of shapes correctly. For example, not many TES correctly answered whether a square is a trapezium (9%), whether a square is a rectangle (13%), and whether a parallelogram is a trapezium (18%). On the whole, these studies showed that although simple tasks could be developed and construction of shapes were mostly accurate, TES have weaknesses in understanding basic properties relating to geometry and geometrical figures. These findings align with other weaknesses associated with geometrical figures that have been identified. In particular, TES have been reported to display a lack of ability to identify and explain student errors relating to properties of quadrilaterals. For example, in examining TES' ability to provide feedback on their students' work, Şahin and Başgül (2020) found that TES could not explain students' mistakes in a mathematically accurate way. In this case study, TES (n=83) were presented with six scenarios to respond to mistakes made relating to definitions and properties of quadrilaterals. Findings showed that although TES were aware that mistakes were made, they had difficulties identifying the errors. Additionally, very few were able to explain the mistake or provide any solution recommendations. For example, no one was able to provide a recommendation for the elimination of errors in three out of the six scenarios. Correct recommendations for elimination of errors on the other three scenarios ranged from 1.2%-10.8%.

A lack of ability to identify and explain errors has also been reported in other MG content areas, such as symmetry. For example, when exploring identification and explanation of mistakes in reflection symmetry, Hacisalihaglu Karadeniz et al. (2017) reported that TES could only identify errors and could not describe or explain the cause of the errors. In this study, 26 middle school mathematics TES were asked to examine six written solutions containing one or more errors. TES were asked to identify the right and wrong steps the students had made and propose possible explanations for the cause of the error, including ways to correct them. Although the TES generally gave a simple description of the error on most solutions, results showed that the TES' ability to identify errors varied. For example, in the first solution provided, 81% of TES were able to identify the error, and in the second solution, only 31% identified the mistake. However, the most significant finding in this study was that very few TES were able to provide explanations for the causes of errors. In most cases, the explanations provided were very general. This study also found that most TES had little command of mathematical terminologies in their descriptions. On the whole, these studies show that TES' ability to identify errors relating to the MG concepts varied, and when they were able to highlight correct and incorrect answers, they lacked the depth of knowledge to explain the cause of errors.

TES' inability to explain student errors has also been identified in concepts associated with measurement (Livy et al. 2012). In this study, in addition to reporting TES' inability to identify and explain student errors, it was also reported that TES displayed a specific weakness in making connections between different MG concepts. More specifically, it was reported that TES displayed misconceptions in understanding the difference between the perimeter and area concepts. In this exploration, Livy et al. (2012) examined content knowledge of three cohorts of TES from two universities. Participants were required to interpret and respond to primary students' work samples, explain what their responses mean about their thinking, explain what a teacher might do to address any issues, and explain measurement concepts and relationships between perimeter and area. Overall, findings showed that TES displayed similar misconceptions to their students. Other results from this study indicated that TES could not describe what happens to the area of a rectangle when its perimeter is increased. Interestingly, this specific finding has also been reported by other researchers. For example, in the study of Şahin and Başgül (2020) discussed above, it was also identified that TES could not describe or explain the relationship between perimeter and area. Similarly, Holm (2018) also reported on the same weakness. In this study, a survey was administered to 53 TES, and results showed that 85% could not correctly answer whether the area of a rectangle increases when the perimeter increases.

TES' inability to combine information of different MG concepts has also been reported by Ozdemir and Goktepe Yildiz (2015) who concluded that TES could only evaluate independent situations in spatial reasoning skills. In this exploration, a test was administered to 81 TES, and an interview was conducted with 6 TES to examine their spatial orientation skills and classify their abilities according to levels (low-middle-high). Overall, findings showed that TES could not combine their information within a consistent structure (Ozdemir & Goktepe Yildiz, 2015). Furthermore, this study also found that TES only have a surface level (low) understanding and not a deep level (high) understanding of spatial reasoning skills. In summary, these studies show that TES' ability to combine MG concepts and make connections between the concepts is limited, suggesting a basic level of TES' understanding of MG concepts.

Overall, these studies suggest a limitation to TES' capability to develop students' knowledge of MG concepts essential for numeracy applications. For example, geometry knowledge is necessary for understanding bearings, scales, maps, and plans. In addition, reflection symmetry knowledge is required for investigating patterns and designs, and perimeter and area concepts are beneficial for achieving the most efficient use of resources, for example, when packing or stacking items. Since the findings suggest that there are limitations to TES' ability to develop students' knowledge, including a lack of ability to explain the cause of errors, use appropriate mathematical terminology, and make connections between different concepts, numeracy applications of these MG skills may be limited.

In the mathematical strand of SP, specific weaknesses have been identified. For example, when examining statistical reasoning and probability knowledge, Karatoprak et al. (2015) identified misconceptions in TES' understanding. This study collected data from 173 TES in Turkey, including 82 TES training as elementary teachers (grade 5-8) and 91 as secondary teachers (grade 9-12). All participants were given a written test including 20 multiple-choice items examining statistics and probability. Examples of concepts assessed included: interpretation and computation of probabilities, understanding averages, distinguishing between correlation and causation, and interpreting two-way tables. The results from this study indicated that TES from both groups could not select appropriate measures of centre. For example, only 31% of the elementary TES and 34% of the secondary TES chose an appropriate average on the first item assessing this concept. In another item assessing the same skill, only 46% of the elementary TES and 37% of the secondary TES selected an appropriate measure of centre. In addition, many TES from both groups incorrectly selected mode rather than mean as the most appropriate average. Other findings indicated that the TES displayed misconceptions about measures of spread; they could not identify outliers in the data set and could not identify variability in the data when comparing different groups.

In assessing TES' understanding of probability concepts, Karatoprak at al. (2015) found that although TES were able to successfully apply simple probability skills, many were unable to compute probabilities requiring combinatorial reasoning. In the test items assessing these concepts, the percentage of correct responses ranged between 11% and 32%, and specific weaknesses were identified in combinatorial reasoning. In agreement with these results, Forgasz and Hall (2019) also reported a weakness in combinatorial reasoning. In this study, surveys were conducted across three years to examine TES' understanding of mathematical concepts and their confidence to incorporate them into their teaching. Results showed that correct responses on the questions assessing knowledge of combinations ranged from 42%-58%. Further, only 38%-48% expressed confidence in their response. Overall, these results indicated a specific weakness in both knowledge and confidence of combinations concepts.

Furthermore, Karatoprak et al. (2015) also reported that TES lacked reasoning with sample spaces and identified that they could not explain their reasoning of probability outcomes. Although the TES in this study could determine the probability of simple events, a lack of understanding of more difficult probability concepts was evident. Aligning with these findings, Afamasaga-Fuata'i et al. (2007) also reported that TES had difficulties with probability concepts. In this study, two tests were administered to TES (n=46) assessing various concepts. Overall, it was identified that students demonstrated particular difficulties with concepts that involved understanding successive probability concepts.

The weaknesses identified in the SP content areas imply several numeracy weaknesses. For example, the weaknesses displayed in statistics suggest that numeracy applications such as understanding data presented in the news and drawing reasonable conclusions from the data may be quite limited. Furthermore, the weaknesses found in probability also suggest limitations in numeracy capabilities, such as applying knowledge to make appropriate predictions and understand the consequences of events. Competence in the skills of statistics and probability allows for careful consideration of data that can inform decision-making. Therefore overall, there appears to be limitations to TES' numeracy potential involving SP concepts.

In summary, research has found that TES lack capabilities in mathematical areas that span the mathematical strands of NA, MG, and SP. It is essential to acknowledge the weaknesses identified in these areas and consider the implications of TES having a limited ability to apply mathematical skills. A considerable amount of literature exists on specific TES' mathematical weaknesses; therefore, it is reasonable to suggest that TES' potential to be numerate may be quite limited.

2.6 TES' Competency Tests

Teachers' personal mathematical competency and their ability to teach these skills have been documented to impact students' learning (Darling-Hammond, 2000; Hattie, 2009; Shirvani, 2015; Tchoshanov et al., 2017). In addition, it has been identified that to facilitate understanding in mathematics that allows students to make connections, justify, argue their claims and validate their answers, teachers must have an in-depth and thorough understanding of the subject matter themselves (Ball, 1990). Therefore, it is plausible to think that teachers' numeracy understandings should also be comprehensive to facilitate student numeracy understanding. To that end, it is necessary for TES' competencies to be assessed, including numeracy capabilities, to ensure they have the necessary skills and knowledge before becoming teachers.

Standardised testing of TES has existed internationally for many years, and TES have been taking these assessments to meet teacher certification requirements. In 2000, in the UK, the Department for Education acknowledged that teachers must be competent in numeracy and other core skills to fulfill their roles as teachers, and compulsory testing of literacy, numeracy, and information and communications technology (ICT) skills were introduced. Specifically, the Professional Skills Test (Department for Education, 2018a) ensured that all teachers were competent in these skills before being recommended for qualified teacher status. After approximately ten years, the pass mark was set higher in these tests resulting in the pass rate dropping from 98% to 88%. Also, the ICT skills test was removed around this time, and the Professional Skills Test became a pre-requisite for entry into an initial teacher education program rather than a teacher registration requirement (Department for Education, 2011).

The numeracy skills test component of the Professional Skills Test (Department for Education, 2018b) is approximately 48 minutes. It is divided into two areas: a mental arithmetic section that does not permit the use of a calculator and a written section that allows the use of an on-screen calculator. The mental arithmetic section is an audio test with 12 questions that are individually timed. It covers topics such as basic arithmetic, financial mathematics, fractions, decimals, percentages, proportions, measurement, and conversion of units (between different currencies and between fractions, decimals, and percentages) (Department for Education, 2015). The written section involves interpreting and using data and solving arithmetic problems (Department for Education, 2015). Each test is not identical to another, and therefore specific pass marks vary; however, the tests are computer marked, and a printed score report is available once a test is completed. The skills test statistics are available online, and brief analysis shows that more students fail the numeracy test each year than the literacy test; however, the overall pass rates are quite high (Department for Education, 2018c).

Interestingly, as a strategy to boost recruitment into the teaching profession, as of April 2020, new entrants into teacher education in the UK are no longer required to complete the Professional Skills Test (Department for Education, 2020a). Instead, it is now the responsibility of education program providers to ensure that graduates meet the benchmark for fundamental literacy and numeracy skills (Department for Education, 2020b).

In the USA, teacher certification requires passing a standardised test in almost every state. The most commonly administered tests used in more than 40 US states are The Praxis Series exams (Education Testing Service, 2018a). These exams include the Praxis Core Academic Skills for Educator Tests measuring three basic reading, writing, and mathematics skills. The core skills mathematics test is a 1-hour 25-minute test consisting of only selected-response questions. Score reports are provided 10-11 business days after the test, and passing depends on the corresponding state requirement. These pass marks are available online, and the score of 150 is the pass mark required in most states on the mathematics core skills test (Education Testing Service, 2018a).

In other US states, different approaches are used to test teacher competency. For example, in California, the California Basic Educational Skills Test (CBEST) covers reading, writing, and mathematics (Pearson Education Inc., 2018b). The mathematics component of the CBEST assesses three broad areas: estimation, measurement, and statistical principles; computation and problem solving; and numerical and graphic relationships (Commission on Teacher Credentialing, 2017). Oregon and Arizona similarly do not participate in the Praxis but administer the National Evaluation Series, which tests Essential Academic Skills in Reading, Writing, Mathematics, and Technology Literacy (Pearson Education Inc., 2018d).

Many other international approaches have existed for some time to test TES' competency. Interestingly, only some approaches include a measure of mathematics or numeracy component. Examples of other international approaches are in Singapore, where an Entrance Proficiency Test (Ministry of Education Singapore, 2018) is administered, and in Hong Kong, where Language Proficiency Assessment for Teachers of English (The Government of the Hong Kong Special Administrative Region, 2018) is required.

The most recent introduction of an international teacher competency test was in March 2020, when Ontario became the first province in Canada to implement a Mathematics Proficiency Test that TES must pass to achieve teacher certification (Ontario College of Teachers, 2020). The Mathematics Proficiency Test assesses a mathematics content component and a mathematics pedagogy component. To pass the Mathematics Proficiency Test, 70% must be achieved in both the content and pedagogy component. The mathematics content component of the test is based on Year 3-9 level mathematics and makes up 70% of the test. A calculator is not permitted for the first 5 questions that focus on number sense. An on-screen calculator is allowed for the remaining 45 questions covering number sense, relationships and proportional reasoning, and measurements (Education Quality and Accountability Office, 2020). The pedagogy component of the test assesses general understanding of the mathematics curriculum, such as planning, assessment and evaluation practices, and strategies to engage all learners (Education Quality and Accountability Office, 2020).

In Australia, until recent years, the only competency test was for teachers who gained their teacher qualification from non-English speaking countries. The NSW Education Standards Authority (NESA) (2017) outlined that teachers must pass a language proficiency test such as the International English Language Testing System, the International Second Language Proficiency Rating, or the University of NSW Institute of Language-developed Professional English Assessment for Teachers. In more recent years, the requirements to enrol in a teacher education program and the registration to be a teacher in Australia have become more prescriptive. For example, to enrol in an education degree in NSW, prospective students must have achieved a minimum of three Band 5s in the Higher School Certificate. This is the equivalent of achieving 80% or higher in three subjects at the senior secondary level. Also, English must be included in the three Band 5s (NESA, 2020). Regarding teacher registration at the completion of a degree, more specific approaches to assess core skills, including numeracy, have recently been adopted to determine teacher competency.

2.6.1 Literacy and Numeracy Test for Initial Teacher Education

The LANTITE is currently used in Australia to assess initial TES' personal literacy and numeracy skills to meet teaching demands. All TES must sit and pass the computerised test, consisting of a literacy and numeracy component, prior to graduation. ACER (2018c) outlined several purposes of the test, including that it ensures consistency in accrediting graduating teachers. Additionally, the test is intended to ensure that graduate teachers meet the literacy and numeracy standards of approximately the top 30% of the Australian population. This is expected to assist higher education providers, schools, and the public to have confidence in the skills of all teachers who have graduated (ACER, 2018c).

The LANTITE is relatively new in Australia. The NSW Government first introduced in 2013 that TES were required to pass the LANTITE to graduate from their initial teacher qualification. In 2014, TES from NSW tertiary institutions and colleges were involved in an initial trial to assess whether the test content, format, and administration procedures were reliable to provide accurate data (NSW Council of Deans of Education, 2015). Soon after, in August/September 2015, the first test was deployed, which gave TES from all universities and colleges the opportunity to take the test fee-free. After the success of this trial, there was an announcement that from 1 July 2016, all TES across Australia would need to pass the LANTITE prior to graduation. Early in 2016, other trial tests were administered, such as the NSW Literacy and Numeracy for Classroom Readiness test and the ACT Canberra Teacher Recruitment Assessment (NSW Education Standards Authority, 2018). These tests were accepted as approved tests, and students who successfully passed any of these tests are not required to sit the LANTITE.

In order to develop a framework and assessment items for the first two years of the LANTITE's implementation, ACER was contracted by the Australian Institute of Teaching and School Leadership (AITSL). ACER (2018c) outlined that the Australian Core Skills Framework (ACSF) and the Programme for the International Assessment of Adult Competencies were extensively drawn upon developing the concepts to be assessed in the test. More specifically, three content areas were adopted for the development of the numeracy test (NA, MG, and SP), three numeracy processes were adopted (identifying mathematical information and meaning in activities and texts, using and applying mathematical knowledge and problem-solving processes, communicating and representing mathematics), and three context domains were adopted (personal and community, workplace and employment, education and training). Using prescribed targeted proportions of these content areas, numeracy processes, and context domains, the numeracy test was developed to include 65 items to be completed in 2 hours. A calculator is allowed for 52 items, and the use of a calculator is not allowed for 13 items.

It is estimated that thousands of TES sit the test each year across the four available testing windows. However, very limited information on the percentage of TES who pass/fail each test window is released because the NSW Council of Deans of Education and the Government have agreed not to publish results. Instead, a commonly recited statement is that 100% of TES who graduate have passed the LANTITE. Interestingly, the pass rates of each Australian university for each of the literacy and numeracy components of the test were released a few years ago by Campus Morning Mail (2018). This information showed that the pass rate varied widely across the states and universities. It is unknown whether the results released in this article were from one of the four test windows offered in a year or whether the results were the combined results from several test windows for a selected period of time. Nevertheless, the results are quite useful and interesting. Of particular interest, the numeracy results displayed pass rates of individual Australian universities ranging from 50% to 100%, with the average pass rate across the 52 Australian universities being 90.8%. Therefore, based on these results, it is expected that approximately 9% of students who attempt the test are failing the numeracy component. However, it is important to note that TES are allowed three attempts on each LANTITE test (five in some cases, if granted by their institution). Therefore, it would be necessary to fail all attempts for TES to not eligible to graduate from their teacher education programmes.

Further, of the approximate 91% passing, details of the TES' specific strengths and weaknesses are not available. In other words, there are TES that are meeting registration requirements; however, no information about the numeracy skills they may be lacking is provided. At present, an overview is only provided for results in content domains NA, MG, SP, and in Non-Calculator and Calculator (ACER, 2017), with further analysis of the skills within these domains not currently being released. However, a more granule understanding of TES' strengths and weaknesses would be beneficial for both the students and the university providers to understand the skills teachers are taking into the classroom and the skills that need further development.

2.7 Developing Teachers Numeracy

Fortunately, it has recently become a requirement for institutions to include the teaching of numeracy concepts within their teacher education programmes (NESA, 2017). Interestingly, some institutions choose to dedicate specific units to teaching these concepts, while others justify their implementation by embedding these skills across many units within their courses. However, specific expectations of the numeracy skills, knowledge, and teaching strategies should be developed within initial teacher programmes. For example, in NSW, it is outlined that programmes should ensure that graduate teachers understand and analyse the curriculum's numeracy demands. Additionally, it is outlined that they must understand the importance of and role of numeracy in everyday life, understand and analyse the diverse numeracy abilities of learners, and have knowledge of a range of resources that can be drawn upon for student numeracy support and have the ability to effectively use them (NESA, 2017) .

The introduction of the initial teacher program requirements is specifically related to the fact that numeracy is a skill prescribed as a focus area within one of the domains of teaching outlined in the Australian Professional Standards for Teachers (Australian Institute for Teaching and School Leadership, 2011). As part of the Professional Knowledge domain of teaching, one of the standards (Standard Two) describes that a teacher should “know the content and how to teach it” (AITSL, 2011, p. 6). Therefore, the expectation is that initial teacher programmes prepare graduate teachers for the numeracy implementation expected in their teaching.

Considering the onus is now on the university providers to ensure graduate teachers are well-prepared to understand and teach numeracy, it is worth exploring what institutions currently offer to develop these skills. Although all institutions have not published this information, it has been found that many institutions have chosen to dedicate specific units to the teaching of these numeracy concepts and also offer additional numeracy support through specialised services. Table 2.7 summarises some examples of the units of study in selected Australian teacher education programmes (mainly in NSW) dedicated to the teaching and learning numeracy concepts. Additional support services provided by the institutions are also displayed.

Table 2.7.1 Summary of numeracy implementation at institutions

Institution	Dedicated Unit of Study	Dedicated Support Resource	Example of Support Offered
University of Notre Dame Australia	Mathematics and Numeracy	Study Support	Individual support, online practice tests
Western Sydney University		Mathematics Education Support Hub (MESH)	Online resources, help sessions
Australian Catholic University	Literacy and Numeracy Diagnostic		Practice tests
University of Sydney	Literacy and Numeracy Diversity		
University of New South Wales	Language, Literacy and Numeracy	UNSW School of Mathematics and Statistics drop-in Centre	Individual support
University of Wollongong	Mathematics and Numeracy in Education	UOW Moodle	Online support materials
Monash University	Numeracy for Learners and Teachers	Mathematics Self-Help Kiosk	Various resources including quizzes
Curtin University	The Numerate Educator	UniSkills Online Modules (Numeracy Fundamentals)	Online learning resources
Swinburne University	The World of Maths/ Numeracy Across the Curriculum	Support Services	Help from a student advisor

Since the emphasis to develop TES' numeracy is a relatively new concept, there is currently no clarity on best practice. However, some research has recently been conducted to understand the impact of the different approaches that institutions have used (Callingham et al., 2015; Forgasz & Hall, 2019; Sellings et al., 2018). One example is the study of Forgasz and Hall (2019) who investigated TES' numeracy skills, and confidence in their numeracy capabilities, through surveys and interviews conducted before and after completing a newly implemented 'Numeracy for Learners and Teachers' unit at Monash University. The unit was designed for TES to develop an understanding of what numeracy is and how it is related to mathematics, recognise numeracy opportunities across the curriculum, and identify methods to engage students in relevant activities to build numeracy skills. Pre and post-unit surveys were conducted with cohorts in 2015, 2016, and 2017 to examine changes in TES' confidence to implement numeracy into their teaching and gather their feedback on the unit. Overall, the authors reported that TES' understanding and confidence to implement numeracy into their teaching improved after the unit. Nearly all TES (88%) reported that the unit had made an impact on their views of numeracy. However, of those who responded about their overall impression of the unit, 23% did not comment positively. There were 9% who provided a negative comment, and 14% provided a mixed response. Most interestingly, all TES who responded to the overall message they would take away from the unit acknowledged the importance of numeracy for all teachers.

Another example is the study by Callingham et al. (2015) who explored TES' numeracy capabilities through a unit conducted in the first semester of a teaching degree at the University of Tasmania. Numeracy understanding was evaluated through three assessments; a numeracy competency test addressing everyday mathematics, a brief report identifying numeracy in media articles, and an annotated presentation explaining the numeracy of their main teaching area. After the course, some concerning outcomes suggested that not all TES recognised the nature of numeracy and its prospects of its development, and many students stated they would simply get a textbook to progress their numeracy understanding.

Similarly, Sellings et al. (2018) also evaluated an institution initiative developed to improve TES' numeracy skills. The initiative consisted of support interventions offered to TES who did not meet mastery level on an administered test. In total, there were 22% (156 out of 711) of TES who did not meet mastery level and were offered additional support in the way of additional skill sessions, one-to-one support, and individual numeracy plans were provided in some cases. Months later, these students repeated the test with different questions; however, the same skills and difficulty levels were assessed. One concerning finding from the evaluation of this initiative was that 46% of those failing TES (74 out of 156) still did not reach mastery level. Furthermore, some students did not improve their performance at all. For example, eleven students recorded the same mark while nine students performed lower on the subsequent test.

Internationally, there is a gap in the literature exploring university initiatives that have implemented to ensure graduating teachers have adequate numeracy capabilities. Instead, the international focus is on the implementation of teacher competency tests that measure TES' numeracy skills before teacher certification and the relevance of these tests (Angrist & Guryan, 2008; Graham, 2013), TES' attitudes towards numeracy (Stables et al., 2004), in-service teachers' overall numeracy skills (Golsteyn et al., 2016), and in-service teachers' self-efficacy beliefs about numeracy (Golsteyn et al., 2016). With the lack of international research to determine best practice, further research into institution initiatives would be beneficial.

Overall, the research suggests that institutions can play a major role in developing TES' numeracy abilities and their confidence to teach numeracy skills. However, not all current initiatives are successful for all TES (Callingham et al., 2015; Forgasz & Hall, 2019; Sellings et al., 2018). Furthermore, since stand-alone courses are short-term, their longer-term impact on TES' numeracy skills remains unknown. Therefore, it would be beneficial to explore other methods of developing numeracy skills that can also track learning and progress. It is important to track progress to determine whether longer-lasting impacts of numeracy learning are possible. This is especially important considering that teachers are not only required to graduate with adequate numeracy skills, they are also required to sustain their skills so they can be implemented in their classroom teaching for the duration of their teaching careers.

2.8 Diagnostic Tests

Complimentary to the inclusion of numeracy in teacher education programmes, the use of diagnostic tests may be beneficial in evaluating the progress of TES' numeracy skills. It is particularly important to be able to track TES' numeracy learning to determine progress, and this is likely possible through the use of diagnostic tests considering they have been successful in accurately measuring ability in mathematics (Carmody et al., 2006). For example, Carmody et al. (2006) conducted a study to compare diagnostic test results with the end-of-semester unit results of first-year mathematics students at the University of Technology, Sydney. The study included administering a test designed to determine if the students had the required background to complete the unit. The test consisted of 20 multiple-choice questions examining algebra, curve sketching, and calculus and was given to students in the first week of the unit. Findings showed that there was a high correlation between diagnostic test results and students' final exam results. Most students (82%) who scored above 70% on the diagnostic test also performed well on the final exam; however, it was found that the test was a better predictor for those who will likely fail the final exam. Furthermore, Carmody et al. (2006) also collected qualitative data from this study to gauge students' feelings about the diagnostic test. Findings indicated that it gave many students confidence, helped them identify their weaknesses, and provided them with the opportunity to be challenged. Overall, findings from this study concluded that the diagnostic test helped determine ability and alerted students who were not adequately prepared for the mathematics unit of study (Carmody et al., 2006).

Diagnostic tests have also been useful in identifying TES' mathematical knowledge on entry into teacher education programmes. For example, in Ireland, Fitzmaurice et al. (2021) administered a diagnostic test to 365 TES between 1997 and 2013 to examine fundamental knowledge across a range of different content areas, for example, arithmetic, algebra, geometry, trigonometry, and calculus. The test consisted of 40 questions, where using a calculator was not allowed, and TES were provided with the test without warning in the first mathematics lecture of their unit of study. Overall, it was found that the diagnostic test assisted in identifying that TES' mathematics knowledge had declined over the years; however, no specific areas of weakness were determined as there was a decline evident in all areas (Fitzmaurice et al., 2021). This study suggested that diagnostic tests were successful in determining mathematical ability and changes in ability over time.

2.8.1 Online Diagnostic Tests

While pen and paper methods are the more traditional learning style, there has been a recent emergence of computerised learning. In particular, computerised methods have been more beneficial in improving learning than pen and paper methods in mathematics (Alcoholado et al., 2016). For example, to understand the difference in student mathematical learning when using computers or pen and paper to solve arithmetic problems, Alcoholado et al. (2016) compared the results of 81 third graders who were separated into groups. All groups performed the same arithmetic exercise; however, the main difference between the groups using computers and the group using pen and paper was in the immediacy of the feedback given. The groups using computers were provided with instant feedback and could not advance to the following questions until they answered the current one correctly. Alternatively, the pen and paper group had to complete all exercises before using an answer booklet to discover mistakes. Learning was measured through pre and post-tests. Although all groups were found to progress in their learning, findings showed that the results significantly favoured the groups using computers instead of the group using pen and paper. An explanation provided was due to the instant feedback that the groups using computers were provided.

Similarly, although pen and paper tests are the more traditional style of testing instruments, computerised online tests are increasingly being used to assess student performance and track progress (Snekalatha et al., 2021). For educators, although there is a higher workload associated with the design and development of online tests, a reduction in work time is possible when managing the tests. For example, collecting multiple test papers is not necessary, and automatic marking functions can replace hours of marking time. For students, there are also many benefits, such as the ability to use a mobile phone to take the tests, the significant assistance they provide in learning the subject, and the immediate feedback they receive (Snekalatha et al., 2021).

It is also important to acknowledge that online tests have shown comparable results to traditional pen and paper-based exams and therefore are a reliable measure of students' knowledge. For example, Blanco et al. (2009) have shown the reliability of online tests, reporting that online test results show a positive correlation with results from subsequent paper-based tests. Similarly, Metz (2008) also reported that student results from online tests displayed a statistically significant and positive correlation with overall course achievement. In Metz's (2008) study, weekly online tests were given to 215 undergraduate students in the US as part of a biology unit. The online tests represented 12% of the overall course grade, and other course components represented the remaining percentage. At the completion of the unit, comparisons of students' average online test scores and their overall results at the end of the course were made. Overall, it was found there was close to a 1:1 correspondence between online test results and overall course performance, suggesting that the online tests were a reliable indicator of students' knowledge.

In addition to being a reliable measure of knowledge, online assessments have also been a more valuable form of assessment than traditional pen and paper tests for TES (Linsell & Anakin, 2012). To determine whether their two forms of numeracy diagnostic instruments (written and online) were reliable assessments of number and algebra content knowledge, Linsell and Anakin (2012) analysed and compared results from the two tests. The written test was given to 153 TES in 2010, and the online test was given to 122 TES in 2011. To determine how reliable the assessments were, sets of questions that were considered indicators of foundational knowledge were selected from each assessment. Performance on these indicator questions was then compared to the overall test results to determine how reliable the tests were in determining ability. Overall, an interesting finding from this study revealed that it was possible to pass the written assessment without foundational content knowledge. Additionally, Linsell and Anakin (2012) noted that their findings raised questions about how valuable the written assessment was as a diagnostic assessment of TES compared to the online version. Therefore, the online test was considered the more reliable indicator of knowledge.

It must be acknowledged that online tests can be problematic. One example is the increased cheating behaviour that has occurred when online tests are unsupervised (Fask et al., 2015). However, other studies have explored the validity of online tests, specifically when students are unsupervised, and have found that lack of supervision is not an issue (Ladyshewsky, 2015; Metz, 2008). For example, while exploring the differences in post-graduate students who completed a supervised multiple-choice in-class paper and pencil test and those who completed an unsupervised multiple-choice online test, Ladyshewsky (2015) found that the mean test scores were not significantly different. Similarly, in the study of Metz (2008) discussed above, it was found that academic dishonestly was not a determinant in the unsupervised weekly online test performance. This conclusion was drawn because scores were seen to decrease over the test access time. That is, most students who sat the test earlier performed better than those who sat the test later. Therefore, even though it was predicted to happen, sharing information from students who had completed the test earlier did not seem to occur (Metz, 2008). While these findings do not suggest that caution is unnecessary when implementing online assessment, they suggest that academic dishonesty is not a great concern for online tests.

It is also important to consider other factors affecting the validity of online diagnostic tests used to assess mathematics and numeracy. For example, online diagnostic tests such as PISA, TIMSS, NAPLAN typically include a range of item types such as multiple-choice, true/false, and short response questions. Given the range of test features available and used to design online tests, it is worth exploring whether test performance could be affected by the different item types that exist in a test. One such example was conducted by Abida et al. (2011), who explored whether it is necessary to have multiple-choice items only, short response items only, or a combination of both item types for maths proficiency assessments. In their study, a large-scale mathematics proficiency test that included questions with both item types was administered to 268 grade 9 students from 134 schools in Pakistan. Their study found that the mean scores for multiple-choice items were lower than the short response items and, therefore, lower than the overall maths proficient test mean score. The authors' explanation was that the multiple-choice items may have been more difficult or that students may have been more familiar with the short response items (Abida et al., 2011).

Several studies exist that specifically explore item type performance in mathematics assessments (Lindberg et al., 2010; Lui & Wilson, 2009; Taylor & Lee, 2012) and other tests relevant to numeracy, for example, economics tests (Becker et al., 1990; Walstad & Robson, 1997). Interestingly, these studies all focus on and identify gender differences in item type performance. More specifically, the research suggests that multiple-choice items favour males and short response items favour females. The differences are so obvious that Siegfried and Wuttke (2019) suggested that females may be disadvantaged in the overall result when there is a large number of multiple-choice items because they are more likely to score lower on these items simply due to their gender. They also suggested that due to the test format, these gender-based differences may potentially cause misjudgements on ability. With this in mind, it is therefore important to consider the test item formats of assessments when making a decision about results and abilities, particularly when gender differences are involved.

Although gender gaps are not a particular focus in this study, it is still worth exploring aspects other than content knowledge that may affect the validity of online tests and overall numeracy test performance. Furthermore, despite the literature acknowledging females' lower performance on multiple-choice items, research suggests that females are better able to sustain their performance during tests due to the test-taking strategies they may adopt and therefore decrease this gender gap (Balart & Oosterveen, 2019). Reasons cited by Balart and Oosterveen (2019) are that females have more self-discipline and have an advantage in cognitive planning abilities. Therefore, it is justified to suggest that longer tests may decrease this gender gap and should be considered when determining the validity of online tests.

Although diagnostic tests are commonly used to assess performance and to inform education interventions (Siegfried & Wuttke, 2019), it has been acknowledged that assessments should not merely be used as tools for measurement. Alternatively, self-testing has recently emerged as a validated teaching and learning tool that scholars believe should be used to improve performance (Wiggins, 1993). Furthermore, Wiggins (1993) identified that timely feedback is an important component of ensuring the learning process takes place through assessment. More specifically, the time taken for students to receive feedback has been found to considerably affect student learning outcomes. For example, when Alcoholado et al. (2016) aimed to understand the difference between learning outcomes when solving arithmetic problems comparing the use of computers and pen and paper, the results significantly favoured the groups using computers because they were provided with instant feedback from their responses as opposed to the students with pen and paper whose feedback was delayed.

Since developing personal numeracy skills is a long-term endeavour, the literature suggests that online diagnostic tests could support this development. Considering the benefits of online learning (Alcoholado et al., 2016), the accuracy online diagnostic tests provide in diagnosing skills (Linsell & Anakin, 2012) and also the ability they have to be able to track learning and progress (Snekalatha et al., 2021), the development of numeracy skills could be achieved through the use of such tests. Therefore, through the use of an online diagnostic test that encourages self-testing and provides instant effective feedback, TES' numeracy skills could be accurately evaluated and improved to ensure they are graduating with sufficient skills. Furthermore, there may be longer-lasting impacts of learning through the use of online diagnostic tests that will allow TES to retain their skills to take into their classroom teaching. This is especially important to ensure they can confidently implement numeracy into their teaching for the duration of their teaching careers.

Chapter Three: Theoretical Framework

3.1 Introduction

Two main elements of this research have been considered when adopting an appropriate theoretical framework. They are the development of TES' numeracy skills in an online learning environment and the development of numeracy learning through an online test. Therefore, an appropriate framework for this research will need to encapsulate the development of TES' numeracy skills as well as self-paced online learning through assessment. Since there is currently no theoretical framework that addresses both of these areas, theories that explore online learning and learning through assessment will be examined separately, to discern overlaps that can be used for this research study.

Research suggests that several elements inform preparation and teaching. For example, Shulman's (1987) Model of Pedagogical Reasoning indicated that in order to inform preparation and teaching, teachers draw on sources of knowledge from several categories. These categories include content knowledge, curricular knowledge, pedagogical knowledge, knowledge of aims and purposes, knowledge of learners, and knowledge of educational contexts, settings, and governance (Gudmundsdottir & Shulman, 1987). Specifically, Shulman's (1987) model described how a teacher's understanding of content is transformed to make it teachable. Within the literature and the Australian Professional Standards for Teachers (AITSL, 2011), content knowledge is the first category necessary for successful teaching. Therefore, key to this research is the focus on developing content knowledge to inform successful teaching.

The primary focus of this research is placed on developing TES' numeracy content knowledge through the use of an online test. Therefore, this chapter will initially provide a systemic review of considered theoretical frameworks in the study of online learning. Following this, the chapter will review the role and types of assessments and their influence on student learning. More specifically, this chapter discusses the roles of assessment, types of assessments, developments of A*f*L, and Black and Wiliam's (1998) development of the A*f*L theory. Next, the chapter will argue for why and how the A*f*L theory was adopted for the development of an online diagnostic test. Finally, the chapter concludes with a discussion of the overlaps identified between the online learning models and the A*f*L model to be adopted in the design of the testing instrument and also provides a discussion of Tan’s (2013) extension to the A*f*L model that is used for the interpretation of findings in this thesis.

3.2 Theoretical Frameworks and Models in the Study of Online Learning

Online learning links learning with technology. More specifically, online learning is the achievement of knowledge through technology as the enabler of learning (Aparicio et al., 2016). The research described in this thesis involves the evaluation and improvement of TES’ numeracy skills through the use of an online test. Therefore, there is an element of learning through the use of technology involved in this study. As such, models and theories used in the study of online learning were evaluated to determine if they were appropriate for this research. Within the research literature, the Technology Acceptance Model (TAM), the Analysis, Design, Development, Implementation, and Evaluation (ADDIE) Model, and the e-learning systems framework of Aparicio et al. (2016) are the most frequently cited.

Several online learning studies have adopted the TAM (Ranellucci et al., 2020; Song et al., 2017); therefore, it was considered a possible theoretical framework for this research. The TAM considers that an individual’s perception of how easy to use and how useful the technology is, determines their attitude towards the use of technology and behavioural intention to use it (Ranellucci et al., 2020). Hence, determining whether they accept or reject the technology. To explore the most widely used external factors of the TAM regarding online-learning acceptance, Salloum et al. (2019) conducted a literature review. The literature analysis showed that computer self-efficacy, subjective norm, perceived enjoyment, system quality, information quality, content quality, accessibility, and computer playfulness were the most common TAM factors.

Most studies that adopted the TAM explored technology user adoption behaviours and predicted behavioural intentions and use of technologies (Ranellucci et al., 2020; Song et al., 2017). As reported by Salloum et al. (2019), who examined university students’ acceptance of online learning by adopting the TAM, found that the system quality, computer self-efficacy, and computer playfulness significantly impacted online learning systems' perceived ease of use. Similarly, from examining TES’ attitudes toward computers, Teo et al. (2008) found that perceived usefulness, perceived ease of use, and subjective norm were the most significant determinants of TES' computer attitudes. Furthermore, from examining environmental factors that might play a role in influencing a person's desire to perform a task (e.g., technical support, the physical environment, lack of access), Teo et al. (2008) found that these facilitating conditions affected the perceived ease of use but did not influence TES' overall attitude towards the technology.

Overall, the TAM is considered a good model to evaluate technology adoption but does not determine users' level of learning through interacting with the technology, which is what the research described in this thesis seeks to determine. Therefore, the element of technology adoption and the variables considered successful were considered through use of the TAM in this study; however, other models were explored to determine users' learning.

Another theoretical framework considered for this study was the ADDIE Model. The ADDIE Model addresses the shortcomings of the TAM and, more specifically, evaluates the technology program's effectiveness (Trust & Pektas, 2018). One main goal of using the ADDIE Model is to promote online students' focus on learning and active teaching and learning (Castro & Tumibay, 2019). Scholars suggest that the model is most commonly used as an instructional design model used to guide processes that instructional designers use (Castro & Tumibay, 2019). The first stage is the Analysis phase, where designers identify the goals to be achieved and know the intended user, the learning environment, and the materials that need to be taught. The second stage is the Design phase, where the designer carefully constructs a task analysis that includes steps the learner must take. The third phase is Development, where the performance objectives are written, and assessments are created to provide feedback about the learner's performance. The fourth phase is Implementation, where the overall plan is actioned. The final phase is Evaluation which consists of formative and summative evaluation. The formative evaluation plays a role in each stage, while summative evaluation is used for feedback to improve the program. Although the ADDIE Model is typically represented as a linear model, the five components are considered interconnecting (Trust & Pektas, 2018).

In 2016, The ADDIE Model was used by Nichols Hess and Greer to incorporate best practice in teaching and learning into an online university course. As a result, the authors found that the ADDIE Model could be used to achieve several elements of instruction . For example, they found that it provided structure that allowed for the development of a variety of interactions, promoted more consideration of student engagement, learning, and assessment, and assisted in linking content standards with other learning guidelines such as high-impact practices and e-learning best practices. Similarly, Trust and Pektas (2018) used the ADDIE Model to guide the development of an open online course to meet the learners' diverse needs. Analysis of post-course surveys suggested that most students who completed the course could achieve the intended learning objectives. Overall, the ADDIE Model is a successful model used by instructional designers to encourage and explore users' learning. It is a systemic design process that is believed to lead to instruction that is effective, efficient, and significant (Gustafson & Branch, 2002).

More generally, online learning theories can be broken down into three core elements. According to Dabbagh (2005), these elements can be defined through a theory-based framework that links learning technologies, instructional strategies, and pedagogical models. Furthermore, Dabbagh's (2005) framework considers how people learn, together with the learning strategy and the technology. Dabbagh's (2005) framework and the work of other scholars, such as Mason and Rennie's (2006) classification of online learning perspectives, were reviewed and presented by Aparicio et al. (2016) to provide the theoretical background for online learning research strategies. Subsequently, the e-learning systems theoretical framework was constructed. More specifically, the e-learning systems framework of Aparicio et al. (2016) was developed with a more holistic view and summarises the three dimensions of online learning. The framework aims to identify the intended participants and users, whether the technology used is intended to provide content, communication, or collaboration, and whether the services related to online learning include pedagogical models or instructional strategies.

For the research described in this thesis, elements of the discussed online learning theoretical frameworks models all contributed to guide the study (Figure 3.2.1). Firstly, the TAM was adopted through consideration of the participants' perception of how easy to use and how useful the technology is. The users were carefully considered in the test design to promote a positive attitude towards the use of the online test. Secondly, considering the ADDIE Model describes best practice in learning design (Nichols Hess & Greer, 2016), this model was specifically considered in the development of the online test used to collect the data for this research. The ADDIE Model phases of Analysis, Design, Development, Implementation, and Evaluation were specifically adopted in the design and development of the online test to ensure best practice in learning design. Finally, the research design also considered the e-learning systems theoretical framework elements of: users, technology, and services. The specific intention of the use of the technology in this research was carefully considered through these three dimensions. That is, the technology was intended to be used by a specific TES sample, aimed to provide content to the users, and include services that included some instructional strategies. However, in addition to online learning, the research described in this thesis has a focus on the specific role assessment plays in enhancing learning. Therefore, it was important to also explore the concept of assessment.

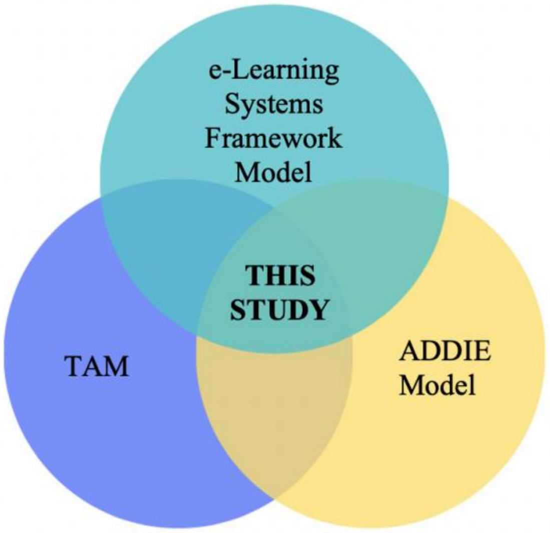


Figure 3.2.1 Interaction of the online learning theoretical frameworks in this study.

3.3 Roles of Assessments

Assessment is a central component of education and is used for a variety of reasons. For example, to gather, interpret, and use evidence to make judgements about one's achievement (Harlen, 2007). The specific role of assessment refers to the function and intended purpose of an assessment in which evaluation generally plays a role. The term evaluation was previously considered a more preferable term as it was thought to portray less negativity (Taras, 2005). In education, there are two common forms of assessments, each having distinctive roles. One form of assessment is known as summative assessment which is concerned with measuring the degree to which students have achieved curricular objectives and captures evidence of learning up to a given point (Taras, 2005; Yorke, 2003). It has been suggested that summative assessments represent all the negative aspects that the more general term of assessment used to hold (Taras, 2005). This is because summative assessments generally take place at a fixed time and involve assessing learning that often leads to a final mark or grade and are often associated with high-stakes examinations (Allan, 2016; Harlen, 2007).

There are many types of summative assessments. For example, examinations, essays, presentations, group activities, and projects (Allan, 2016). However, in a study by Taras (2008) that surveyed 50 lecturers to explore clarity in the definitions and relationship between different types of assessments, findings showed that educators were only able to provide examples of summative tasks that included external examinations and assignments. Furthermore, the lecturers generally used the terms 'final' and 'end' to define summative assessment. This highlights the way that summative assessments are generally understood and used by educators.

The other form of assessment is known as formative assessment. The purpose of formative assessment is to contribute to student learning through the provision of information about performance (Yorke, 2003). It occurs during the learning process, is intended to inform students of their progress, and provide an indication of how work can be improved (Allan, 2016; Taras, 2005). Boud (2000) described formative assessment as the process that guides us in how to learn what we want to learn that then tells us how well we are progressing to get there. Examples of formative assessments include quizzes, peer assessments, and essay drafts. Although there are many opportunities to implement formative assessments in the classroom, the purpose and how formative assessments should be implemented is sometimes misunderstood. For example, it has been reported that many lecturers provide examples of tasks that were not even related to any characteristics which make a task formative (Taras, 2008). Taras' (2008) study also showed that many lecturers lacked understanding of the definition of formative assessment, and only 28% mentioned that feedback was central to the definition. This is in contrast to the research literature, which often associates formative assessment with feedback. For example, the definition of formative assessment widely used in the literature highlighted that effective feedback is a major determinant that can enable students to develop evaluative skills to improve (Sadler, 1989). These results suggest that educators are not entirely clear about the role of formative assessment tasks in supporting student learning. Indeed, this is supported by Wiliam’s (2011) argument that teachers may not entirely understand and practice formative assessments.

3.4 Types of Assessments

Another way to view the different types of assessments is through the educator’s perspective, which is focused on the role of assessment as a tool to promote learning. In this light, assessments can be considered as: Assessment *of* Learning (A*o*L), Assessment *as* Learning (A*a*L), and A*f*L. Firstly, A*o*L is a summative assessment regarded as an essential part of education and necessary to record and report on learning outcomes (Harlen, 2007). A*o*L takes place after the learning has occurred to determine the extent of the learning (NSW Government, 2021), and can be used to assess achievement against specific standards or outcomes (NESA, 2021). NESA (2021) outlined that A*o*L provides evidence of achievement that can be used to plan future goals for learning.

The emphasis is shifted from summative to formative assessment in the other two types of assessment: A*a*L and A*f*L. These two formative assessments are to be used complementary to each other and are often described together because they both share the characteristic of assessing what students know while also supporting learning (Berry, 2008). Furthermore, they both treat learning as an internal endeavour and emphasise the importance of feedback (Berry, 2008). However, there are some differences between the two. A*a*L considers how students evaluate their own learning, including how they use the feedback provided (NESA, 2021). A*a*L recognises that students must engage in interactions and make judgements that allow them to advance their learning (Dann, 2014). In A*a*L, students are their own assessors, and they choose from and use a range of strategies to determine what they know and can do (NESA, 2021).

On the other hand, A*f*L are designed to specifically promote students’ learning and are used throughout the learning process to clarify students' understanding (NESA, 2021; Wiliam, 2011). In fact, the term A*f*L is often used interchangeably with the term formative assessment (NESA, 2021). Interestingly, Wiliam (2011) noted that many scholars have stopped using the term formative assessment and now only use the term A*f*L. Some characteristics of A*f*L that have been highlighted by NESA (2021) including clear goals, providing effective feedback, reflecting a belief that improvements are possible, encouraging self-assessment, and are inclusive of all learners. According to Berry and Kennedy (2008), A*f*L allows students to make the decisions that matter most by allowing them to gain continuous information about their learning, information that describes where they are succeeding, where they should focus efforts for improvements, and what strategies they need to consider to improve performance. Similarly, Clark (2012) identified that the purpose of A*f*L is to monitor learners' progress towards desired goals, which seeks to close the gap between the learner's current and desired levels. In summary, A*f*L provides a way for students to know what can be done in response to a task.

However, it is important to acknowledge that divisions can occur when assessment is viewed from a teacher-dominated view of teaching and learning compared to a student-centred view. If assessment is viewed from a student-centred perspective, then improvements can only occur when the student undertakes an assessment process that includes accepting and using teacher feedback. The general purpose of assessment is to provide data or evidence of student learning, and teacher feedback provides data for students to use and act upon. The research described in this thesis considered the different views of assessment and considers how A*f*L can be used to assess TES' numeracy knowledge while also supporting their numeracy learning and development.

3.5 Assessment for Learning

The idea of using assessments to enhance learning would likely have existed for a long time; however, the term A*f*L is a relatively recent one. Since its introduction to the educational community, which is often credited to Gipps (1994), many developments have been made to improve the quality of assessments to enhance learning. For example, contributions to the development of A*f*L have been made by introducing the idea that formative assessment is effective when benefits are long-term and allow for the preparation of students for sustainable learning (Boud, 2000). More specifically, Boud (2000) outlined the idea that assessment has a double function of meeting specific and immediate goals and establishing a ground for students to perform self-assessments in the future. Boud (2000) also asserts that if a sustainable assessment is not part of education at all levels, students will not be able to fully participate in a learning society.

Contributions to the development of A*f*L have also been made through the identification of the need for self-regulated learning to be encouraged through formative assessment (Clark, 2012). In Clark's (2012) study, 199 publications were examined on assessment, learning, and motivation to present a detailed view of formative assessment's values, theories, and goals. In particular, the study noted the relationship between feedback in formative assessment and how it develops and encourages self-regulated learning strategies. Another finding in the study by Clark (2012) was that the development of self-regulated learning strategies supported sustained motivation for life-long learning. While many other developments of A*f*L have been made, it is the contributions of Black and Wiliam (1998) to promote improvements in the quality of formative assessments that have been particularly influential in the field.

3.6 Black and Wiliam's Theory on Assessment for Learning

In 1998, Black and Wiliam published a seminal article which shifted the view on formative assessments from a more general view to one that focuses on the interactions between learning and assessment, thus providing a clear view of what A*f*L entails. More specifically, Black and Wiliam (1998) conducted a comprehensive review of formative assessment in the research literature to explore the perceptions of students, their role in self-assessment, and strategies incorporated in order to provide a detailed and theoretical perspective of A*f*L.

To serve as the basis for their review, Black and Wiliam (1998) examined articles they considered to be influential work in the field of A*f*L, which were written by Natriello (1987) and Crooks (1988). These two papers were the centrepiece of Black and Wiliam's search. Firstly, because their work was in the same field, and secondly, because of the difference between the reviews. That is, Natriello's (1987) review covered a full range of assessment purposes, while Crooks' (1988) review had a narrower focus, which focused on the evaluation practices of students. The gathering of further articles was produced through identifying publications that cited these two articles, using keyword searches in a research database, and examining the reference lists of the articles found. This process produced 681 publications in total and 250 of those were considered important to read in full. From these publications, the key elements coded, reviewed, and discussed. These elements include self-assessment, response and reception, goal orientation, self-perception, the use of tests, and the quality of feedback (Black & Wiliam, 1998). These elements are discussed below.

3.6.1 Self-Assessment

Several factors relating to self-assessment and self-evaluation were discussed in Black and Wiliam's (1998) review. This included the benefits of self-assessment methods that were reported by Fontana and Fernandes (1994). In Black and Wiliam's (1998) evaluation of these studies, they noted that there were significant gains in groups of students who used daily self-assessment methods over those who did not. Further, Black and Wiliam (1998) explored self-evaluation in the research by Schunk (1996), who compared measures in skill, motivation, and self-efficacy between students who performed self-evaluations and students who did not. It was found that the effect of frequent self-evaluation produced higher motivation and achievement outcomes. Additionally, from reviewing Carroll's (1994) study on the development of self-assessment capabilities, Black and Wiliam (1998) noted that when worked examples of algebra problems were provided for students to study, it improved their overall performance. The most significant improvements were seen in low achievers.

3.6.2 Response and Reception

Another aspect Black and Wiliam (1998) discuss in their review, is response and reception. It was stated that there are complex links between the way an assessment message is received, the way it is perceived, and the way it motivates a course of action (Black & Wiliam, 1998). Therefore, a focus in the review considered the factors which influenced the way that the assessment messages are received and the different ways in which positive action is taken. One example provided was that students are often reluctant to seek assistance because they interpret seeking help as confirmation of them having a low ability (Blumenfeld, 1992). Further, research conducted by Newman and Schwager (1995) on the effects of guidance on improving students' mathematical problems solving ability, showed that although different forms of guidance made a difference, the frequency of students seeking assistance was low. These findings led Newman and Schwager (1995) to conclude that there was a need to encourage more help-seeking by students.

3.6.3 Goal Orientation

In addition, Black and Wiliam's (1998) reviewed goal orientation as an area that had been widely explored. For this, Black and Wiliam (1998) drew on the research conducted by Ames and Archer (1988) who enquired into the goals of 176 students. In this study, it was reported that students' goals could be divided into two distinct groups. One group spoke of the importance of learning, believed in the value of effort, and were generally positive about learning. The other group displayed more negative attitudes, was associated with lack of ability and under-achievement, and focused on the importance of out-performing others. Furthermore, from the research of Cameron and Pierce (1994) and Kluger and DeNisi (1996), it was noted that feedback that focuses on building self-esteem, including giving praise, can have a negative effect on performance and also attitudes. Black and Wiliam (1998) noted that a similar conclusion was also made in the research by Lepper and Hodell (1989), who argued that interest and motivation can be undermined with reward systems.

3.6.4 Self-Perception

Self-perception was also explored in Black and Wiliam's (1998) review; however, it is noted that this section was selective and did not cover all aspects of attitude and motivation. Nevertheless, it was acknowledged that many studies have found that students' achievement can be affected by their beliefs about their own capacity as learners. For example, research has suggested that achievement is linked to students' sense of their own control over their learning (Fernandes & Fontana, 1996) and self-directed learning styles produced better learning outcomes (Grolnick & Ryan, 1987). It was also noted by Black and Wiliam (1998) that personal features have effects on learning. These personal features can be either good or bad depending on the way that formative feedback is provided to the student and the context of culture and beliefs about ability and effort within which feedback is interpreted (Black & Wiliam, 1998).

3.6.5 Frequency of Assessments

Given the nature of A*f*L activities, the impact of regular assessment was also reviewed. Overall, frequent testing can lead to improved learning, which was highlighted in a meta-analysis of 40 studies conducted by Bangert-Drowns et al. (1991a). In this study, it was also reported that performance positively correlated with assessment frequency. In other words, performance improved as the frequency of assessments increased. Importantly, in studies where more frequent assessments did not improve performance, students indicated that they preferred having frequent assessments. As such, it could be argued that frequent testing can have benefits such as improving performance and/or increasing motivation to learn (Black & Wiliam, 1998). It should also be considered that results from these assessments are providing feedback to students on their learning progress.

3.6.6 Feedback

A major aspect of Black and Wiliam's (1998) review centred on the notion that feedback is the critical feature in formative assessment. Firstly, Black and Wiliam (1998) identified four elements required to make up the feedback system: the information on the actual level of a measurable attribute, information on the reference level of that attribute, a mechanism for comparing the two levels, and a mechanism by which the data can be used to alter the gap. Interestingly, Kluger and DeNisi (1996) only regarded the first element necessary for feedback to exist. Also, Ramaprasad (1983) defined feedback as the information about the gap between the actual level and reference level and suggested that for feedback to exist, the information about the gap must be used to alter of inform how to alter the gap. However, Black and Wiliam (1998) took the broader view when discussing effective feedback and considered that the four listed elements are all necessary for A*f*L.

Black and Wiliam (1998) noted that one of the most important reviews of the effectiveness of feedback was conducted by Kluger and DeNisi (1996). In particular, Kluger and DeNisi (1996) conducted an extensive review of 131 reports of the effects of feedback on performance. They found, on average, the effect size of 0.4, which they report was equivalent to raising the achievement of the average student to the 65th percentile. However, due to some negative feedback effects, they explored the variability in the reported effect sizes and the specific interventions of the feedback provided were found to be of significant importance. Examples of impacting factors were that the individual must have a high commitment to achieving the reference level and must believe in their own success to achieve the reference level. Most importantly, it was also noted that the reference level goal must be made clear to the individual (Kluger & DeNisi, 1996).

The idea of providing feedback through A*f*L tasks to highlight the reference level goal has more recently been discussed and extended upon. For example, Taras (2005) suggested that for an assessment to be formative, feedback is required that explicitly identifies the required standard and indicate a gap between the level of current achievement and the required standard. Later, Hattie and Timperley (2007) developed a model to more clearly highlight that feedback must answer the students' questions of *where am I going*, *how am I going*, and *where to next*. More recently, Allan (2016) extended upon these three questions and outlined that the feedback provided to students through formative assessment tasks should specifically address how well they are doing, their strengths and weaknesses, whether they are progressing, what areas they need to work on, and how they compare to others.

Exploration into the quality of the feedback was also conducted in Black and Wiliam's (1998) review to determine other requirements for feedback to be effective. In particular, Black and Wiliam (1998) noted the study by Bangert-Drowns et al. (1991b), who conducted a meta-analysis of 58 experiments from 40 reports. This study reported that for feedback to be of high quality, it must encourage a student's reaction as a response to the feedback. Additionally, the study reported that the quality of the feedback had a considerable influence on student learning and was most effective when it was designed to encourage the correction of errors (Bangert-Drowns et al., 1991b). Importantly, it was found that feedback was best when it gave details of the correct answer instead of simply indicating whether the student's response was correct or incorrect.

It is now widely accepted that feedback is a major determinant of effective formative assessment and it has been broadly discussed in the literature since Black and Wiliam's (1998) review. For example, in an exploration of formative assessment in higher education, Yorke (2003) emphasised the importance of feedback and outlined that, from a student's perspective, feedback is only effective if it contributes to learning. Similarly, Sekulich (2020) conducted an extensive literature review and highlighted that formative feedback processes guides students to learn and motivates them to close achievement gaps.

Another example is the study of Havnes et al. (2012), which conducted surveys and interviews with teachers and students across different subject areas in five schools in Norway. The intention was to explore concepts of tests and assessments, feedback practices, and how teachers and students perceive the practices. Results showed that most students found the feedback they received on tests and assignments helpful as it provided them with information about how well they performed and what was expected of them. Although this study found that the participating schools had not yet developed a culture of A*f*L, the results highlighted the overall benefits of feedback in assessment practices.

Overall, Black and Wiliam's (1998) theory of A*f*L is a comprehensive and highly regarded frame in the research literature. According to Yorke (2003), Black and Wiliam's review confirms a theory that formative assessment is essential for student learning. Furthermore, Taras (2010) outlined that Black and Wiliam's (1998) A*f*L theory motivates and inspires teachers and academics worldwide.

3.7 Use of Assessment *for* Learning in this Research

An appropriate framework for this research needed to encapsulate the development of TES' numeracy skills as well as self-paced online learning through assessment. Therefore, overlaps were identified between the models explored to embed in the research. As such, the TAM, the ADDIE Model, the e-learning systems theoretical framework, and Black and Wiliam's (1998) theory of A*f*L were all evaluated to discern overlaps. Interestingly, overlaps were apparent between the online learning theories and Black and Wiliam's (1998) theory of A*f*L. In particular, a common element was identified to be the element of the users *and* participants. The online learning models outline that consideration must be given to the participants' perception of how easy to use and how useful the technology is. This is necessary to promote a positive attitude towards the technology. Similarly, Black and Wiliam's (1998) theory includes elements of self-assessment, response and reception, goal orientation, and self-perception. All of these elements are also user/participant focused.

Another common element identified between the online learning models and Black and Wiliam's (1998) theory of A*f*L was the provision of feedback about the learner's performance. In summary, the online learning models highlight that the development of an online design need to include services that intend to involve some instructional strategies, such as providing feedback. This links with Black and Wiliam's (1998) theory which places a strong emphasis on providing feedback to the learner for A*f*L to occur. These overlapping elements were considered necessary to embed in the research.

Black and Wiliam (1998) described the practices that must exist for A*f*L to occur, which informs the theory. These practices, considered the determinants of effective formative assessment, together with the online learning theoretical framework elements, are specifically utilised in the research design that included the development of a user-friendly online diagnostic test used to evaluate TES numeracy skills. In order to promote learning, the design of the test considered the intended users and participants and, in particular, the determinants of self-assessment and self-evaluation, response and reception, goal orientation, self-perception, and the use of tests. Of most significance, the element of feedback is also considered, and a major emphasis is placed on providing effective feedback through the online testing system. The benefits of providing feedback through technology are supported by research in higher education. For example, Winstone and Carless (2020) explored how technology can be used as an approach to providing feedback and found that capabilities for students to engage with feedback to promote learning and development could be enhanced by technology. The specific test implementation of feedback and also the consideration of the other elements are further discussed in Chapter Four.

Recently, an extension of the original A*f*L framework was proposed by Tan (2013). Specifically, Tan (2013) proposed a framework for A*f*L that emphasised assessment design, student feedback practices, and clarity of standards as the three main components of A*f*L that are able to influence each other. Tan (2013) believed that these three components must work together, be understood together, and when triangulated together provide the elements for assessment to be used to optimise students' learning.

As such, Tan (2013) developed the triangulated model (Figure 3.6), where the component of standards depicts the vertical axis and displays the identified gaps that the assessment intends to close for each student. The task design element is represented by the horizontal axis which displays the pace of learning from when the gaps are first exposed to when learning is assessed in order to determine if improvements had been made. The feedback component represents the trajectory of the triangulated model and the level of incline reveals how ambitious the feedback is (Tan, 2013).

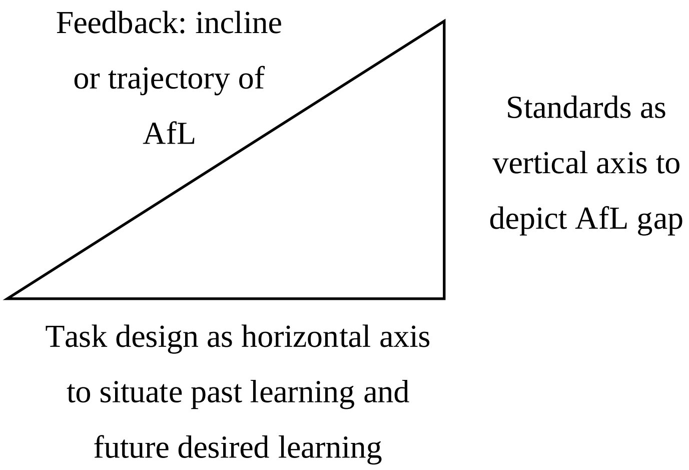

Feedback: incline or trajectory of AfL
Standards as vertical axis to depict AfL gap
Task design as horizontal axis to situate past learning and future desired learning

Figure 3.7.1 Tan's (2013) triangulated model of AfL.

The research described in this thesis adopts the A*f*L theory of Black and Wiliam (1998) in order to develop the online diagnostic test to be used as an A*f*L tool. In addition, the research will use Tan's (2013) recently proposed triangulated A*f*L framework to provide an interpretation of the findings. This interpretation includes investigating the relationship between the three main elements of A*f*L to understanding the trajectory of students' learning. In summary, the research described in this thesis intends to build on the A*f*L theory by exploring A*f*L in higher education, specifically through the use of online diagnostic tests support TES' life-long learning of numeracy skills in a sustainable manner.

Chapter Four: Methodology and Methods

4.1 Introduction

This chapter describes the chosen methodology and methods that was adopted for this research and outlines the specific research design. In addition, this chapter explains how the research is situated in the positivist paradigm, outlines the research methodology adopted, and describes the specific design of the testing instrument and analysis methods. The chapter concludes with the limitations to the research design and a reflection of the ethical considerations.

4.2 Research Design

It is understood that researchers should consider the research questions and the guidance to the investigation that they may provide (Cohen et al., 2011; DeCuir-Gunby, 2008). Therefore, the research questions in this study, exploring the extent to which TES' numeracy skills can be evaluated and improved through a practice test, guided the design. More specifically, the research design included adopting an appropriate paradigm, a suitable research methodology, and a compatible research approach that involved developing a testing instrument and suitable methods of analysis.

4.2.1 Paradigm: Positivism

Methodologies that guide research efforts include many theoretical perspectives that are informed by a range of epistemological positions. Each epistemological position attempts to explain how we know what we know and attempts to determine the position taken that informs the understandings that we reach (Crotty, 2020). It is believed that ontological assumptions bring about epistemological positions which in turn give rise to the methodological considerations in research (Cohen et al., 2017). Furthermore, in addition to ontology and epistemology is axiology, which involves the beliefs and values one holds. It is believed that axiology shifts the thinking of regarding research as simple, to being informed by how the world is viewed and what we see as the purpose for understanding (Cohen et al., 2017).

This research was guided by an objectivist approach that involved adopting realism ontology and the epistemological position of positivism. The views of realism and positivism are guided by the belief that real life is scientific and phenomena are measurable (Brown & Baker, 2007). The world of a positivist is appropriately described as a mathematical world (Crotty, 2020) because it seeks objectivity, predictability, measurability, and controllability (Cohen et al., 2017) . Similarly, methodological positivism refers explicitly to a set of scientific research practices that involve concepts of knowledge, social reality, and science (Riley, 2007). Thus, methodological positivism can be adopted for undertaking research to solve a problem while taking a positivist stance.

The research described in this thesis is positioned within the positivist paradigm. It adopts a perspective defined by Cohen et al. (2017) as the science that provides the most accurate view of knowledge. The research itself follows the practice of methodological positivism since the methods adopted are objective and quantitative. Furthermore, the study is not interested in human behaviours such as intentions, thoughts, or attitudes in which the adoption of other paradigms would be more relevant. Instead, this research is interested in abilities that can be observed and measured and fit appropriately in the positivist paradigm. More specifically, there are two elements to the research that place it in the positivist paradigm. The first element is the quantitative methods that are adopted and outlined in the next section. The second element is the acknowledgment of objectivity, validity, and generalisability to the quantitative findings.

A criticism of positivism is that it regards human behaviour as determined and controlled, and as a consequence ignores intention, freedom, and individuality (Cohen et al., 2017) . While acknowledging that this criticism is valid, the research in this study aims to make generalisations and therefore does not require consideration of individualism.

4.2.2 Research Strategy

Several research methodologies were considered for this study, acknowledging that different approaches are necessary for different research purposes (Cohen et al., 2017). For example, one intention of the research described in this thesis was to make comparisons to determine changes in ability over time. Case studies and multiple case studies are commonly used for comparative purposes (Corbetta, 2003); therefore, the case study methodology was considered a possibility for this study. However, case studies are more specifically defined as establishing cause and effect where the effects can be observed in real contexts (Cohen et al., 2017). The intention of this study was not to establish cause and effect but rather to diagnose and evaluate capabilities, confirming that case study was not the most appropriate methodology for this study.

Another methodology considered for this research was action research which is commonly known as the methodology used to generate knowledge about initiatives that are implemented to encourage change and improvement (Cohen et al., 2017; Somekh, 2006). A significant component of this research was implementing an initiative to enhance learning and bring about improvement; therefore, the action research methodology was examined and contemplated. However, in addition to exploring how the online test can improve understanding, this research also had the purpose of measuring TES' numeracy capabilities to assess performance and diagnose strengths and weaknesses. Therefore, action research was not considered the most appropriate methodology for this purpose.

After careful consideration of the purpose of the research, to evaluate skills and explore improvements, a more objective methodology was necessary to adopt. Accordingly, Testing and Assessment was decided as the preferred methodology for this study because its specific purpose is to measure achievement and potential, diagnose strength and weaknesses, and assess performance and abilities (Cohen et al., 2017) . Another characteristic of the Testing and Assessment methodology is that it enables comparisons to be made. Therefore, considering the intention of this study to test, measure, assess, and compare students' performance between one attempt and another, this methodology was considered the most appropriate for this research. Many characteristics of the Testing and Assessment methodology outlined by Cohen et al. (2017) further justify its appropriate adoption for this study. For example, Testing and Assessment includes a design to provide scores that can be aggregated, it allows in-depth diagnoses, and it will enable performance to be measured (Cohen et al., 2017) . These characteristics further confirm the suitability of testing and assessment methodology for this study.

Testing and Assessment has commonly been adopted in educational research to diagnose skills and make conclusions about abilities. Its use is widely evident in studies exploring mathematical skills in higher education students. For example, Mulhern and Wylie (2006) administered tests in their research to investigate the mathematical abilities of psychology students in the United Kingdom (UK). This methodology allowed for valid conclusions about the deficiencies in the students' mathematical thinking. Another example is Nortvedt and Siqveland (2019) study, which utilised Testing and Assessment to examine mathematical understanding and skills in engineering and calculus undergraduate students in Norway. This methodology allowed for comparisons of students' overall capabilities based on their gender.

Furthermore, Testing and Assessment has been adopted to explore TES' numeracy skills. For example, Afamasaga-Fuata'i et al. (2007) and Linsell and Anakin (2012) used tests in their study to inform conclusions about the numeracy capabilities in TES in Samoa and New Zealand, respectively. Similarly, in Australia, Sellings et al. (2018) used Testing and Assessment to explore the numeracy capabilities of TES. Overall, the adoption of this methodology allowed for the diagnoses of skills and conclusions to be made about TES' numeracy capabilities.

4.2.3 Quantitative Research Methods

It is believed that many data can be collected in quantitative ways, even those that do not appear in quantitative forms (Muijs, 2010). However, the specific adoption of the Testing and Assessment methodology in this study allowed the data to be collected and appear in an obvious quantitative form. Therefore, it is logical to adopt a quantitative approach in this research and use quantitative data collection and analysis methods. Furthermore, quantitative methods typically stem from the positivist paradigm because of the connection between positivism and objectivity (DeCuir-Gunby, 2008). Given that this study takes a positivist paradigm stance, a quantitative approach was particularly suitable for this research to allow for the findings' measurability, controllability, and generalisability.

Quantitative research can be defined as collecting numerical data and analysing the data using mathematical methods (Muijs, 2010). There are many strengths and advantages of using quantitative methods; for example, they are considered especially important because of the connections that can be made between observations and mathematical explanations (Hoy, 2010). Furthermore, they are beneficial for studying large samples (DeCuir-Gunby, 2008). Therefore, it is appropriate for this study to employ a descriptive quantitative design to explain the characteristics of a large sample through statistical calculations. This included statistical and comparative methods that relied on observations alone.

4.3 Design of Testing Instrument

Muijs (2010) argued that using an appropriate research design and data collection instrument is even more crucial than the data analysis tools. For this reason, the development of the testing instrument was a significant component of this quantitative research. Furthermore, it is believed that the design of a study should align with the epistemological position (Corbetta, 2003); therefore, the epistemological position of positivism was considered in the design of the testing instrument. More specifically, the development of the instrument required the collection of data that allowed for TES' skills to be quantitatively measured and observed through an objective lens.

Online assessments have been successful in quantitatively and objectively evaluating skills (Alcoholado et al., 2016; Snekalatha et al., 2021). Therefore, the development of an online diagnostic test was included in the research design to allow the data to be quantitatively gathered and analysed. The researcher developed the online diagnostic test with two other researchers, all of whom are qualified secondary mathematics teachers and thus have mathematics education expertise. Furthermore, the online diagnostic test development was made using the tests tool available through the Blackboard LMS.

4.3.1 Adoption of the Theoretical Framework in the Testing Instrument

Specific elements were considered in the design of the testing instrument, justifying the adoption and embedding of the overlapping elements identified in the TAM, the e-learning systems framework, the ADDIE Model, and Black and Wiliam's (1998) theory of A*f*L. This was particularly important considering that, in addition to aligning with the epistemological position, it has also been acknowledged that the research design must correspond with the theories and theoretical framework chosen (Corbetta, 2003). Thus, overlapping elements identified in the online learning models and Black and Wiliam’s (1998) theory of A*f*L were used to inform many aspects of the online test design.

Firstly, the identified overlapping elements between the online learning models and Black and Wiliam’s (1998) theory of A*f*L were identified. In particular, it was identified that the consideration of users and participants, the intention of the test/technology used, and the services that needed to be involved were overlapping elements. These three dimensions were considered and embedded throughout the design.

Furthermore, considering the ADDIE Model describes best practice in learning design, this model was carefully considered in the development of the online test. In particular, the elements that were also identified in Black and Wiliam’s (1998) theory of A*f*L were embedded. Firstly, the initial ADDIE Model phase of Analysis was used and goals to be achieved were identified to promote a positive learning experience. In Black and Wiliam’s (1998) theory, the promotion of a positive learning experience is referred to in the discussed element of goal orientation. In this research, the intention was for TES to be provided with a stress-free test experience to promote a positive learning environment.

Next, specific content to be assessed were identified to accurately assess TES numeracy capabilities. Hence, the second ADDIE Model stage of Design was then adopted where the steps the learner must take were carefully constructed in the task. That is, the test was designed to clearly assess the intended content through a testing system that was user-friendly for all students and could promote self-assessment. This element overlaps with the element of self-assessment in Black and Wiliam’s (1998) theory. In their theory, students who self-assess display higher motivation than those who do not. Therefore, the design of the online test allowed for self-assessment and self-evaluation.

The third ADDIE Model phase of Development was then used where the performance objectives were considered. In this phase, the element of feedback of the TES’ performance was carefully considered and methods of providing effective feedback were developed.

Feedback is also identified as a major element in Black and Wiliam's (1998) theory of A*f*L and was considered a crucial component in the test design.

The fourth ADDIE Model stage, Implementation, involved the overall actioning of the plan. That is, the test items were developed and entered into the testing system, the worked solutions were developed and implemented as the method of feedback, and the test settings were set to allow for accessibility for all TES in the sample.

The final stage of Evaluation, which consists of formative and summative evaluation, was used. This stage was more specifically used in the development of the Pilot Test. That is, in order to ensure quality of the test, usability of the testing systems, and to gather TES feedback on the test, the Pilot Test was developed to gather summative and formative evaluations. Summative evaluations consisted of overall scores to determine appropriate test difficulty, and formative evaluations consisted of qualitative responses to specific feedback questions asked of the participants at the completion of the Pilot Test. The evaluations informed the development of the main Diagnostic Test common element was identified to be the element of the users *and* participants. The online learning models outline that consideration must be given to the participants' perception of how easy to use and how useful the technology is. This is necessary to promote a positive attitude towards the technology.

Furthermore, the elements in Black and Wiliam's (1998) theory of A*f*L displaying links to the online learning theories were more specifically adopted and implemented into the test design. The specific adoption of these elements aimed to encourage learning through use of the test. Firstly, the test was designed to allow and encourage self-assessment and self-evaluation to promote increased motivation and achievement outcomes. More specifically, the test was designed to present TES with their overall score and indicated correct or incorrect answers for each item. This capability allowed TES to self-assess and self-evaluate areas of weakness and areas of strength. This was especially important to motivate TES to make appropriate decisions about improving their skills in targeted areas.

Secondly, the A*f*L element of response and reception discussed in Black and Wiliam's (1998) review was carefully considered to ensure that TES would receive results that would increase motivation and encourage seeking assistance. In particular, the test was designed to present specific information according to whether answers were correct or incorrect. Correct answers were acknowledged with a praise, such as ‘Correct. Well done!” and inaccurate answers were presented with detailed solutions. It was anticipated that the presentation of worked solutions would motivate students to learn the correct methods or seek further assistance.

The third A*f*L element considered in the design of the online test was goal orientation. Black and Wiliam’s (1998) review found that goal orientation encouraged TES to have positive attitudes about learning. Furthermore, they reported that negative attitudes were displayed in groups that associated lack of ability with failure and in groups that focused on the importance of out-performing others. Therefore, to encourage positive attitudes, information allowing comparisons between students’ test results was not available. In other words, other students’ scores and the test average were not available. To further encourage positive attitudes, the test was named the “Numeracy Practice Test”. It was expected that this name would emphasise that the test was intended to be used for practice and not as a formal assessment.

Fourthly, the A*f*L element of self-perception discussed in Black and Wiliam’s (1998) review was considered in this study to ensure that TES had control over their learning. This was important given that achievement has been linked to students’ sense of control over their learning (Black & Wiliam, 1998). Therefore, no interventions were made other than the worked solutions that were provided. It was anticipated that this would allow TES to have control over their learning and choose their learning methods to encourage self-directed learning rather than follow prescribed procedures.

The fifth A*f*L element considered in the design of the online test was the specific use of tests. In Black and Wiliam’s (1998) review, there was evidence suggesting that frequent testing improved learning and was favoured by students. In alignment with this, the online test allowed TES to make unlimited attempts. The specific use of the tests in this research allowed for the flexibility of enabling TES to attempt the test once, twice, or more times. Furthermore, the test was designed to allow TES to sit the test at any time as there were no restrictions (day or time) to when they could take the test. It was assumed that TES would benefit from being able to attempt the test whenever or wherever they wish, and as frequently as they wish. This was designed to enhance motivation and self-paced learning.

The last and most crucial element considered in developing the online test used in this research was the quality of feedback to be provided. Black and Wiliam (1998) considered this element the critical feature of formative assessment and provided several aspects of effective feedback in their review. For example, they found that feedback was most effective when it was designed to encourage the correction of errors with details of the correct answer instead of the general indication of whether the student’s response was correct or incorrect. Therefore, the test used in this research was designed to provide feedback that consisted of full worked solutions with easy-to-follow steps. Furthermore, the worked solutions were presented immediately after a test attempt. This ensured instant feedback, allowing TES to promptly identify the gap between their current achievement level and the standard required to master a particular skill.

The design of the test, allowing for the implementation of the specific elements outlined above, was made possible using the Blackboard LMS tests tool. The research design process that included a Pilot Test and Diagnostic Test will now be discussed in more detail.

4.3.2 Pilot Test Development

The first stage of this study consisted of a Pilot Test to benchmark the style and difficulty of questions for the main testing instrument**.** The development of the Pilot Test was made by the main researcher with the assistance of two other researchers, both with mathematics education expertise. The development consisted of the development of items, development of worked solutions, and selection of the most suitable items to be included in the test. Items were initially developed according to the LANTITE Assessment Framework (ACER, 2017) to inform the items’ overall style, content, and difficulty.

Next, the researchers examined NAPLAN papers and textbooks for ideas of questions that covered a variety of topics and contexts. The development process consisted of the development of several items specifically designed to be allocated into one of the four test categories: NA, MG, SP and NC. Items were allocated according to the mathematical content strand for Calculator-allowed questions (NA, MG, or SP), or where items were intended to be answered without the aid of a calculator; they were categorised into the NC category. Items were developed in the format of Multiple Choice, True/False, or Fill in the Blank (requiring numeric responses).

Finally, worked solutions for each item were developed. Iterative processes ensured the quality of the items and involved reviewing, examining, modifying, and improving the produced items. Once the items were finalised, ten items from each test category (NA, MG, SP, and NC) were agreed upon and selected for the 40-questions Pilot Test.

Furthermore, three open-ended questions were developed to gather student feedback. This gathered feedback intended to assess the overall test accessibility and functionality to inform the development of the main Diagnostic Test. The three questions were:

1. What did you find most useful about the site?
2. What improvements do you think should be made to the site?
3. Do you have any other general feedback about the site?

The test was designed to be presented in a specified order. The 30 calculator-allowed items (i.e., NA, MG, and SP) were presented first. Following these questions, the below message was presented to students:

> **Section 2 of the quiz is a NON-CALCULATOR section.**
>
> **The use of a calculator is NOT allowed.**
>
> **Do you agree to not use a calculator in this next section of the quiz?**
>
> □ Yes □ I prefer to use a calculator to practice

After this, the 10 NC items were displayed. Finally, after the 40 skills questions, the open-ended questions were presented. At the completion of a test attempt, TES were presented with their overall mark (out of 40). Their performance on each item was indicated as correct, or in the case of incorrect answers, the worked solutions were presented. No time limit was set for the Pilot Test, although students could see the time they took to complete the test.

4.3.3 Diagnostic Test Development

Once the Pilot Test had been administered and the results collected, an analysis was conducted on the test performance and responses to the open-ended feedback questions. This analysis allowed for a revision of the items to be undertaken and the development of the Diagnostic Test. The LANTITE Assessment Framework (ACER, 2017) was further used as an external objective measure to inform the Diagnostic Test's style, content, and difficulty of the test items.

4.3.3.1 Development of Test Items

The researcher and an external advisor, both of whom are qualified secondary Mathematics Teachers, developed 272 items in total. All questions were then reviewed by the primary supervisor of this study, who has expertise in mathematics education, teacher education, and is a qualified secondary Mathematics Teacher. Initially, items were developed in one of two test sections, according to whether the use of a calculator was allowed or not. Additionally, following the Australian Curriculum (ACARA, n.d.), each item was developed in one of three strands, according to their content strand (NA, MG, or SP). The Calculator-allowed section included a spread of items across all strands (NA, MG, and SP). However, due to the nature of the topics in the SP strand and the smaller number of topics that exist in the SP strand, no SP items were developed in the NC section. Instead, emphasis was placed on assessing NC skill capabilities in the more extensive categories. Therefore, the NC section included only items from the NA and MG strands. Furthermore, items in each strand were developed in sub-pools according to the mathematics topic that the item assessed:

Table 4.3.3.1.1 The mathematics topics assessed in each strand

Number & Algebra	Measurement & Geometry	Statistics & Probability
Algebra	Angles	Combinations
Basic Arithmetic	Area	Interpreting Data
Decimals	Capacity & Volume	Probability
Financial	Distance &Perimeter	Statistics

Fractions	Estimating, Reading & Converting
Percentages	Space, Shapes & Symmetry
Rates & Ratios	Time & Timetabling

These topics were chosen from the list of content considered appropriate for numeracy assessment provided in the LANTITE Assessment Framework (ACER, 2017).

Items were developed in three item types: Multiple Choice, True/False, or Fill in the Blank, aligning with the response formats outlined in the LANTITE Assessment Framework (ACER, 2017). Firstly, Multiple-Choice items were developed with four response options corresponding with the LANTITE Assessment Framework (ACER, 2017) and also acknowledging that four response options are most common in multiple-choice items in online tests. However, recent research argued that it is only necessary to provide a small(er) number of response options (Haladyna & Downing, 1993). Therefore, some items were developed to only include two response options. These items were developed as True/False items. The intent for this was to assess whether there was a difference in performance in questions with two or four choices, noting that the smaller number of options might allow students to easily guess the correct answer. The third item type developed in the test was Fill in the Blank, following the LANTITE Assessment Framework's (ACER, 2017) inclusion of constructed response items and supporting the research that suggests the inclusion of short response items improves test reliability (Abida et al., 2011). The Fill-in-the-Blank items were developed to require short responses or numeric responses. Furthermore, this item type was included to verify that TES were appropriately achieving correct answers and to determine whether responses on the other item types were genuine or likely provided by guessing.

In alignment with the LANTITE Assessment Framework (ACER, 2017), items were also developed in one of three context domains: Personal and Community, Workplace and Employment, or Education and Training. Items were designed to include contexts relating to these domains. For example, Personal and Community items included questions relating to everyday activities such as shopping, travel, or budgeting. Workplace and Employment items included questions relating to the teaching profession, such as problems involving schools, teachers, or students. Education and Training items included questions relating to professional learning, such as attending conferences, interpreting standardised test performances, and participating in further education opportunities.

Finally, items were developed in one of four ACSF levels (Department of Employment, 2015): Level 2, Level 3, Level 4, or Level 5. There were no Level 1 questions included as they were considered below the standard benchmark, according to the LANTITE Assessment Framework (ACER, 2017). In brief, a TES at Level 2 can identify mathematical information, selects familiar mathematical problem solving strategies, and uses informal and some formal mathematical language. A TES at Level 3 selects and interprets mathematical information, selects from and uses a variety of developing mathematical and problem solving strategies, and uses a combination of both informal and formal mathematical language. A TES at Level 4 extracts and evaluates mathematical information, selects from, and applies an expanding range of mathematical and problem solving strategies, and uses a range of informal and formal mathematical language and symbols. A TES at Level 5 analyses and synthesises highly embedded mathematical information in a broad range of tasks, selects from, and flexibly applies, a wide range of highly developed mathematical and problem solving strategies, and uses a wide range of formal, some informal mathematical language and representation (Department of Employment, 2015).

The ACSF Framework (Department of Employment, 2015) was used as well as professional judgement to develop items according to appropriate difficulty levels. Furthermore, items were developed in each level to reflect close proportions to those outlined in the LANTITE Assessment Framework (ACER, 2017). That is, most items were developed as Level 3 and Level 4 items.

An iterative process was used to ensure the quality of the items and items were reviewed, examined, modified, and improved. Finally, worked solutions were developed for each item with easy-to-follow steps. Once the items and solutions were finalised, they were all entered into the online testing system on the Blackboard LMS. Table 4.3.3.1.2 outlines the number of items developed in each test section, mathematical strand, item type, context domain, and ACSF level. Table 4.3.3.1.3 outlines the number of items developed in each content area. Examples of Calculator-allowed questions (Table 4.3.3.1.4) and NC questions

(Table 4.3.3.1.5) are presented below together with their corresponding context domain, ACSF level, mathematical strand, and content area.

Table 4.3.3.1.2 The number of items developed for the diagnostic test

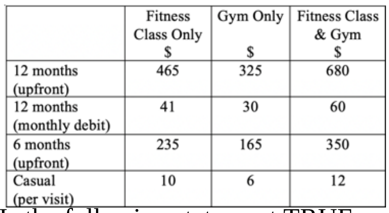
Section	**Number of Items**
Calculator Allowed	203
Non-Calculator	69
Mathematical Strand	
Number & Algebra	136
Measurement & Geometry	76
Statistics & Probability	60
Item Type	
Fill in the Blank	74
Multiple Choice	181
True/False	17
Context Domain	
Personal & Community	184
Workplace & Employment	68
Education & Training	20
ACSF Level	
ACSF Level 2	19
ACSF Level 3	157
ACSF Level 4	86
ACSF Level 5	10

Table 4.3.3.1.3 The number of items developed in each content area for the diagnostic test

Number & Algebra Content	**Number of Items**
Algebra	9
Basic Arithmetic	47
Decimals	16
Financial	15
Fractions	14
Percentages	18

Rates & Ratios	17
Measurement & Geometry Content	
Angles	7
Area	7
Capacity & Volume	6
Distance & Perimeter	7
Estimating, Reading & Converting	27
Space, Shapes & Symmetry	7
Time & Timetabling	15
Statistics & Probability Content	
Combinations	6
Interpreting Data	29
Probability	12
Statistics	13

Table 4.3.3.1.4 Examples of calculator-allowed items

<table>
<tr><th>Question</th><th>Context Domain</th><th>ACSF Level</th><th>Item Type</th><th>Strand and Content Area</th></tr>
<tr><td>Oliver turns 3 years old in 4 months.
What is Oliver's age (in months) now?</td><td>Personal & Community</td><td>Level 2</td><td>Multiple choice</td><td>MG
Estimating, Reading & Converting</td></tr>
<tr><td>Here is the schedule for Gym and Fitness Class membership at a sports facility.

<table>
<tr><th></th><th>Fitness Class Only
$</th><th>Gym Only
$</th><th>Fitness Class & Gym
$</th></tr>
<tr><td>12 months (upfront)</td><td>465</td><td>325</td><td>680</td></tr>
<tr><td>12 months (monthly debit)</td><td>41</td><td>30</td><td>60</td></tr>
<tr><td>6 months (upfront)</td><td>235</td><td>165</td><td>350</td></tr>
<tr><td>Casual (per visit)</td><td>10</td><td>6</td><td>12</td></tr>
</table>
Is the following statement TRUE or FALSE? It is better to pay the casual rate if you are going to the sports facility (gym and fitness class) four times every month, than to pay for the 12 months (upfront) package.</td><td>Personal and Community</td><td>Level 2</td><td>True/False</td><td>NA
Financial Mathematics</td></tr>
</table>

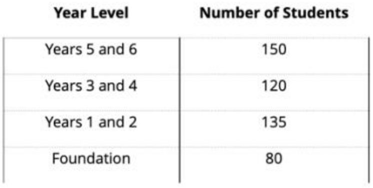
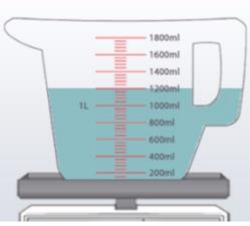
<table>
<tr><td>At a school, lunchtime is divided into two parts of equal duration. Lunchtime is scheduled between 12:40 pm and 1:30 pm. What is the duration of each part?</td><td>Workplace & Employment</td><td>Level 3</td><td>Multiple Choice</td><td>MG
Time & Timetabling</td></tr>
<tr><td>The table below shows the number of students in different year levels at a primary school.

<table>
<tr><th>Year Level</th><th>Number of Students</th></tr>
<tr><td>Years 5 and 6</td><td>150</td></tr>
<tr><td>Years 3 and 4</td><td>120</td></tr>
<tr><td>Years 1 and 2</td><td>135</td></tr>
<tr><td>Foundation</td><td>80</td></tr>
</table>
The information in the table is to be presented graphically after entering it into a spreadsheet. Which type of graph should be selected to produce the most appropriate representation?</td><td>Education & Training</td><td>Level 3</td><td>Multiple choice</td><td>SP
Interpreting Data</td></tr>
<tr><td>To make up the cordial for the school camp, Miss Richards knows that she has to mix 1 part orange concentrate to 5 parts water. If she has to make 36 litres of cordial, how much orange concentrate (in litres) will she need?</td><td>Workplace & Employment</td><td>Level 4</td><td>Fill in the blank</td><td>NA
Rates & Ratios</td></tr>
<tr><td>A restaurant offers a three-course meal with a choice of two entrees, two main courses and two desserts. The choices offered are:
ENTREES: Caesar Salad, Chicken and Corn Soup
MAIN COURSES: Char-grilled, Eye Fillet, Roasted Chicken
DESSERT: Apple Pie, Trio of Sorbets
How many different combinations of three-course meals can be served?</td><td>Personal & Community</td><td>Level 4</td><td>Fill in the blank</td><td>SP
Combinations</td></tr>
</table>

Table 4.3.3.1.5 Examples of NC items

Question	Context Domain	ACSF Level	Item Type	Strand and Content Area
This jug has water in it.  1800ml 1600ml 1400ml 1200ml 1L 1000ml 800ml 600ml 400ml 200ml Richard empties 365mL of water out of the jug. How many mL of water remains in the jug?	Personal & Community	Level 3	Fill in the blank	MG Estimating, Reading & Converting
A Combined Primary and Secondary Education Conference has the following registration costs. All registration costs include 10% GST. Saturday Registration Member $250 Non-member $245 Sunday Registration Member $230 Non-member $325 Saturday & Sunday Registration Member $460 Non-member $565 Conference Dinner $150pp Mark is a member and is attending the Conference on both days as well as going to the Conference Dinner. Find the total cost of Mark's registration.	Education & Training	Level 3	Fill in the blank	NA Basic Arithmetic
A student has 24 green pencils and 16 pink pencils. What fraction of the student's pencils are pink?	Workplace & Employment	Level 3	Multiple choice	NA Fractions
Staff meetings at Rose Hill High School are held on the first Monday of each month. There was a meeting held on 7 May. What is the date of the June meeting?	Workplace & Employment	Level 4	Multiple choice	MG Time & Timetabling
Miss Peterson gives 5 pencils to each of her 24 students. She has 4 pencils left over. She wants each student to have 7 pencils in total. How many more pencils does Miss Peterson need?	Workplace & Employment	Level 4	Fill in the blank	NA Basic Arithmetic

4.3.3.2 The Structure of Each Test Attempt

The test was designed to draw a select number of items from each of the pools. The first section of the test included 30 Calculator-allowed items across each strand category (NA, MG, and SP), with a specified number randomly drawn from each topic (Table 4.3.2.2). Thus, students received the same spread of questions but were exposed to different items on each test attempt.

Similar to the Pilot Test, after the 30 Calculator-allowed items were presented, the test displayed a screen requiring TES to agree not to use a calculator for the remaining questions. Following this, ten randomly selected NC items were presented and drawn from the pool of 69 NC items.

Table 4.3.3.2.1 Number of items drawn from each content area in the calculator-allowed section of the main diagnostic test

Content	Number of Items in Test
Algebra	1
Angles	1
Area	1
Basic Arithmetic	3
Capacity & Volume	1
Combinations	1
Decimals	1
Distance & Perimeter	1
Estimating, Reading & Converting	3
Financial	1
Fractions	1
Interpreting Data	5
Percentages	1
Probability	2
Rates & Ratios	2
Space, Shape & Symmetry	1
Statistics	2
Time & Timetabling	2

Furthermore, the test was designed to allow TES to skip questions, return to previous questions at any time, and pause a test attempt at any point and resume it at a later time or date. However, a test needed to be saved and submitted before results were received. After a test was saved and submitted, TES received their overall mark (out of 40). In addition, their performance on each item was indicated as correct, or in the case of incorrect answers, the worked solutions were presented. No time limit was set for the Diagnostic Test, although students could see the time they took to complete the test.

4.4 Data Collection Methods

All students enrolled in an initial teacher education program from two Australian universities were invited to participate in the study. Participants included a mix of undergraduate and postgraduate students. The testing instrument, available through each of the Institution's LMS, was available to all students within this population for the duration of the study. The population size was approximately 1500 for Institution A and 2000 for Institution B. Applying a 95% confidence level and 5% margin of error, the ideal sample size was 306 for Institution A and 323 for Institution B (Australian Bureau of Statistics, n.d.). In total, there were 878 students from Institution A and 405 students from Institution B who participated in this study, which exceeded the minimum sample size required based on the above power calculation.

It is important to note that there were differences between the two Institutions. For example, different program structures exist at each institution as different education degrees are offered. There were also different demographical factors between the two institutions, such as the proportion of domestic and international students enrolled and different cultural backgrounds. Although these factors were not captured in this study, their potential influence and significance were considered and discussed in the results.

Learning analytics data were collected using the testing instrument available through the Blackboard LMS. Data were collected for the Pilot Test between October and mid-December 2017 and between March 2018 and March 2019 for the Diagnostic Test. Both data sets were downloaded through the grade centre section and included selecting the option to download all attempts data by question and user. Specific data collected for each user and attempt had the items administered, students’ responses, and the score for each question (1 for correct and 0 for incorrect). Student names and identification numbers were deleted to ensure that students could not be identified.

4.5 Methods of Data Analysis

4.5.1 Pilot Test Data Analysis

The Pilot Test data acquired from the Blackboard LMS was imported and processed using Microsoft Excel. Statistical analysis was performed using GraphPad Prism (version 8.0.1). Measures of central tendency and spread were determined, including mean ±SEM, median, minimum, and maximum. Assessment of statistical significance was also performed in GraphPad Prism. Statistical significance was assessed using the non-parametric Kruskal-Wallis with Dunn’s multiple comparisons *post-hoc* test to compare the mean of each category in the test. Results were considered statistically significant where $p<0.05$.

4.5.2 Diagnostic Test Data Analysis

Diagnostic Test data were downloaded from the Blackboard LMS and processed using Microsoft Excel. Information on the test sections, mathematical strands, content areas, item types, context domains, and ACSF levels for each data point were then added.

Various analyses were performed for the Diagnostic Test data, including visual and statistical analyses, using GraphPad Prism (version 8.0.1). Measures of central tendency and spread were determined, including mean ±SEM, median, minimum, and maximum. The unpaired non-parametric Mann-Whitney U test was used to determine whether the performance of the institutions was different and the non-parametric Kruskal-Wallis with Dunn’s multiple comparisons *post-hoc* test was used to compare the performance between categories, content areas, items types, context domains, and ACSF Levels at each institution. Results were considered statistically significant where $p<0.05$.

4.5.3 Rasch Measurement Model Analysis

Analyses applying the Rasch Measurement Model were necessary to accurately compare one attempt with another in overall performance and to examine development in ability in test categories, content areas, item types, context domains, and ACSF Levels. The data collected in the study consisted of items with only two possible responses (correct or incorrect); therefore, the Rasch analysis for dichotomous data was applied using Winsteps (version 4.6.1). Data were entered into the Winsteps control setup file and included codes to indicate the identification of attempts, attempt numbers, and the scores for each item (1 or 0) presented on a test attempt.

A crucial expectation of applying the Rasch Model is that the data must conform to the Model's requirements (Bond & Fox, 2015; Iramaneerat et al., 2008); therefore, the steps involved in this analysis were quite complex and involved. In summary, the Rasch analysis in this study ensured Model fit by conducting an item and attempt measure overview, identifying misfitting items and attempts, removing misfitting items and attempts, recalculating measures, calculating item invariance, and generating item anchors. A more thorough description and explanation of these steps are provided below.

4.5.3.1 Attempt and Item Analysis

The initial attempt and item overview analysis consisted of calculating measures of central tendency and spread, including mean ±SEM, minimum, and maximum measures. Logit measures were also calculated, allowing for comparisons between the raw scores and logit measures. Visual observations were also conducted from a Data Matrix to determine if results were produced as expected.

4.5.3.2. Rasch Model Fit Analysis

The next stage included a further investigation into all attempts and items using Rasch analysis to ensure fit to the Rasch Model. This required an analysis of response residuals and fit statistics to help detect differences between the collected data and the Rasch Model prescriptions. A response residual is the difference between the observed score and the expected score. A small residual indicates that a response is close to the Rasch Model's expectation (e.g., a capable person performed as expected on an easy item). A large residual indicates that a response is quite different from the Rasch model's expectation (e.g., a capable person has unexpectedly scored 0 for an easy item). Residuals range from -1 to +1. Negative values are derived from incorrect responses, and positive residual values are derived from correct responses. For example, when a person's ability is the same as the item's difficulty, the residual will be 0.5 (if they gave a correct response) or -0.5 (if they gave an incorrect response). These are considered the same size residual (i.e., -0.5 is not smaller than +0.5). The residuals are squared to result in all positive values to be combined and accumulated without cancelling each other out. Each raw residual is standardised by using the variance of that residual, and it is the standardised residual (z) that is used to calculate fit statistics (Bond & Fox, 2015).

Chi-square fit statistics were used to calculate the residuals to determine how well the data fit the requirements of the Rasch Model. The Winsteps software reported the fit statistics as two chi-square ratios; infit and outfit mean square statistics. The mean square is the unstandardised form of the fit statistic and is the mean or average value of the squared residuals for any item or attempt. The infit and outfit were calculated in the analysis and reported in standardised form (z). Results were considered acceptable when the mean square values were 0.5-1.5, and values of z were 0±2. A thorough analysis of these fit statistics for individual items was the first step to determine whether the data fits the Rasch Model's requirements.

Further analysis of both items and attempts were necessary to ensure Rasch Model fit, beginning with an examination of the specific functioning of the items. To understand the functioning of the items and to visually observe their fit to the Rasch Model, the analysis included the development of an Item Pathway. Item Pathways represent the developmental acquisition of cognitive reasoning ability. In other words, item difficulties can be observed, and the error estimates and their (infit) fit statistics to determine if the items all fit the same developmental pathway.

Considering that Item Pathways only use infit statistics as the indicator of (mis-)fit, the analysis then considered fit statistics that considered the two necessary aspects of fit, infit and outfit. This item functioning analysis included reporting on the item measure as logits, SE, the fit statistics, and the reliabilities of item estimates. The items' statistics were examined, and results were considered acceptable when the mean square values were 0.5-1.5 and values of z were 0±2.

In the case of items being observed in an unacceptable range, the subsequent analysis aimed to determine whether eliminating the misfitting attempts would improve the functioning of the items to fit the requirements of the Rasch Model better. To achieve this, attempt statistics were calculated, including fit statistics and point-measure correlation statistics. Again, results were considered acceptable when the mean square values were 0.5-1.5, and values of z were 0±2. Finally, misfitting attempts were removed, and the *post-hoc* changes in the item and attempt statistics could be analysed and reported.

4.5.3.3 Item Invariance

It was crucial to examine the results from this approach further to determine and confirm that removal of misfitting attempts was appropriate in developing the item anchors to be used as the measurement scale for the test. Therefore, the subsequent analysis included examining item invariance, which is believed to be an essential measurement principle for any testing situation (Bond & Fox, 2015). Invariance involves dividing the sample into two, according to their ability and estimating the difficulties for each group. The goal is to establish item difficulty values such that those values are relatively invariant across the two different subsamples for their intended purpose. However, measurement of invariance cannot be achieved when differential item functioning is present in a scale.

Invariance analysis was performed by dividing the attempts into two subsamples according to ability. High-ability attempts (score ≥31) and low-ability attempts (score <31). Pairs of calibrations were then produced and plotted for each item using the pair Rasch-modelled ability estimate measures in logits. These Rasch item measures were based on the raw score totals for each subsample (high and low ability). Invariance was examined by determining whether the distribution of the plotted points was close enough to the modelled relationship (diagonal line, y=x). Items were considered sufficiently invariant when points lay inside the outer curved quality-control lines (95% based on the standard errors for each item pair).

4.5.3.4 Performance Improvement Analysis

The thorough analyses described above ensured the data fits the Rasch Model and produced reliable item anchors. The item anchors could then be applied for further analysis to enable an accurate analysis of individual test attempts, including measurements, comparisons, and identification of development in ability. This consisted of three primary analyses: overall comparisons of performance for each test attempt, stacking to compare individual students' first attempt to final attempt performance, and racking to identify changes in capabilities on the items, from first to final attempts. In addition, the racking analysis examined changes in capability in individual test categories, content areas, item types, context domains, and ACSF Levels. In addition, mean measures were calculated to determine central tendency, and scatterplots were produced to aid visual analysis of areas of improvement and areas of decline. Finally, paired *t*-tests were performed to assess statistical significance between the first and final attempt results. Results were considered statistically significant where $p<0.05$.

4.6 Limitations of the Design of the Test

It has been acknowledged that collecting online data is efficient and convenient, and specifically allows for the potential of a large amount of data (Lefever et al., 2007). Although a limitation of this research is that the samples in this study only represent two universities, rather than the state of overall TES' numeracy achievement, there are over 3000 education students enrolled between the two institutions. Therefore, the online data collection system allowed for a substantial number of potential participants in this study. Thus, while the specific results of this study may not be generalisable beyond these universities, the sample represents a large number of participants that allowed for valid conclusions to be made about the overall state of TES' numeracy achievement.

Another limitation of the research is that students might see the same questions in different test attempts. However, given the volume of items in each pool, there is a low chance that this will occur. Furthermore, the Rasch Model was applied to ensure that accurate measures of ability could be calculated and valid comparisons could be made despite repeated attempts made on some items.

Another limitation is that the students may have used a calculator in the NC section of the test. If this was the case, student performance in the NC section might be lower than reported in this study. However, given the fact that the test used in the research was not a formal assessment associated with any unit of study, there is no reason to expect that students would need to cheat the system. Students were likely to use the test appropriately and as intended.

4.7 Ethical Considerations

The use of online technologies for educational research has seen the emergence of new ethical issues (Girvan & Savage, 2012) that were necessary to examine because of the specific use of online technologies in this study. Therefore, the complexity of such ethical considerations in the online environment was examined in this study, as well as traditional ethical issues. For example, this research examined ethical issues such as informed consent, privacy protection and identity, and confidentiality.

One of the most critical ethical practices in research is obtaining informed content. However, the issues involved with informed consent for online research consist of how it can be obtained and how it can be validated. This is particularly challenging compared to simply distributing and collecting informed consent forms in physical settings (Girvan & Savage, 2012). Consent was necessary for this study to ensure that students were aware of their participation in the research; however, obtaining consent from hundred and thousands of participants can be challenging. To address this issue, all relevant information was presented to TES on-screen before the commencement of a test attempt to ensure that participants were fully aware of their involvement in the study. More specifically, students were informed that their results would be used and analysed for research purposes and their commencement of the test indicated their consent. To that end, although consent was not physically observed, students' participation beyond this point was considered as consent.

Another crucial ethical practice in research is ensuring privacy to protect participants' identities. The issue of identity is related to both informed consent and privacy protection (Girvan & Savage, 2012). Personal information such as names, student numbers, or any other identifiable information was not used in this research. Furthermore, one of the participating institutions specifically requested to remain anonymous. Therefore it was decided to treat both institutions anonymously by labelling them as Institution A and Institution B. Therefore, all data were de-identified prior to analysis, and reported without compromising the identities of the participants or the institutions. In addition, all collected data was protected and safeguarded in files only accessible by the researcher and primary supervisor.

The Research Ethics Committees reviewed and approved this research at Institution A (Reference Number RH12524) and Institution B (Reference Number 017040S).

Chapter Five: Pilot Test Results

5.1 Introduction

In the context of test development, content validity is the extent to which a test assesses what it was designed to assess (Gipps, 2011). In this research, the development and implementation of a valid test was essential to measure exactly what was intended to address the research questions. In particular, we sought to evaluate the extent by which an online diagnostic test can be used to evaluate TES' numeracy capabilities, their mathematical strengths and weaknesses, and track their numeracy development over time.

This chapter outlines the results of the Pilot Test and discusses how these findings were used to inform the development of the main Diagnostic Test. It is important to highlight that the Pilot Test was not intended to be used to make conclusions about TES' numeracy abilities. Instead, it was used to inform the development of the main Diagnostic Test, including the overall test difficulty. Therefore, data from the Pilot Test was assessed to determine whether each of the four test categories (NA, MG, SP, and NC) were of equal level of difficulty.

To further assist in the development of the main Diagnostic Test, additional data were collected from the TES to evaluate the usability of the Pilot Test. More specifically, feedback was collected and used to evaluate the accessibility and functionality of the test.

5.2 Details of Data Collected

The Pilot Test was developed and trialed through the Blackboard LMS, which was used at both institutions. It was made available to all initial TES at both institutions from October until December 2017. In this test, there were 10 questions in each mathematical strand of NA, MG, and SP presented in mixed order. Students were instructed at the start of the test that a calculator could be used to solve these 30 questions. After these 30 questions, a prompt appeared, which asked whether students agreed to no longer use a calculator for the remaining questions before 10 NC questions were presented. The questions in the Pilot Test consisted of 27 Multiple-Choice items, 11 Fill-in-the-Blank items, and 2 True/False items.

Additionally, three open-ended questions were presented to all TES who completed the Pilot Test. These questions were intended to gather feedback to further inform the development of the main Diagnostic Test. The three questions were:

1. What did you find most useful about the site?
2. What improvements do you think should be made to the site?
3. Do you have any other general feedback about the site?

Once each test attempt had been completed, the system allowed each student to see their result as a mark out of 40 and as a percentage. Furthermore, all of the questions were displayed, indicating whether a correct or incorrect response had been given. For incorrect responses, the correct answer was provided with feedback displaying worked solutions.

In December 2017, learning analytics data from Blackboard LMS were collected from the Pilot Test at both institutions. There were a total of 75 attempts made on the Pilot Test during the administered period across the two institutions. Data collected for each attempt included the questions administered, students' responses, and the score for each question (1 for correct and 0 for incorrect). When a test attempt displayed more than 8 unanswered questions (>20%), it was considered a non-genuine attempt and therefore removed from the analysis. This left 56 genuine attempts, which were analysed and reported in this chapter.

5.3 Results

5.3.1 Overall Test Performance

Of the 56 genuine attempts made by TES at both institutions, the mean score was 27.7±0.96 out of 40 (Figure 5.3.1A). Scores ranged from 6.0 to 39.0, with a median score of 28.0 (Figure 5.3.1B). For the test, a pass was determined where a score of at least 50% was achieved. It is important to note that the questions and difficulty of the test were not exactly like it would be in the LANTITE. This was not the purpose of the test and it was also not intended to specifically model or recreate the LANTITE. Rather, this was an in-house assessment to evaluate TES numeracy skills. Therefore, when considering the pass rate of our test, and comparing to the pass rate of the LANTITE, we acknowledge that the standard of a LANTITE pass is not necessarily scoring 50%. The comparisons are simply observing pass rates. Overall, 87.5% of students passed the Pilot Test (i.e., scored 20 or higher), which is slightly lower than the reported pass rate of approximately 90% in the numeracy component of the LANTITE (Barry, 2017; School News, 2020).

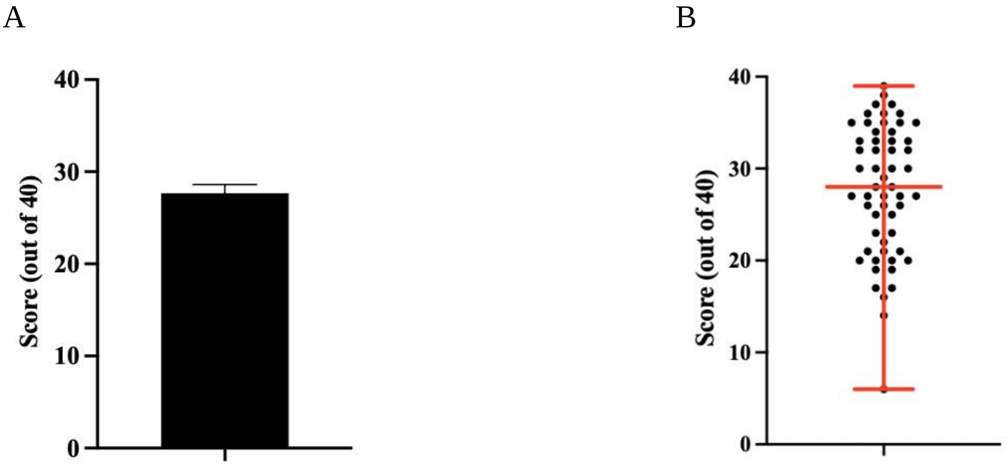


Figure 5.3.1.1 Overall performance in the pilot test. (A) Column represents the mean score ± SEM (n = 56). (B) Dots indicate the spread of individual students' scores. The horizontal red lines represent the minimum, median, and maximum scores.

5.3.2 Performance in Test Categories

The Pilot Test consisted of two sections; Calculator-allowed and NC. The Calculator-allowed section made up 75% of the test (30 questions), and the NC section made up the remaining 25% of the test (10 questions). In the Calculator-allowed section, the three mathematical strands were equally assessed. That is, there were 10 questions for each of the strands (NA, MG, and SP). The NC section consisted of items that included a mix of NA items and MG items. There were no SP questions in the NC. Before students commenced the NC section, they were presented with a prompt informing them that the next section should be completed without using a calculator. Although students must select the option which states that they agree to comply with this request to proceed to the NC section, it was impossible to know whether a calculator was actually used or not. According to these descriptions, the Pilot Test results are discussed in terms of the four categories (NA, MG, SP, and NC). Results for each category were collected and expressed out of 10 marks.

Analysis of TES' mean scores showed that the highest mean was evident in the NA category (7.11±0.26 out of 10), followed by MG (7.00±0.27) and then SP (6.93±0.25) (Figure

5.3.2A). In comparison to the three categories that permitted the use of a calculator, TES performed poorer in the NC category (6.63±0.36) (Figure 5.3.2). There was no statistical difference evident between any of the categories. Assessment of the spread of scores also showed that TES performed similarly across all four categories, with a median of 7 out of 10 (Figure 5.3.2B). These results show that the questions in each category were at an approximately equal level of difficulty.

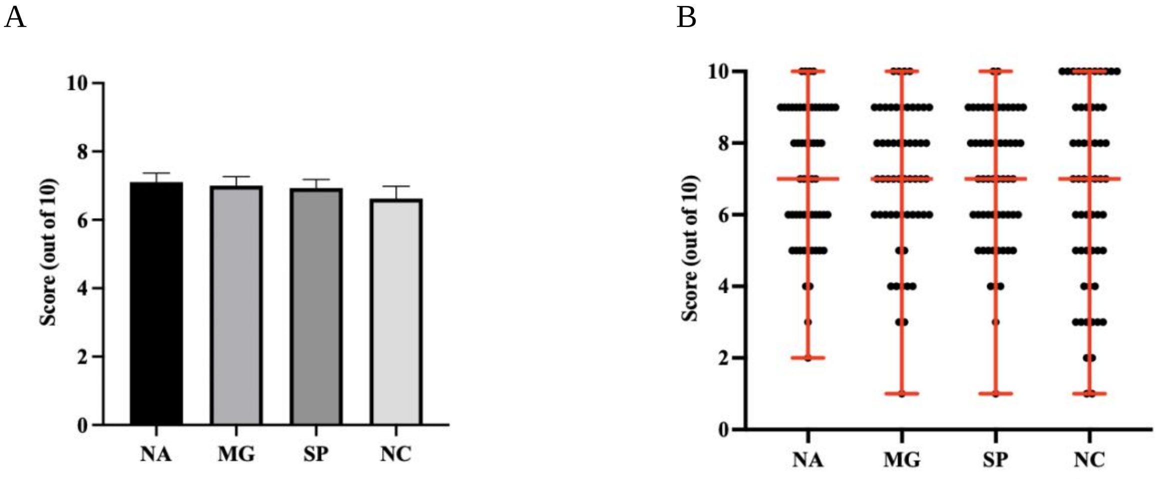


Figure 5.3.2.1 Performance in each category of the pilot test. (A) Columns represent the mean score ± SEM for each category (n=56). (B) Dots indicate the spread of individual students' scores. The horizontal red lines represent the minimum, median, and maximum scores.

5.3.3 TES' Feedback of the Pilot Test

To evaluate the usability of the Pilot Test, feedback was collected from TES through three optional open-ended questions. Firstly, TES responded to what they liked best about the site. Examples of TES' responses were: "the overall variety of questions", that it was "user friendly", the test "helped me with time management", it was "good practice", "it helped me to know what I still need to study", it was "similar to national numeracy test" and it was "very useful".

When asked about improvements they think should be made to the site, many TES did not express that any were necessary. More specifically, 22% of the TES who responded said that they either did not know of any improvements needed or specifically gave a response that indicated no improvements were needed at all. Examples of these responses included "no improvements need to be made", "no improvements are required", "I do not recommend any", and "it is as good as is". However, there were valuable suggestions for improvements provided by some TES. For example, it was evident that some wanted more questions added to the test, the ability to return to previous questions, and more questions on certain topics. Specific examples of these responses included "add more questions to the quiz", "allow us to go back to the previous questions", "add more questions about graphs", and "have more statistics questions".

Finally, TES were asked to provide additional general feedback. Most TES (53%) who responded indicated that they had no general feedback. However, for those who did have general feedback, responses were similar to those from the other feedback questions. For example, the responses included "I prefer more questions", the site was "very helpful" and "it was a great opportunity for me to practice". Additionally, there were positive responses that highlighted the TES' appreciation of using the site. For example, one TES expressed "thanks for the opportunity, I found it a great way to test myself after doing lots of practise" and another TES responded "the site is good, I like it very much. Thank you for this opportunity".

In summary, the feedback regarding the test and its functionality was very positive. In general, the feedback collected showed that students found the test to be beneficial, good practice, and useful to help identify what they needed to practise most. Overall, the evidence suggested that the TES liked the test.

5.4 Discussion

The purpose of the Pilot Test was to assess the suitability and validity of the test and the functionality of the testing environment, which was the Blackboard LMS. This was achieved by evaluating the level of difficulty in the overall test and the test categories, and collecting and evaluating feedback from the students.

Firstly, the analysis of the Pilot Test results showed that most TES (87.5%) passed. This pass rate was close to the reported LANTITE pass rate of 90% (Barry, 2017; School News, 2020). Therefore, the results suggested that the Pilot Test assessed skills at approximately the same level as those assessed in the LANTITE. Given how close the pass rates were, the similarities suggested appropriate suitability of Pilot Test to the TES sample.

In the evaluation of performance in the categories of the Pilot Test, no performance differences were seen between any categories (NA, MG, SP, and NC), indicating that no category was easier or more challenging than any other. Therefore, the results from the Pilot Test suggested a consistent spread of item difficulties in all categories. It was outlined in the LANTITE Framework (ACER, 2017) that a range of item difficulties is necessary to target an anticipated benchmark. This is catered for through a large number of test items that are included in each test. Additionally, the LANTITE Framework (ACER, 2017) outlined the target proportions of difficulty levels to be included in each test; however, these proportions of item difficulty are prescribed for the overall test. That is, there is no description of the proportion of item difficulties within each section of the LANTITE (NC or Calculator-allowed) or in each mathematical strand assessed (NA, MG, or SP). Therefore, it is assumed that the proportions of item difficulties are to be distributed equally among all sections and strands. The findings from the Pilot Test suggest an equal distribution of item difficulties among the categories, supporting the LANTITE Framework (ACER, 2017). This further confirms the suitability of the test to the target TES sample.

Furthermore, given that the Pilot Test results were collected and analysed to determine performance in the separate categories (NA, MG, SP, or NC), this suggested that the test could be used to evaluate skills in greater detail. For example, it was determined that questions could be categorised into content areas and therefore could be examined to identify more specific strengths and weaknesses. Additionally, data could be analysed to allow for skills to be compared between different groups to determine trends in capabilities. Therefore, the specific function of the test, and the data that could be collected and analysed, suggested that the test could be used more extensively to make generalisations about TES' numeracy skills.

The overall test accessibility and functionality were assessed through the Pilot Test results and the feedback gathered from the TES participants. Firstly, the overall performance of TES indicated a high pass rate, suggesting that the test functioned well. That is, there were no indications of any issues with the test accessibility or how it functioned to collect TES responses. Secondly, the consistent results in the categories suggested no obvious errors arose within any test category. Furthermore, in the feedback responses collected, TES did not provide any negative feedback regarding the functionality of the test. In fact, no one commented at all about how the test functioned. This further confirmed that the tests' accessibility and functionality were satisfactory.

Findings from the Pilot Test and the Feedback questions were used to guide the development of the Diagnostic Test. Firstly, the spread of item difficulties in the Pilot Test was deemed suitable to the TES sample and therefore used as the approximate benchmark for the items developed in the main Diagnostic Test. Secondly, findings showed that the test allowed data to be collected and analysed in categories. Extending on this idea, it was decided to develop the main Diagnostic Test to allow for an even more thorough evaluation of skills. More specifically, to allow for skills to be examined in content areas, questions types, context domains, and difficulty levels.

Finally, TES' feedback to the open-ended questions were taken into consideration to guide the development of the main Diagnostic Test. In particular, several TES suggested more questions be added; therefore, it was decided that more questions were to be developed for the main Diagnostic Test. However, rather than include more questions and develop a longer test, it was decided to create a large number of items to be drawn from for each test attempt. This would allow TES to attempt the test multiple times and not be presented with the same questions on each attempt. Furthermore, many TES requested more questions with graphs. This is likely due to the common perception that the LANTITE includes many questions with graphs. Nonetheless, it was decided to develop the main Diagnostic Test with more items involving reading and interpreting graphical representations. Furthermore, the specific request of students wanting the ability to return to previous questions while taking the test was also considered and implemented in the design of the Diagnostic Test.

Overall, the results from the Pilot Test confirmed the tests' functionality, validity, and suitability to assess TES' numeracy capabilities. The data collected from the Pilot Test and the feedback questions were analysed, and the findings influenced and guided the development of the main Diagnostic Test. This was especially important as a quality instrument needed to be developed to allow for TES' numeracy capabilities to be thoroughly evaluated and tracked over time.

Chapter Six: Diagnostic Test Results

6.1 Introduction

Although previous studies have looked at TES' mathematical abilities for specific content areas (Glidden, 2008; Izsák et al., 2010; Livy et al., 2012; Lo & Luo, 2012; Son & Lee, 2016; Tobias, 2013; Zembat, 2010), little research has been conducted on TES' numeracy capabilities. The latter differs as it incorporates the application of mathematics skills in an everyday context. Therefore, this chapter specifically addressed this gap in the literature by evaluating TES' performance in a numeracy Diagnostic Test we developed based on ACER's LANTITE Assessment Framework (2017) and results from our Pilot Test (see Chapter Five).

Based on the Pilot Test results in Chapter Five, the Diagnostic Test was designed to form the basis of the research project and a total of 272 items were developed. For each attempt in the Diagnostic Test, 40 questions were drawn from these items. Similar to the Pilot Test, the Diagnostic Test consisted of two sections: Calculator-allowed and NC. The Calculator-allowed section made up 75% of the test (30 questions), and the NC section made up the remaining 25% of the test (10 questions). Therefore, items were developed as either Calculator-allowed or NC. After the 30 Calculator-allowed questions, TES were presented with a screen that informed them of the next section being NC, and they were advised not to use a calculator. Although they selected an option to agree to this, it was impossible to know whether a calculator was actually used or not.

Following the Australian Curriculum (ACARA, n.d.), each item was developed in one of three strands, according to their content strand (NA, MG, or SP). In the Calculator-allowed section, the three mathematical strands were assessed equally. That is, there were 10 questions from each strand. The NC section consisted of items selected from a pool of questions consisting of a mix of items in the strands of NA and MG. There were no SP strand items in the NC section.

Within each strand, items were developed in pools of mathematical content areas according to content considered appropriate for numeracy assessment in the LANTITE Assessment Framework (ACER, 2017) and following the Australian Curriculum (ACARA, n.d.). The pools of content areas were: Algebra, Angles, Area, Basic Arithmetic, Capacity and Volume, Combinations, Decimals, Distance and Perimeter, Estimating, Reading and Converting, Financial Mathematics, Fractions, Interpreting Data, Percentages, Probability, Rates and Ratios, Space, Shapes and Symmetry, Statistics, or Time and Timetabling. A specific number of items were drawn from each of these content pools in the Calculator-allowed section of the test to ensure that an even spread of mathematical areas was presented on each test attempt (as outlined in Chapter Four). Items in the NC section were randomly drawn from the 69 NC-developed items.

In alignment with the LANTITE Assessment Framework (ACER, 2017), items were developed in one of three item types: Multiple Choice, True/False, or Fill in the Blank, and in one of three context domains: Personal and Community, Workplace and Employment, or Education and Training. Additionally, items were developed in one of four ACSF levels according to the ACSF levels (Department of Employment, 2015): Level 2, Level 3, Level 4, or Level 5.

In this chapter, we present TES' raw performance scores captured through the Diagnostic Test. Specifically, we evaluated TES' performance in the Calculator-allowed and NC section, in the three mathematical strands (NA, MG, & SP), and under different question conditions (item type, context, & difficulty). Through these analyses, we sought to evaluate TES' overall numeracy performance and their mathematical strengths and weaknesses.

6.2 Results

Across the two institutions, there were 1283 attempts made on the Diagnostic Test between March 2018 to March 2019. Of these attempts, 878 were from Institution A and 405 were from Institution B. Data collected for each attempt included the questions administered, students' response, and the score for each question (0 or 1). All participants were enrolled in an initial teacher education degree from a mix of undergraduate or post-graduate programmes.

6.2.1 Overall Test Performance

An initial assessment of attempts from each institution was performed to gauge the distribution of TES' overall test performance (Figure 6.2.1.1). The distribution of scores for both institutions appeared similar, which was negatively skewed with a large lower-end tail. A descriptive statistic of the cohorts showed that the overall mean performance was 21.7±0.4 for Institution A, 24.5±0.6 for Institution B, and 22.6±0.3 when combined. The median was 26.0 for Institution A, 29.0 for Institution B, and 27.0 when combined (Table 6.2.1.1).

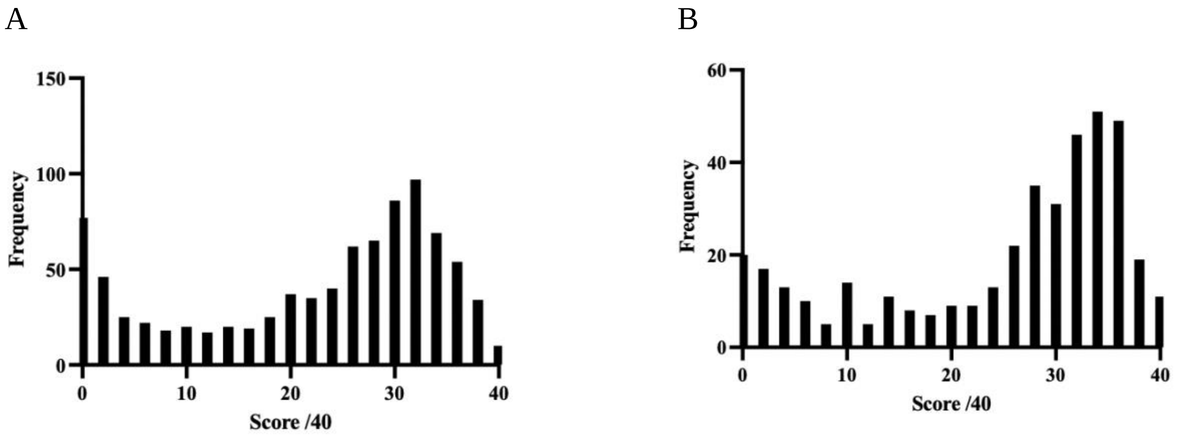


Figure 6.2.1.1 Distribution of scores in the diagnostic test. (A) The graph represents the distribution of scores at Institution A. (B) The graph represents the distribution of scores at Institution B.

Table 6.2.1.1 Summary statistics of all attempts made in the diagnostic test

	Institution A (n = 878)	**Institution B (n = 405)**	**All Results (n = 1283)**
Mean	21.7	24.5	22.6
SD	12.2	11.9	12.2
SEM	0.4	0.6	0.3
Median	26.0	29.0	27.0
Range	40.0	40.0	40.0
Minimum	0.0	0.0	0.0
Maximum	40.0	40.0	40.0

Considering the existence of very low scores (scores <10) and that our Pilot Test showed that 87.5% of TES passed the test, it was important to determine whether these attempts were genuine attempts and if they should be included in further analysis. This was a critical step to ensure that the analysis data accurately represented TES's numeracy ability. To make this distinction, all attempts where 8 or more items were left unanswered (i.e., ≥20% of the test) were evaluated. There were attempts where the student had logged into the test and exited before answering any questions, which was the case for virtually all attempts that scored zero (n=97). In most other cases, unanswered items were seen consecutively, suggesting that these attempts were not genuine. Additionally, some students started the test and only answered a few questions before leaving the test. A number of students answered a scattered selection of questions throughout the test and left all other questions unanswered. Although the reasons why these students left items unanswered are unknown, these observations suggest that these attempts were non-genuine. We considered questions answered incorrectly (i.e., not blank or unanswered) as genuine attempts on those item. Applying these definitions, there were 283 non-genuine attempts from Institution A and 115 non-genuine attempts from Institution B. These attempts were eliminated from the dataset. All other attempts were considered genuine attempts. In total, there were 885 genuine attempts made by 385 students. Of these genuine attempts, there were 595 attempts made by 267 students from Institution A, and 290 attempts were made by 118 students from Institution B.

Upon removing the non-genuine attempts, the data were assessed for normality, which was assessed both visually and through normality tests. Visually, the distribution of scores from both institutions were unimodal, negatively skewed, and appear normally distributed (Figure 6.2.1.2). However, the Shapiro-Wilk Test indicated that the data were not normally distributed (p>0.05). Therefore, non-parametric statistical tests were used in all further statistical analyses.

In the interest of gaining a more accurate understanding of TES' overall test performance, after non-genuine attempt removal, a descriptive statistical analysis for results from Institution A, Institution B, and all results were conducted (Table 6.2.1.2). There were variations in results between Institutions. Institution A had a lower mean score, median, and minimum score than Institution B. The maximum score was the same at both institutions, which was 40 out of 40.

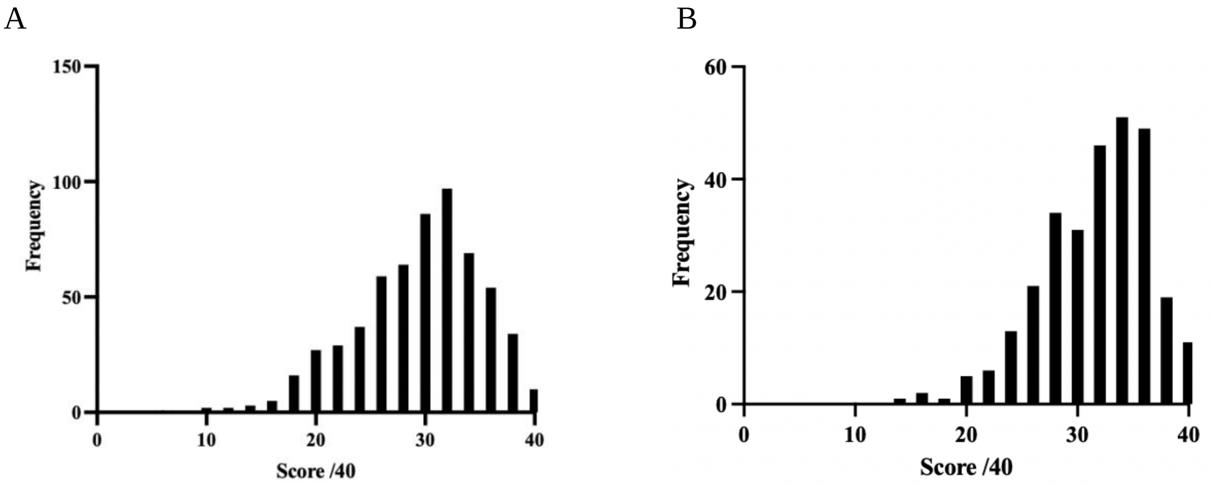


Figure 6.2.1.2 Distribution of scores after removing non-genuine attempts. (A) The graph represents the distribution of scores at Institution A. (B) The graph represents the distribution of scores at Institution B.

Table 6.2.1.2 Summary statistics of genuine attempts made in the diagnostic test

	Institution A (n = 595)	**Institution B (n = 290)**	**All Results (n = 885)**
Mean	29.0	31.1	29.7
SD	5.7	4.9	5.6
SEM	0.2	0.3	0.2
Median	30.0	32.0	30.0
Range	34.0	26.0	34.0
Minimum	6.0	14.0	6.0
Maximum	40.0	40.0	40.0

Considering the different program structure and demographical factors between the two institutions, it is important to determine whether data between the institutions were statistically similar and therefore can be combined into one dataset, or significantly different and should be analysed separately. A Mann-Whitney U test of the data between the two institutions showed that there was a statistically significant difference between institutions (Figure 6.2.1.3B), which informed the decision to analyse the data separately. It is important to highlight that the purpose of this study was to generalise TES' numeracy capabilities as well as strengths and weaknesses, and not to compare results between institutions. Therefore, we focused on identifying trends within each institution and the similarities between the institutions, rather than their differences.

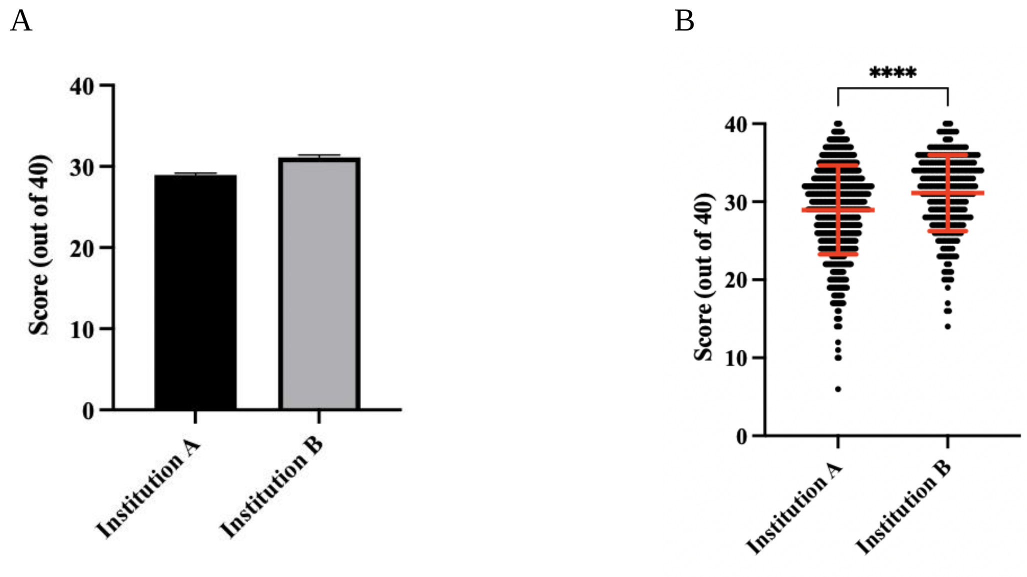


Figure 6.2.1.3 Overall performance of genuine attempts in the diagnostic test. (A) Data represent the mean scores ± SEM for each institution (Institution A, n=595; Institution B, n=290). (B) Data show the spread of scores for each institution. The horizonal red lines represent the median and interquartile range. ****p<0.0001.

6.2.2 Performance in the Calculator-allowed and NC Section

Data were sorted to allow for an evaluation of TES' performance in each section. For Institution A, TES performed best in the NC section (76.7%±0.8) compared to the Calculator-allowed section (70.9%±0.6) (Figure 6.2.2A). This trend was also observed for Institution B, albeit the difference between sections were smaller (NC = 78.7%±1.0; Calculator-allowed = 77.5%±0.7) (Figure 6.2.2B). A statistical difference was evident between the sections of the test at Institution A (Figure 6.2.2C) and Institution B (Figure 6.2.2D).

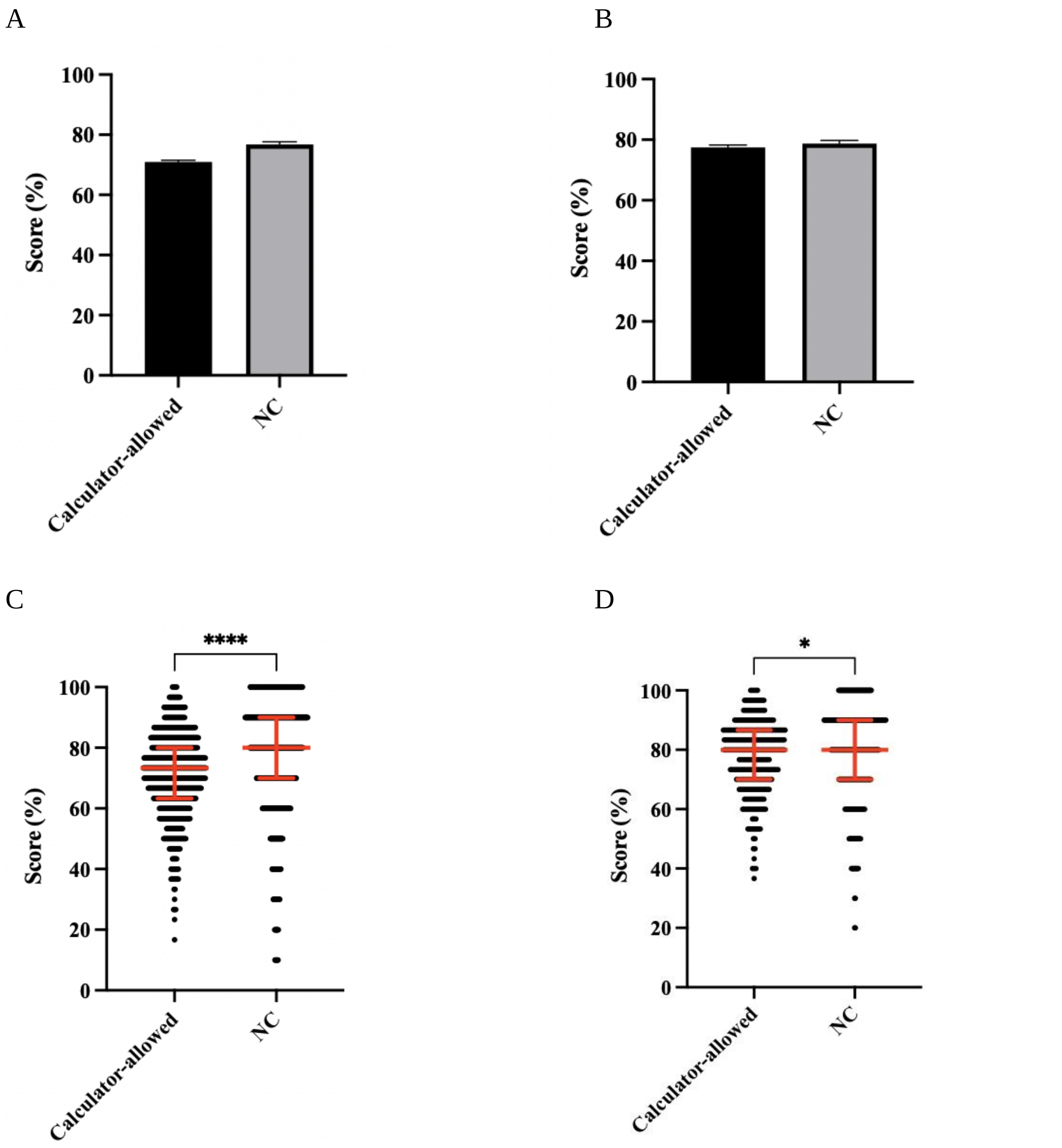


Figure 6.2.2.1 Performance in each test section. Data represent the mean score ± SEM for: (A) Institution A (n=595) and (B) Institution B (n=290). Data show the spread of scores for each strand from: (C) Institution A and (D) Institution B. The red lines represent the median and interquartile range.

6.2.2.1 Performance in the Calculator-allowed Section

Data for the Calculator-allowed section were sorted into the three strands so that TES' performance in each strand could be evaluated. For Institution A, TES performed best in NA (72.8%±0.8), followed by MG (72.2%±0.7) and then SP (67.9%±0.8) (Figure 6.2.2.1A). This trend was also observed for Institution B (NA = 78.5%±1.0; MG = 78.1%±0.9; SP = 75.9%±1.0) (Figure 6.2.2.1B). For Institution A, a statistical difference was evident between NA and SP, and MG and SP (Figure 6.2.2.1C). There was no statistical difference evident between NA and MG. For Institution B, there was no statistical difference evident between any strands (Figure 6.2.2.1D).

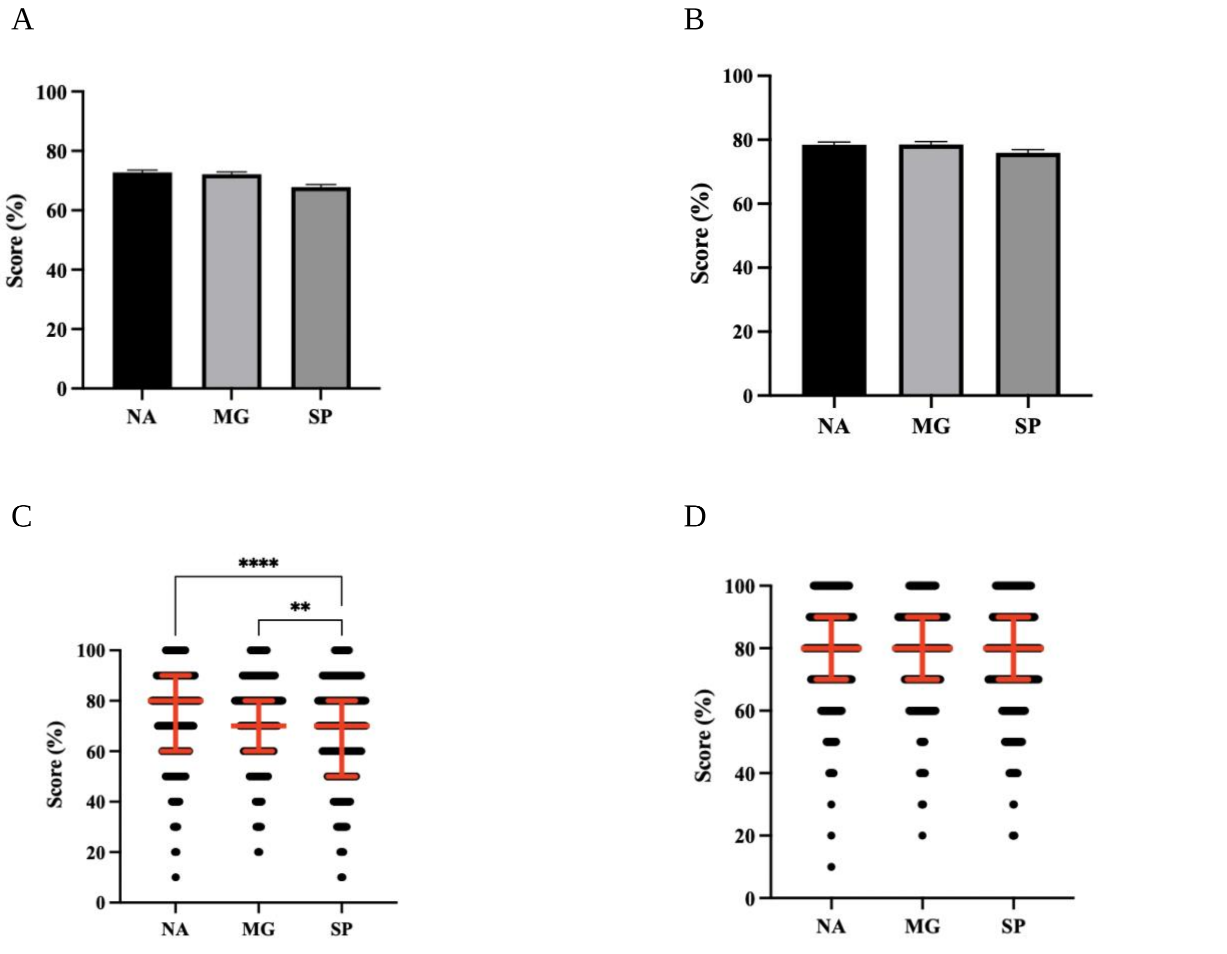


Figure 6.2.2.1.1 Performance in each strand within the calculator-allowed section. Data represent the mean score ± SEM for: (A) Institution A (n=595) and (B) Institution B (n=290). Data show the spread of scores for each strand from: (C) Institution A and (D) Institution B. The red lines represent the median and interquartile range.

6.2.2.2 Performance in the NC Section

Next, we examined TES' performance in the NC section of the test. Owing to the length of this section and the nature of NC-type questions, only NA and MG strands were assessed in this section. Since all NC questions were within one pool and that questions were randomly selected for each attempt, the number of questions for each strand presented varied between attempts. Therefore, students' performance for each strand was expressed as a percentage of the number of questions correctly answered in each strand out of the total number of questions in that strand that was presented in that attempt.

Analysis of TES' mean performance for Institution A indicated that students performed best in the NA category (77.1%±0.9, n=595) compared to the MG category (72.8%±2.0, n=448) (Figure 6.2.2.2A). In contrast, students' median score were higher in MG. This contradiction can be explained by the limited number of MG items presented on each test attempt. In fact, on some attempts, there were no MG questions. In most cases, there were only one or two MG items, which resulted in the large spread of data, with many scores either 0% or 100%. The variation in the number of MG items in this section means that there were fewer opportunities for students to demonstrate their capability. Therefore, rather than attempt to determine the best NC performed category, it is more appropriate to determine whether there was a significant difference or not. There was a statistical significance evident between the NC strands at Institution A (Figure 6.2.2.2B).

At Institution B, students' mean score indicated that they performed better in the MG strand (81.9%±2.4, n=213) compared to the NA strand (78.4%±1.0, n=290) (Figure 6.2.2.2C). Similarly, the median score at Institution B was also higher in MG (Figure 6.2.2.2D). Issues regarding the spread of MG scores was also observed for Institution B, where a large proportion of data were 0% or 100% (Figure 6.2.2.2D). There was a statistical significance evident between the strands in the NC section at Institution B.

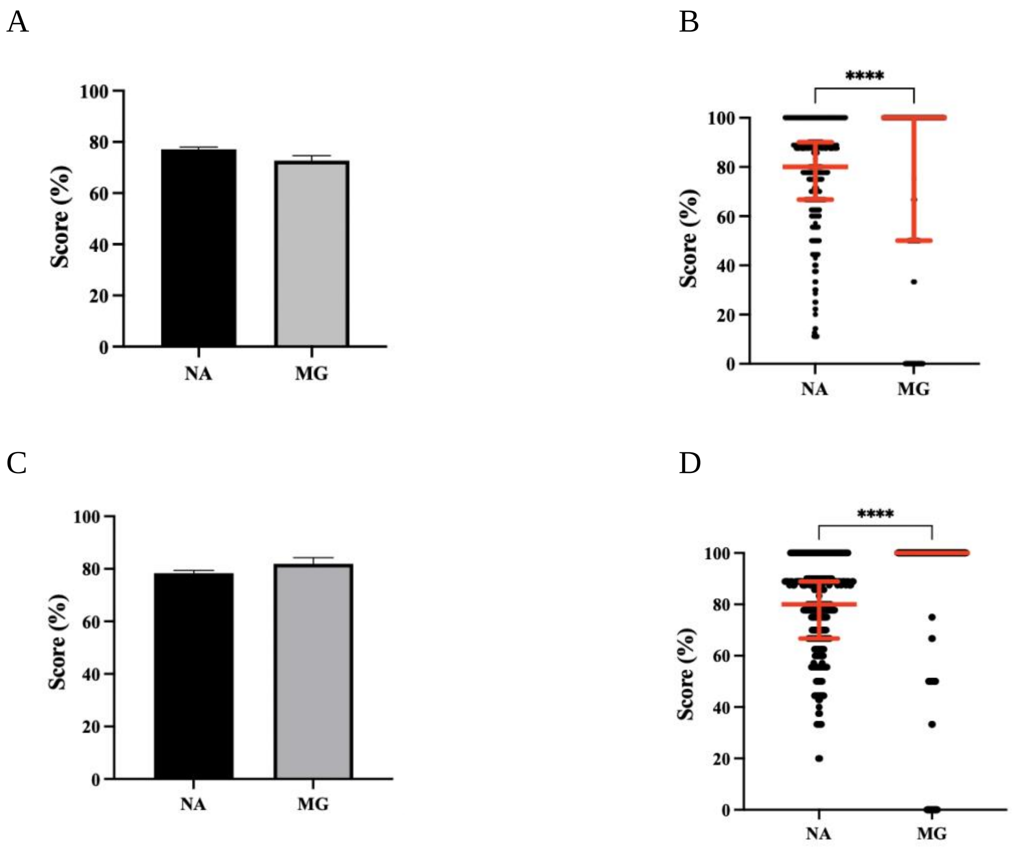


Figure 6.2.2.2.1 Performance in each strand within the NC section. (A) Data represent the mean score ± SEM and (B) spread of scores for each strand for Institution A (n=595). (C) Data represent the mean score ± SEM (C) and spread of scores for each strand (D) for Institution B (n=290). The red lines represent the median and interquartile range.

6.2.3 Performance in Content Areas

In order to determine TES' numeracy strengths and weaknesses, we analysed students' performance in specific mathematics content areas within each of the three mathematical strands. This analysis combined results from the Calculator-allowed section and the NC section.

6.2.3.1 Performance in NA Content

Data were sorted into seven content areas within the NA strand. These are: Algebra; Basic Arithmetic; Decimals; Financial Mathematics; Fractions; Percentages; and Rates and Ratios. TES' performance for each content area was expressed as a percentage of the number of questions correctly answered in each content area out of the total number of questions in that content area presented in that attempt. For Institution A, students performed best in Percentages (81.4%±1.2) and worst in the Rates and Ratios (61.6%±1.5) (Figure 6.2.3.1A). Students at Institution B also performed poorest in Rates and Ratios (68.5%±2.1); however, the content area that they performed best was in Algebra (85.9%±2.1) (Figure 6.2.3.1B).

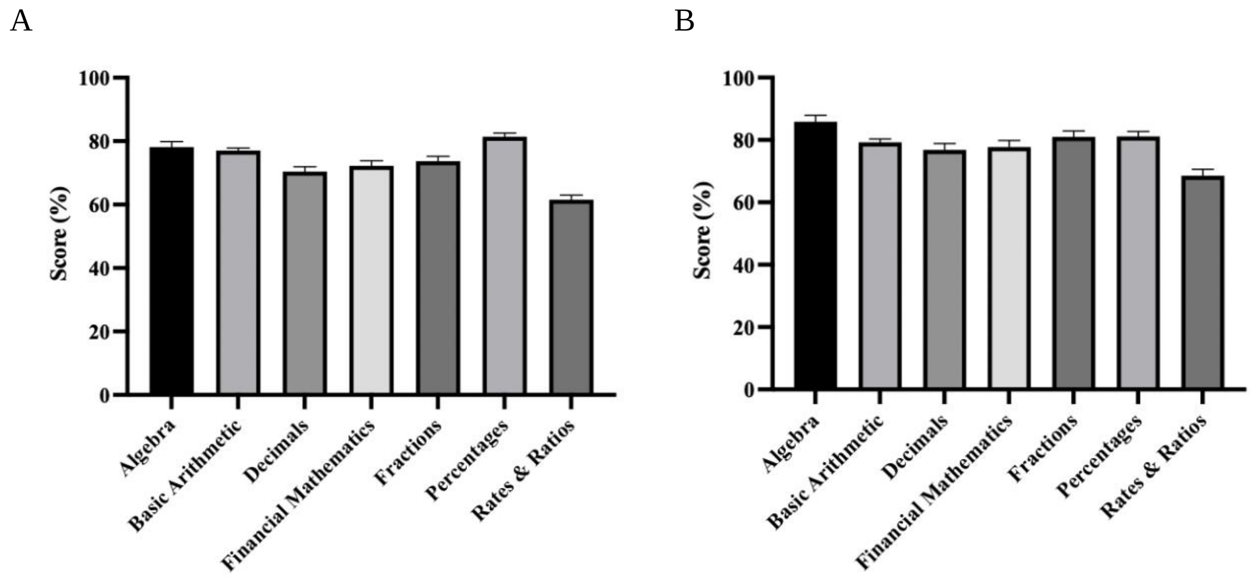


Figure 6.2.3.1.1 Performance in the NA content areas. Data represent the mean score ± SEM for: (A) Institution A (n=595) and (B) Institution B (n=290).

A statistical difference was evident between most NA content areas at Institution A (Table 6.2.3.1.1) and Institution B (Table 6.2.3.1.2). Interestingly, there was no statistical difference between Financial Mathematics and Decimals, Financial Mathematics and Fractions, and Decimals and Fractions for both institutions.

Table 6.2.3.1.1 Statistical significance between NA content areas at institution A

	Algebra	Basic Arithmetic	Decimals	Financial Mathematics	Fractions	Percentages
Algebra						
Basic Arithmetic	****					
Decimals	****	ns				
Financial Mathematics	****	**	ns			

	Algebra	Basic Arithmetic	Decimals	Financial Mathematics	Fractions	Percentages
Fractions	****	**	ns	ns		
Percentages	ns	****	****	****	*	
Rates & Ratios	****	*	****	****	****	****

Note. *p<0.05, **p<0.01, ****p<0.0001.

Table 6.2.3.1.2 Statistical significance between NA content areas at institution B

	Algebra	Basic Arithmetic	Decimals	Financial Mathematics	Fractions	Percentages
Algebra						
Basic Arithmetic	****					
Decimals	****	**				
Financial Mathematics	***	****	ns			
Fractions	**	****	ns	ns		
Percentages	****	***	ns	ns	ns	
Rates & Ratios	****	ns	**	****	****	***

Note. **p<0.01, ***p<0.001, ****p<0.0001.

6.2.3.2 Performance in MG Content

Data for the MG strand were sorted into seven content areas, namely: Angles; Area; Capacity and Volume; Distance and Perimeter; Estimating, Reading and Converting, Space, Shapes and Symmetry; and Time and Timetabling. Some similarities in performance in the MG content items were evident between the institutions. The mean scores indicated that students at both institutions performed the lowest in the Angles content area (Institution A = 52.8%±2.1, Institution B = 63.1%±2.8) (Figure 6.2.3.2A and Figure 6.2.3.2B). However, differences between the institutions were evident in the best performed content areas. The mean scores indicated that students at Institution A performed best in Estimating, Reading, and

Converting (76.9%±0.9), Time and Timetabling (77.0%±1.2), and Distance and Perimeter (73.5%±1.8) (Figure 6.2.3.2A). Students at Institution B performed best in Area (83.5%±2.2), Capacity and Volume (82.8%±2.2), and Distance and Perimeter (82.1%±2.3) (Figure 6.2.3.2B).

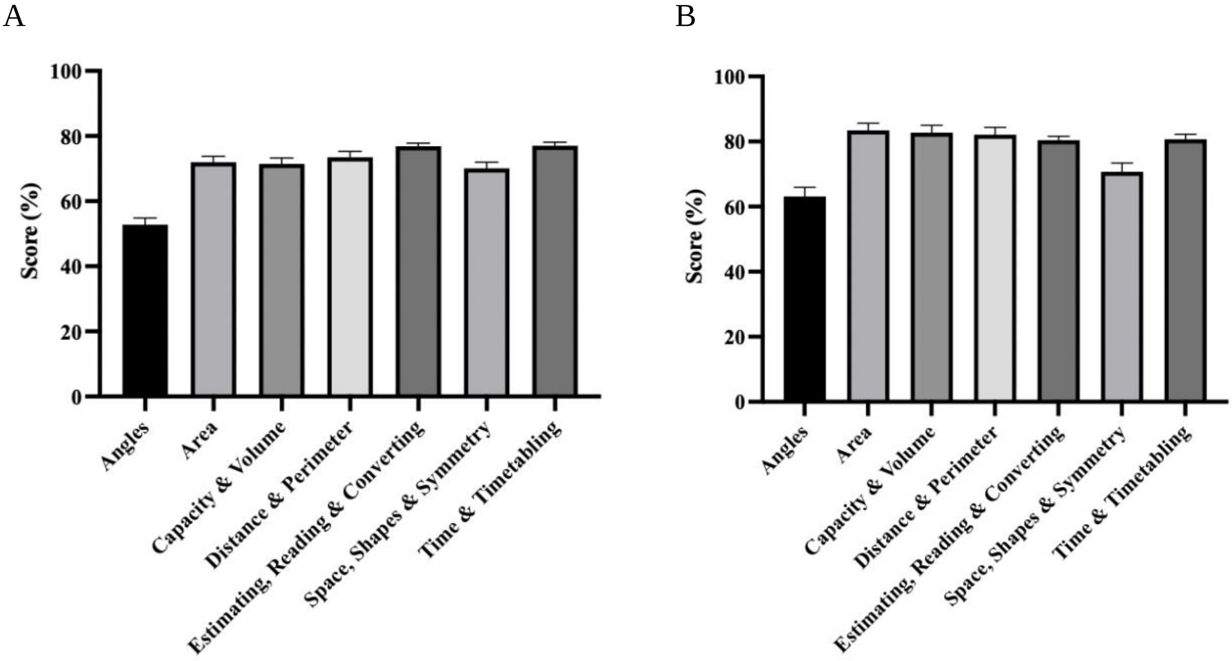


Figure 6.2.3.2.1 Performance in the MG content areas. Data represent the mean score ± SEM for: (A) Institution A (n=595) and (B) Institution B (n=290).

A statistical difference was evident between most MG content areas at Institution A (Table 6.2.3.2.1) and Institution B (Table 6.2.3.2.2). There were some similarities evident between the institutions, both displaying no statistical difference between Angles and Estimating, Reading and Converting, Area and Capacity and Volume, Area and Distance and Perimeter, Capacity and Volume and Distance and Perimeter, and Space, Shapes and Symmetry and Time and Timetabling.

Table 6.2.3.2.1 Statistical significance between MG content areas at institution A

	Angles	Area	Capacity & Volume	Distance & Perimeter	Estimating, Reading & Converting	Space, Shapes & Symmetry
Angles						
Area	****					
Capacity & Volume	****	ns				
Distance & Perimeter	****	ns	ns			
Estimating, Reading & Converting	ns	****	****	****		
Space, Shapes & Symmetry	****	ns	ns	ns	***	
Time & Timetabling	****	ns	ns	ns	*	ns

Note. *p<0.05, ***p<0.001, ****p<0.0001.

Table 6.2.3.2.2 Statistical significance between MG content areas at institution B

	Angles	Area	Capacity & Volume	Distance & Perimeter	Estimating, Reading & Converting	Space, Shapes & Symmetry
Angles						
Area	****					
Capacity &Volume	****	ns				
Distance & Perimeter	****	ns	ns			
Estimating, Reading & Converting	ns	****	****	****		
Space, Shapes & Symmetry	ns	**	**	*	**	
Time & Timetabling	ns	***	**	**	*	ns

Note. **p<0.01, ***p<0.001, ****p<0.0001.

6.2.3.3 Performance in SP Content

Data for the SP strand were sorted into four content areas, namely: Combinations; Interpreting Data; Probability; and Statistics. Similarities in performance in the SP content items were evident between the institutions. The mean scores indicated that students at both institutions performed the best in the Probability content area (Institution A = 73.8%±1.4, Institution B = 81.0%±1.8), followed by Statistics (Institution A = 71.6%±1.4, Institution B = 78.0%±1.8), Interpreting Data (Institution A = 66.9%±1.0, Institution B = 77.8%±1.3) and then Combinations (Institution A = 53.6%±2.1, Institution B = 63.1%±2.8) (Figure 6.2.3.3A and Figure 6.2.3.3B).

At Institution A, statistical significance was evident between Combinations and Probability, Combinations and Statistics, Interpreting Data and Probability, and Interpreting Data and Statistics (Table 6.2.3.3.1). However, there was no statistical difference between Combinations and Interpreting Data, and Probability and Statistics at Institution A (Table 6.2.3.3.1). At Institution B, Interpreting Data was significantly different to the other three content areas (Table 6.2.3.3.2). There was no statistical difference between other content areas (Table 6.2.3.3.2).

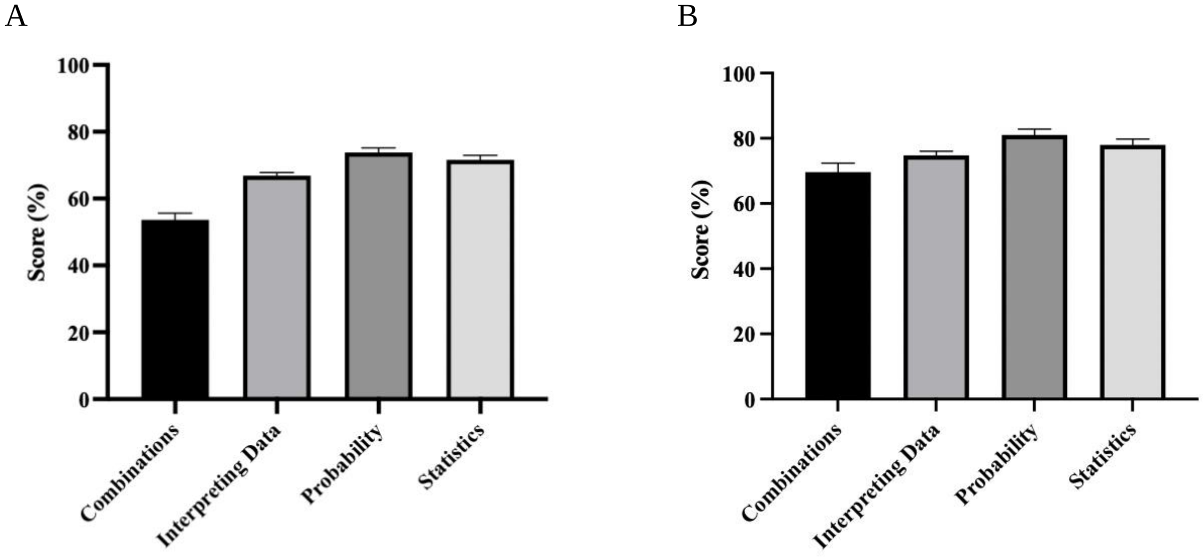


Figure 6.2.3.3.1 Performance in the SP content areas. Data represent the mean score ± SEM for: (A) Institution A (n=595) and (B) Institution B (n=290).

Table 6.2.3.3. 1 Statistical significance between SP content areas at institution A

	Combinations	Interpreting Data	Probability
Combinations			
Interpreting Data	ns		
Probability	****	****	
Statistics	****	****	ns

Note. ****p<0.0001.

Table 6.2.3.3.2 Statistical significance between SP content areas at institution B

	Combinations	Interpreting Data	Probability
Combinations			
Interpreting Data	***		
Probability	ns	****	
Statistics	ns	***	ns

Note. ***p<0.001, ****p<0.0001.

6.2.4 Performance in Item Types

The Diagnostic Test included a variety of item types to provide a comprehensive user experience and assess the different skills required to address these different item types. There were three item types used, which were: Multiple Choice (with four response options – one correct answer and three distractors); True/False; and Fill in the Blank (short response or numeric response). We therefore explored whether there were any differences in performance based on item type. For this analysis, items from all content areas in the NC and Calculator-allowed sections were included.

Amongst the three item types, students performed best in Multiple-Choice items at both institutions (Institution A = 74.9%±0.6, Institution B = 80.7%±0.8), followed by True/False items (Institution A = 70.5%±1.4, Institution B = 73.4±1.9) and then Fill-in-the-Blank items (Institution A = 66.3%±0.9, Institution B = 72.2%±1.0) (Figure 6.2.4.1A and Figure 6.2.4.1B).

At both institutions, a statistical difference was evident between Fill in the Blank and both other item types suggesting there was an obvious weakness in performance in Fill-in-the-Blank item types (Figure 6.2.4C and Figure 6.2.4D). There was no statistical difference at Institution A (Figure 6.2.4C) or Institution B (Figure 6.2.4D) between Multiple-Choice and True/False items, suggesting that performance in these items was similar.

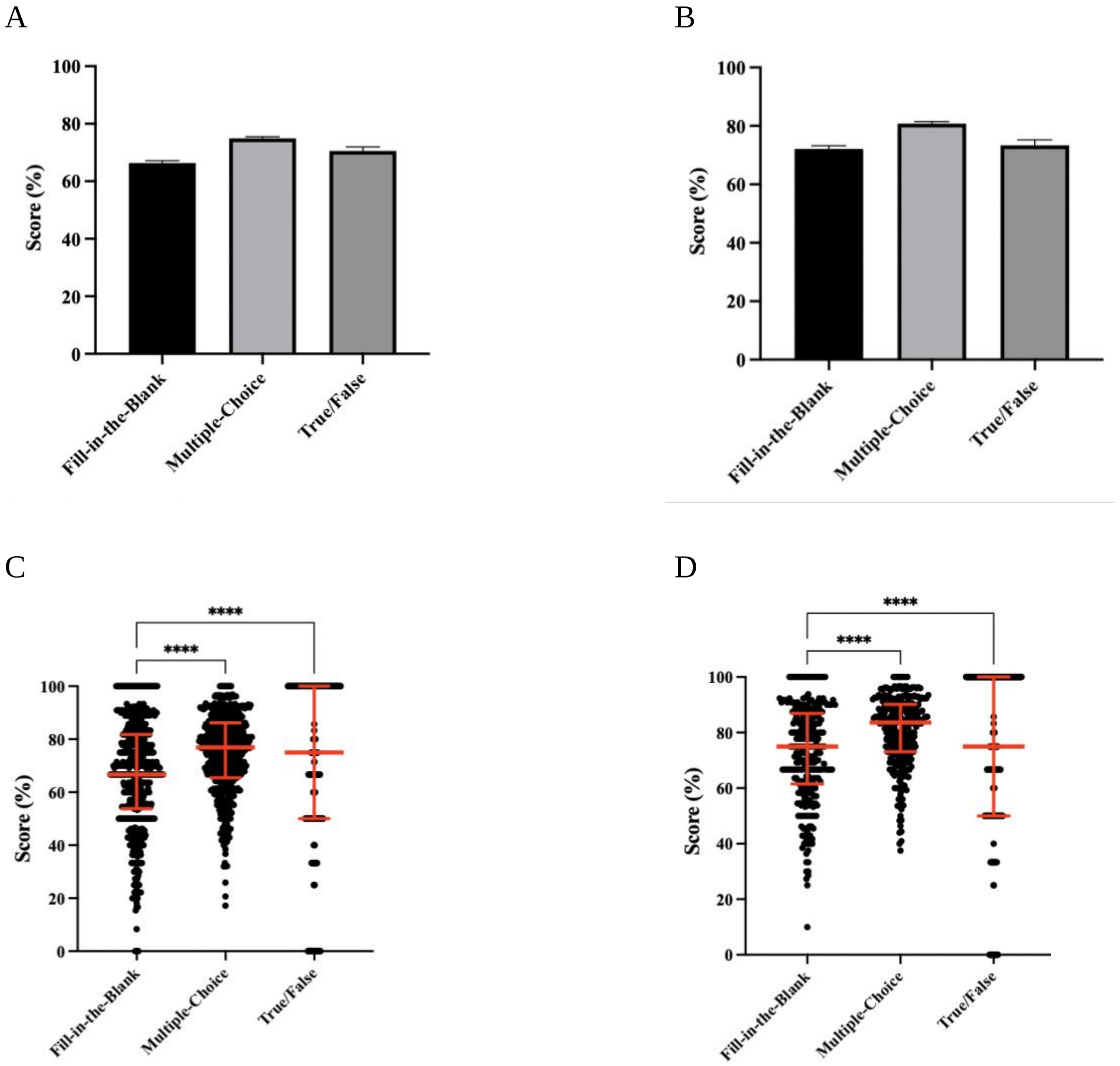
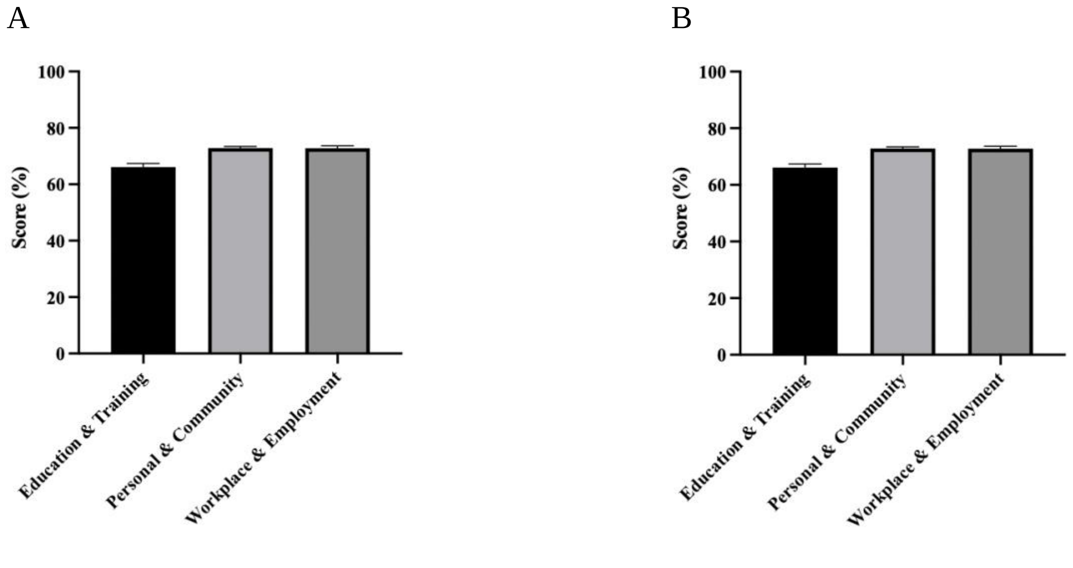


Figure 6.2.4.1 Performance based on item types. Data represent the mean score ± SEM for: (A) Institution A (n=595) and (B) Institution B (n=290). Data shows the spread of scores for each item type from: (C) Institution A and (D) Institution B. The red lines represent the median and interquartile range. *****p*<0.0001.

6.2.5 Performance in Item Contexts

In order to determine whether TES' performance varied between the contexts of the questions, all data for each of the contexts were analysed. The items were based on three contexts domains, these were: Education and Training; Personal and Community; and Workplace and Employment. For this analysis, all items from both the NC and Calculator-allowed sections were included.

The mean for both institutions showed similar variation between item contexts; however, the variation between contexts were small (Figure 6.2.5A and Figure 6.2.5B). Students at Institution A performed best in Personal and Community items (72.8%±0.6) and students at Institution B performed best in Workplace and Employment items (78.7%±1.0). For both institutions, students performed lowest in Education and Training items (Institution A = 66.1%±1.3, Institution B = 73.1%±1.8). At Institution A, a statistical difference was only evident between Education and Training and Workplace and Employment items (Figure 6.2.5C). There was no statistical difference evident between any of the context domains at Institution B (Figure 6.2.5D).



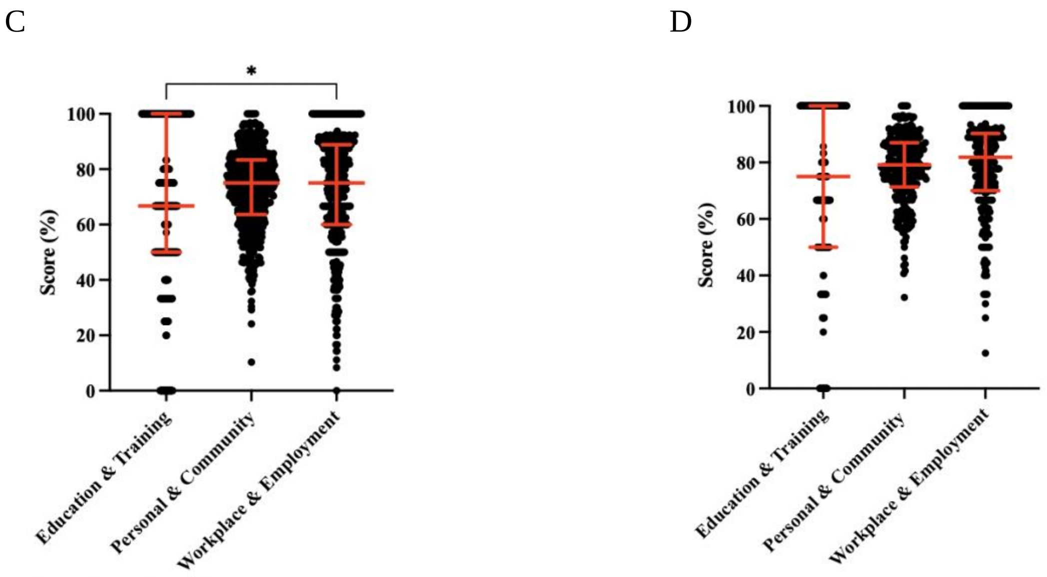


Figure 6.2.5.1 Performance based on context domains. Data represent the mean score ± SEM for: (A) Institution A (n=595) and (B) Institution B (n=290). Data shows the spread of scores for each context domain from: (C) Institution A and (D) Institution B. The red lines represent the median and interquartile range. **p*<0.05.

6.2.6 Performance in ACSF Levels

We next assessed TES’ performance based on the ACSF levels of the items. This included all items from both the NC and Calculator-allowed sections. Items in the Diagnostic Test ranged from ACSF Level 2 to Level 5. At both institutions, Level 3 and Level 4 items appeared on all attempts (Institution A, n=595; Institution B, n=290). The number of Level 2 and Level 5 items in each attempts were smaller (Table 6.2.6). Results were expressed as a percentage, calculated by the number of ACSF level questions answered correctly on each attempt out of the total number of that ACSF items in each test attempt.

Table 6.2.6.1 Attempts made for each of the ACSF levels for institution A and institution B

ACSF Level	Institution A	Institution B
2	573	273
3	595	290
4	595	290
5	504	251

The trend in mean performance based on ACSF levels were the same at both institutions (Figure 6.2.6A & 6.2.6B). Unsurprisingly, students performed best in the lowest level, which were Level 2 items (Institution A = 89.9%±0.8, Institution B = 93.5±1.0). Students' performance slightly decreased between Level 2 and Level 3 (Institution A = 81.8%±0.6, Institution B = 86.4%±0.7). A more noticeable decline in performance was observed between Level 3 and Level 4 (Institution A = 56.4%±0.9, Institution B = 64.2%±1.1). For both institutions, fewer than half the attempts at Level 5 were correct (Institution A = 34.3%±1.8, Institution B = 42.1%±2.5). Statistical significance was evident between all levels at Institution A (Figure 6.2.6C). For Institution B, statistical difference was observed between all levels, except between Level 4 and Level 5 (Figure 6.2.6D).

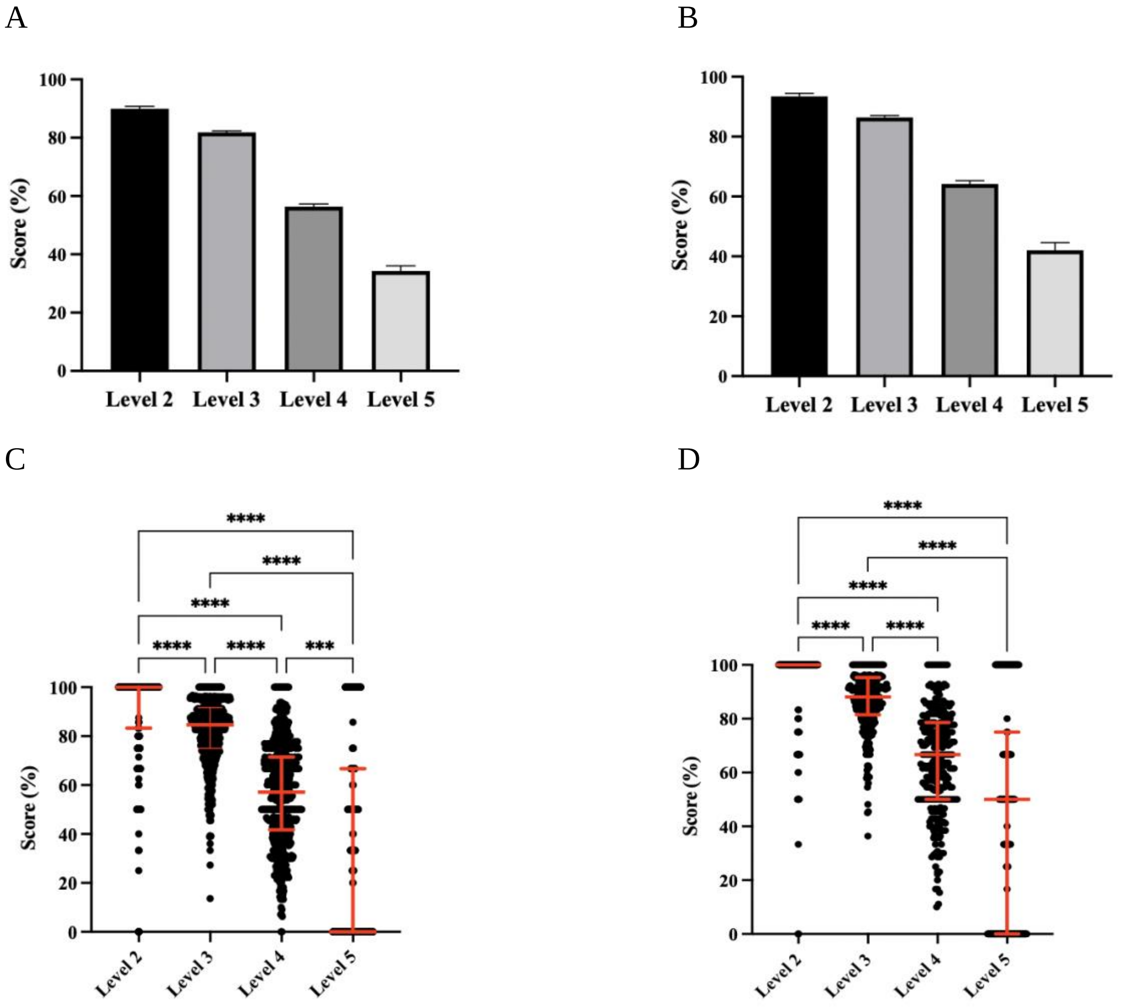


Figure 6.2.6.1 Performance based on ACSF levels. Data represent the mean score ± SEM for: (A) Institution A and (B) Institution B. Data show the spread of scores for each ACSF level

from: (C) Institution A and (D) Institution B. The red lines represent the median and interquartile range. ***$p<0.001$, ****$p<0.0001$.

6.3 Discussion

As noted earlier, the purpose of this study is to make generalisations about Australian TES' numeracy capabilities, and not to compare institutions. Given the differences in student demographics, institutional context, and the statistical difference between the cohorts' overall performance in the Diagnostic Test, we analysed the data separately and focussed on the similarities between the institutions. Findings in this chapter demonstrated that TES have similar numeracy strengths and weaknesses at both institutions, which are summarised and discussed below.

Overall, TES at both institutions performed significantly better in the NC section compared to the Calculator-allowed section (Institution A, $p<0.0001$; Institution B, $p<0.05$). TES' mean performance was 1.2-5.8% higher in the NC section compared to the Calculator-allowed section. Interestingly, other relevant research findings are in contrast to our results. For example, when examining LANTITE results data for 694 TES, Hall and Zmood (2019) found that performance was poorer on the NC questions compared to those for which a calculator was available. The authors suggested that this was evidence of a lack of TES' ability in basic computational skills. In this case, our findings indicated that TES' basic computational skills were adequate. Differences in our results and the findings of Hall and Zmood (2019) could be explained by the differences in test conditions provided. For example, Hall and Zmood (2019) suggested that the poorer performance in the LANTITE NC questions may have been affected by the timed-test conditions. Specifically, the NC questions in the LANTITE were presented at the end of the test and it is possible that TES may have run out of time before completing them, or they may have felt rushed and made mistakes. In contrast, our Diagnostic Test had no time limit, which may have provided the opportunity for TES to more accurately demonstrate their skills and, therefore may have provided a more valid indication of their NC abilities. Given that the Diagnostic Test is unsupervised, it is possible that TES could have used a calculator to check their answers in the NC section. However, as the Diagnostic Test was promoted by the researchers as a resource to practice and improve numeracy skills, it should have motivated TES to complete the NC section without a calculator. Also, this test is a non-compulsory test that is not associated with any unit of study, which would have only attracted participants eager to develop their numeracy skills. Therefore, it is unlikely that TES would choose to cheat the system.

A breakdown of the Calculator-allowed section into mathematical strands showed no noticeable difference between NA and MG results. This result aligns with Hall and Zmood's (2019) findings, which also found no noticeable differences in these two strands in the LANTITE results data they examined. However, in contrast to Hall and Zmood's (2019) findings, which showed that TES performed best in SP, we found that TES performed marginally worse in SP, compared to NA and MG. Indeed, others have also reported that TES perform poorly in SP. Specifically, in Turkey, Karatoprak et al. (2015) observed that TES performed poorly in statistical reasoning, which included understanding averages, interpreting data, and computation of probabilities. The reason for this discrepancy in when and why TES perform well in SP in certain instances and poorly in others requires further investigation.

In the Australian Mathematics (F-10) Curriculum, the largest proportion of content is evident in the NA strand, followed by MG (ACARA, n.d.). The smallest proportion of content is seen in the SP strand. Since the emphasis in mathematics is placed more on NA and MG compared to SP, one might contend that this finding is not of great concern. However, when exploring the particular numeracy demands across the curriculum, many of these demands require the application of SP strand content, such as interpreting data, constructing data displays, and performing statistical analyses. Therefore, the findings in this research present a particular concern that warrants further research. In fact, the findings suggest that TES potential to implement required SP content across the curriculum at all year levels may be limited.

Furthermore, in the LANTITE, which aims to assess TES' numeracy capabilities, the proportion of items drawn from the SP strand is 25-35%. This is approximately the same, or slightly higher, than those drawn from MG (20-30%). The remaining 40-50% are drawn from NA (ACER, 2017). Therefore, there is a high emphasis placed on the assessment of SP, which suggests that adequate capabilities are expected in this strand for TES to reach the standard required by teachers. Our findings indicate that TES may not have the capabilities to achieve the required standard overall because of their specific numeracy weaknesses in the SP strand. Further analysis of the strands within the NC section showed that although there was a statistically significant difference between NA and MG, there was conflict between the mean and median results. Further work is required to discern this difference.

When comparing TES' performance between content areas in NA, a common strength was seen in Algebra. It was the best performed content area at Institution B and the second best performed content area at Institution A. This finding supports the research of Guler and Celik (2018). Although the overall conclusion in their study were that most students performed below an accepted achievement level when assessed on algebraic content knowledge, their findings showed that performance on concepts involving algebra expressions and equations was good. Therefore, considering our Diagnostic Test specifically assessed algebra concepts relevant to numeracy applications (e.g., concepts involving algebra expressions and equations rather than inequalities and functions), our findings support their work. Noting that Guler and Celik also found a positive correlation between TES' algebra content knowledge and algebraic pedagogical content knowledge, it is likely that TES in our study also possess good algebraic pedagogical content knowledge or can readily develop this skill with little academic support.

Another strength in the NA strand was evident in Percentages. It was the best performed NA content area at Institution A and the second best performed at Institution B. The strengths evident in the Percentages content area in this research support the findings of other scholars (Fitzmaurice et al., 2021; Ngu, 2019). It is important to note the relationship between the content areas of Percentages and Financial Mathematics. In particular, Ngu (2019) noted that percentage capabilities are especially significant for understanding percentage change problems in financial related mathematics, such as interpreting and calculate pay rises, calculating price increases or decreases, and calculating GST amounts. This relationship is also confirmed in the Financial Mathematics questions in our Diagnostic Test that involved concepts as mentioned above. Therefore, considering Percentages was a strength in this study, we expected similar results to be displayed for Financial Mathematics. However, similarities in performance between these two content areas was only evident at Institution B. Interestingly, TES at Institution A performed significantly better in Percentages compared to Financial Mathematics ($p<0.0001$) with the mean performance calculated to be 9.1% higher in the Percentages content area. These results suggest that strengths in Percentages are evident overall; however, there a difference in the TES' ability to transfer these skills for Financial Mathematics applications.

A weakness in the NA strand was evident for Rates and Ratios. It was the poorest performed NA content area at both institutions. The extent of this weakness was especially evident when examining the differences between the mean scores. In particular, there was a large mean difference between Rate and Ratios and the second poorest performed NA content area at both Institutions. This difference was 8.8% at Institution A and 8.3% at Institution B. The difference between all other content areas and their next closest ranked content area ranged between 0.2-3.3% at Institution A and between 0.2-4.7% at Institution B. Unsurprisingly, performance was significantly different between Rates and Ratios and all other content areas at Institution A, and all but one content area (Basic Arithmetic) at Institution B. This observation supports the findings of Afamasaga-Fuata'i et al. (2007), who found that TES were challenged with even the most simple ratio questions as well as questions involving computing rates, such as average speeds.

The extent to which our findings demonstrated Rates and Ratios as a weakness and the findings from other research showing that simple Rates and Ratios questions are challenging for TES, present a major concern. Rates and Ratios concepts are specifically required for applications in many areas in life. For example, they are needed to accurately make comparisons, calculate quantities, determine trip lengths, and calculate speeds. Furthermore, these skills are necessary for numeracy implementation across the curriculum. For example, when using and applying scales in Design and Technology, and understanding the body's reactions to physical activity (such as breathing rates and heart rates) in Health and Physical Education. Therefore, the findings in this research present a concern about TES' personal weaknesses in Rates and Ratios, which may limited TES' ability to adequately teach these skills.

Interestingly, we did not observe Fractions to be a notable area of weakness in the NA strand, which is often cited in the research literature. In particular, studies have reported that TES struggled with fraction multiplications in different contexts (Son & Lee, 2016), have difficulties conceptualising the metalanguage of fractions (Tobias, 2013), and cannot correctly complete word problems involving fractions (Lo & Luo, 2012). However, in this research TES' Fractions capabilities were not found to be a weakness. In fact, performance on Fractions questions was ranked in the middle of all of the content areas, within NA and amongst all content areas. However, we must consider that although performance of Fractions was in the middle of all content areas, it is still possible that it is an area of weakness and that our study found areas that TES are even weaker in.

When comparing TES' performance between content areas in MG, there were differences in the best performing content areas. The best performed content areas were Estimating, Reading, and Converting, and Time and Timetabling at Institution A. Whereas Area, and Capacity and Volume were the best performing content areas at Institution B. However, a common strength was evident in Distance and Perimeter, which was the third-best performed MG content area at both institutions.

The results in the MG strand present some interesting findings that should be considered when making comparisons to the findings of other studies. In particular, we saw Distance and Perimeter as a content area strength and there was no significant difference evident between Distance and Perimeter and Area at both institutions. In fact, Area was the best performed MG content area at Institution B. Therefore, we saw strengths in both Distance and Perimeter, and Area. However, other studies have found that TES displayed misconceptions relating to the difference between Perimeter and Area (Holm, 2018; Livy et al., 2012). Therefore, it is possible that our results suggest that TES can only use these skills when they are independent from each other. Interestingly, this lack of interconnectivity has been identified in the MG strand in the Australian Curriculum by Lowrie et al. (2012). In fact, these scholars have even provided suggestions for improving the MG curriculum to provide opportunities for interconnections between the concepts of perimeter and area. Therefore, it has been acknowledged that, in general, there is a need to improve understanding of the difference between these two concepts. Although TES' ability to make connections between the Perimeter and Area of a shape was not explored in our study, our data show that when assessed independently, TES possess an adequate understanding of these content areas.

A weakness in the MG strand was evident for Angles. It was the poorest performed MG content area at both institutions. Mean performance showed that there was a difference of 17.3% between Angle and the second-poorest-performed MG content area (Space, Shapes & Symmetry) at Institution A and a mean difference of 7.6% between Angles and the second poorest performed MG content area (also Space, Shapes & Symmetry) at Institution B. This difference is considerably large in comparison to the other differences between means observed that on average were 1.4% at Institution A, and 2.5% at Institution B. The weakness observed in Angles is consistent with existing literature. For example, in Turkey, Yigit (2014) found that TES had limited knowledge of Angles when exploring their mental constructions of the concepts of angles and angle measurement through interviews. Specific weaknesses identified in Yigit's study were that TES were not able to identify an angle from a straight line because they expressed that they either needed two lines or a vertex, and they were not able to identify angles in circles. This finding is in line with our study's results because the angle questions in the Diagnostic Test involved the calculation of angles where rays were either not visible, or not obvious in the image presented. Furthermore, when discussing strategies for improving TES Angles skills, Yigit suggested using the technology tool GeoGebra and discussed benefits of using the tool such as helping to determine differences, effects and angle properties of difference objects. Considering the similarities between our findings and those of Yigit, the use of technology that provide hands-on and visual support should be considered to improve angles skills in Australian TES.

Another weakness in the MG content area was evident for Space, Shapes and Symmetry. It was the second poorest performed MG content area at both institutions. This particular area of weakness is also evident in other studies who have explored mathematical capabilities involving shapes. For example, Fujita and Jones (2006) found that overall TES capabilities were low with regards to defining and classifying quadrilaterals (e.g., parallelograms, squares, rectangles, & trapeziums). In fact, the authors described the results from this study as disappointing because although most TES were able to draw the quadrilaterals, they displayed a lack in ability to define them. Similarly, Şahin and Başgül (2020) reported that TES' pedagogical content knowledge on quadrilaterals were not at the required standard. In their study, Şahin and Başgül observed that although many TES were able to identify the errors, very few were able to explain the mistakes or provide any solution recommendations.

In line with the specific weaknesses we found in the MG content areas (Angles & Space, Shapes & Symmetry), Ozdemir and Goktepe Yildiz (2015) found that TES only have a superficial understanding of spatial reasoning skills. It was asserted that TES were only able to evaluate independent situations and not able to combine their information within a consistent structure (Ozdemir & Goktepe Yildiz, 2015). This weakness is of particular concern, especially because spatial reasoning is outlined in the Australian Curriculum as one of the numeracy elements to be implemented across the curriculum (ACARA, n.d.). A knowledge weakness in this content area suggests that TES have limited potential to adequately implement these skills into their teaching is quite limited.

The need to improve spatial reasoning skills has also been identified by Lowrie et al. (2012). In their examination of the Australian Curriculum, the authors identified a lack of spatial reasoning skills within the MG strand. This is despite the inclusion of spatial reasoning as one of the numeracy elements as part of the general capabilities to be implemented across the curriculum. Furthermore, Lowrie et al. (2012) highlighted the demand of these skills in the technology-driven world that now exists. Therefore, our findings are support Lowrie and colleagues' (2012) assertion that spatial reasoning as an area requiring general improvements in Australia.

When comparing TES' performance between content areas in SP, a common strength was seen with Probability. At both institutions, TES performed best in the Probability content area. This finding is in contrast with results from other studies that suggested a weakness in TES' probability skills (Karatoprak et al., 2015). Upon thorough investigation, we noted some Probability questions in our Diagnostic Test specifically required applications of Percentages, which was found to be a strength in our study. Furthermore, several of our Probability questions specifically required applications of Fractions. Although Fractions skills were not found to be a specific skill strength in our study, they were also not a weakness of the participants in our study. Therefore, considering we saw that skills of Percentages and Fractions were adequate for our TES sample, it is not surprising that Probability skills were found to be adequate.

In the SP strand, Combinations was the poorest performing content area at both institutions. Our finding is in line with a study by Forgasz and Hall (2019), which reported that TES' performance on Combination questions displayed the lowest rate of accuracy compared to other questions. When TES were asked to comment on the questions, students' responses were "Literally I have no idea", "Lots", and "Not sure" (Forgasz & Hall) . Likewise, in our study, we saw responses that displayed a lack of skills and understanding in this area. For example, where a response of 8 was the correct response, we saw responses of 216, 40, and 36. In another question requiring students to determine how many ways three people could be arranged in a line, we saw a response of 729. These responses in our study highlight the extent to which Combinations is a weakness and support the findings made by Forgasz and Hall. .

Interestingly, in conclusion to the lowest scores seen in Combinations in the study of Forgasz and Hall (2019), they highlighted a possible item-type explanation for the poor performance. More specifically, in their study, the Combinations question was the only question requiring an open-ended Fill-in-the-Blank response and this was a possible reason provided by the researchers explaining the poor performance. When we explored performance in the different item types, we also saw that Fill-in-the-Blank was a weakness. It was the poorest performed item type at both institutions. This may be due to the lower probability of correctly guessing a Fill-in-the-Blank items or the fact that this item type is believed to require more thinking in mathematics tests and are generally harder (Abida et al., 2011). In general, studies have found that Fill-in-the-Blank items are the most poorly performed of the item types in mathematics assessments, especially for male participants (Lindberg et al., 2010; Lui & Wilson, 2009; Taylor & Lee, 2012).

Multiple-Choice was the best performed item type, closely followed by True/False at both institutions. There was no significant difference between performances on these item types at Institution A or Institution B. Considering the element of chance that is involved with both of these item types, correctly guessing may have contributed to this result. However, the Multiple-Choice items had four response options compared to the True/False items with only two response options. Therefore, if chance was the main reason for this result, we would have expected best performance to be for True/False items, which has a higher theoretical probability of correctly guessing the answer. Therefore, we do not believe the element of chance greatly influenced the results. In fact, we believe that our Multiple-Choice items and True/False items accurately assessed the same traits as the Fill-in-the-blank items, supporting the research of Abida et al. (2011), who found that the inclusion of both Multiple-Choice items and Fill-in-the-Blank items verify a test as a valid and reliable assessment.

It is important to acknowledge that challenges exist in the development of test items in the different item formats. For example, research has suggested that maximum expertise is needed to construct short response Fill-in-the-Blank items and Multiple-Choice items require the inclusion of incorrect response options that are realistic distractors (Abida et al., 2011).

Thus, time and effort are required to develop solutions that addresses common misconceptions. Therefore, overall it is clear that both expertise and careful consideration is necessary in the development of all item type formats to ensure that the assessment is valid and reliable.

These challenges are outweighed by the many benefits that exist for the inclusion of the different item type formats which allowed us to accurately assess TES' numeracy knowledge. For example, it is believed that Fill-in-the-Blank item types allow for the assessment of procedural knowledge and problem solving, and Multiple-Choice items have been determined best to assess content (Abida et al., 2011) and enhance learning (Butler, 2018). Furthermore, in a study that specifically explored the most suitable number of response options for Multiple-Choice items and have concluded that only 2-3 response options need to be developed to sufficiently assess knowledge (Haladyna & Downing, 1993). Our findings showed no significant difference between performance on Multiple-Choice items and True/False items, which supports this notion.

However, the specific inclusion of Multiple-Choice items with 4 responses options was beneficial in this study. Through the development of distractors as the responses, which included potential misconceptions, the most common misconceptions were able to be identified. As an example, for Fractions, in the question: *Felicity has 24 green pencils and 16 pink pencils. What fraction of Felicity's pencils are pink?*, we found that many attempts indicated an incorrect response (121 out of 278). Of those incorrect responses, most (93 out of 121) were found to choose the response of one-third. This indicated that TES calculated the fraction as 16 (pink pencils) out of 24 (green pencils), instead of calculating the total number of pencils to be 40 and then calculating the fraction as 16 out of 40. Therefore, the distractors that were developed allowed us to determine the most common misconceptions of TES. It is important to note that the specific misconception discussed was identified in Fractions, despite our findings that indicated Fractions was not a weakness. However, because our analysis included all attempts on the test (e.g., first attempt, second attempt etc.), it is possible that improvements in this area may have been made over time. This will be explored further in Chapter Seven.

We also assessed performance in three context domains. This aligned with the context domains assessed in the LANTITE (ACER, 2017) and also supported the domains of communication outlined in the ACSF Framework (Department of Employment, 2015) as the broad contexts in which core skills are used. We aimed to determine if the contexts of the questions in these domains affected performance. Overall, we found that Education & Training showed the lowest mean result at both institutions. However, the only statistical difference between the context domains was seen between Education & Training, and Workplace & Employment at Institution A. This suggested that overall, performance in the item contexts was fairly similar with no obvious strength or weakness. Limited literature exists on the effect of specific contexts on numeracy performance; however, these results align with other studies that suggest that story contexts have no effect on performance in certain numeracy areas (Vappula & Clausen-May, 2006). Although our work aligns with these findings of Vappula and Clausen-May (2006), there are noticeable differences between our study and theirs which are important to note. For example, their study explored Year 6-9 students which suggested a limitation for making comparisons between the two studies. Furthermore, Vappula and Clausen-May (2006) considered graphical images and diagrams as a context, and they explored their impacts on numeracy performance. In particular, they found that diagrams improved performance in certain numeracy areas, such as subtraction involving fractions. As outlined previously, we determined the contexts in domains of the broad contexts core skills may be used, however it is possible that contexts could have been distinguished differently, for example, in terms of stories or images. If this was the case, it is reasonable to expect different findings regarding the effects of contexts on performance. This would be interesting to explore further in future studies.

We developed items according to the ACSF levels and we analysed results to determine the levels of numeracy performance of TES. The ACSF levels range from Level 1 to Level 5, although and we did not developed items at Level 1 considering the simplistic nature of this skills outlined at this level. The Level 1 exclusion also aligned with the LANTITE Assessment Framework (ACER, 2017). Therefore, questions were developed between Level 2 and Level 5 and the capabilities provided in the descriptions, as the level indicators, were considered in the development. The indicators described exit performance at each level which allowed us to determine the general numeracy performance of TES. Our analysis found that nearly all attempts were successful at Level 2 and Level 3, considering the high mean performance on these levels at both institutions. Based on the ACSF level descriptors, this suggests that in general, TES have mathematical skills that involve identifying, selecting and interpreting mathematical information, selecting and using appropriate problem solving strategies, and using mathematical language and representation to communicate mathematically. Our results also showed that approximately half of the attempts achieved Level 4 and approximately one-third achieved Level 5. These results suggested that TES' mathematical skills involving evaluation, analysis, synthesis are limited and their ability to use a wide range of highly developed problem solving techniques was limited.

It is difficult to determine the ACSF level required for TES. However, the ACSF level descriptions suggest that Level 5 involves more specialised mathematical skills that would not be required by all TES. Furthermore, the LANTITE assesses content mostly at level 3 (30-40%) and level 4 (40-50%). Therefore, we considered the required numeracy standard of TES to be in the range of Level 3 and Level 4. Although our findings showed that most attempts were successful at Level 3, we observed that only about a half were successful at Level 4. Therefore, our results suggest that many TES are not achieving numeracy standards at the required upper level. These findings are in line with other research in Australia that has also found that many TES are not reaching mastery level in numeracy (Sellings et al., 2018). Similarly, our findings also support international research that has suggested TES are not meeting mastery level (Afamasaga-Fuata'i et al., 2007) and are not meeting professional numeracy standards (Linsell & Anakin, 2012).

Overall, our findings suggest that TES may not be achieving the required numeracy standard, and we identified specific numeracy strengths and weaknesses in our results as well as impacts on performance due to certain test conditions (item type, context domain, and item difficulty). It is expected that knowledge from our findings can be used and targeted interventions can be put in place to improve identified areas of weakness; therefore, improving overall numeracy capabilities to ensure TES meet the required numeracy standards.

Chapter Seven: Applying the Rasch Measurement Model

7.1 Introduction

In the previous chapter, raw scores were analysed to determine general numeracy strengths and weaknesses. We analysed the institutions separately, and our findings displayed trends of capabilities. The focus of this chapter was to extend upon the findings developed in Chapter Six by providing a further evaluation of TES' numeracy capabilities. However, the main purpose of this chapter was to provide an assessment performance between normalised attempts to allow for accurate comparisons between one test attempt and another. This included exploring changes in performance between first and final attempts in mathematical strands, content areas, item types, context domains, and ACSF Levels.

Raw score analysis is common practice in quantitative research. However, in this study, it would suggest that the results of our Diagnostic Test produced scores on a scale of equal units 0-40. Using the raw scores in this way assumes a linear scale where raw scores are simply added up to compare levels of achievement. However, this method does not consider that it is likely that all items are not of equal difficulty. In fact, the items in this test are deliberately developed not to be the same difficulty because they were developed among ACSF Levels 2-5. Therefore, in our study, a sum of raw scores could not be used to achieve reliable comparisons of TES' progress. Further, each attempt only included 40 out of 272 possible items. Therefore, the challenge of comparing results of attempts with different items was acknowledged, and a more meaningful way of measuring ability to make comparisons was necessary for this element of our research.

The Rasch Measurement Model is a mathematical model of measurement used to overcome this issue (Andrich & Marais, 2019). This Rasch Model provides a way to convert raw scores into linear measures known as logits (Bond & Fox, 2015; Iramaneerat et al., 2008). Thus, Rasch analysis allows raw test scores to be expressed in terms of performance on a linear scale that accounts for the unequal difficulties across all test items. Using the Rasch Model allows for an estimate of the person's ability based on the items they attempted by making comparisons to other persons and items (Bond & Fox, 2015).

7.2 Results

7.2.1 Attempt and Item Overview

Applying the first calibration of Rasch analysis produced an initial overview of attempts (Table 7.2.1.1) and items (Table 7.2.1.2). Estimates of person (attempt) ability and item difficulty were produced, together with their respective error estimates, and reported as raw scores and logit measures. Comparisons of raw scores with logit measures were made and measures of central tendency and spread were calculated. The mean attempt measure was 1.50 logits, and the range was 7.91 logits. For item measures, the mean item estimate is theoretically set to zero logits under the Rasch Model. Therefore, item difficulties are appropriately placed above or below the mean measure, and the spread, including the minimum and maximum item measures, of item difficulties is of most significance. The range of the item measures was 6.94 logits.

Table 7.2.1.1 Summary of 885 attempts

	Score	Count	Measure	SE
Mean	29.7	40	1.50	0.46
Max.	40	40	5.87	1.86
Min.	6	40	-2.04	0.34

Note. Overview of attempts with mean raw score, mean logit measure with standard error, and minimum and maximum attempt scores and measures.

Table 7.2.1.2 Summary of 272 items

	Score	Count	Measure	SE
Mean	96.5	130.1	0.00	0.26
Max.	163	194	3.52	0.72
Min.	23	74	-3.42	0.17

Note. Overview of items with mean raw score, mean count, mean logit measure (set at zero), minimum and maximum item scores, counts, and logit measures.

The next stage of the attempt and item overview consisted of comparing all items to their allocated ACSF levels. A visual representation was produced to observe the correspondence between actual item difficulty versus the orders predicted by the allocated ACSF levels (Figure 7.2.1). Overall, it was observed that the items increased in item difficulty according to their corresponding ACSF level. The mean logit measure was -1.43 for Level 2, -0.52 for Level 3, 1.06 for Level 4, and 2.13 for Level 5.

However, there were some overlaps of item measures between the ACSF levels identified. That is, some items were observed to have produced logit measures out of the range of what was expected for their allocated ACSF level. For example, there were 2 out of 19 Level 2 items observed to be above the Level 3 mean logit measure. At Level 3, there were 2 out of 157 items observed to be above the Level 4 mean logit measure and 13 items observed to be below the Level 2 mean. At Level 4, there were 2 out of 86 items observed to be above the Level 5 mean logit measure and no items observed to be below the Level 3 mean logit measure. At Level 5, 1 out of 10 items was observed to be below the Level 4 mean logit measure. Overall, there were 7% of items observed to be out of the acceptable range of logit measures according to their allocated ACSF level (20 out of 272). Overall, this demonstrates the validity of the Diagnostic Test to distinguish performance.

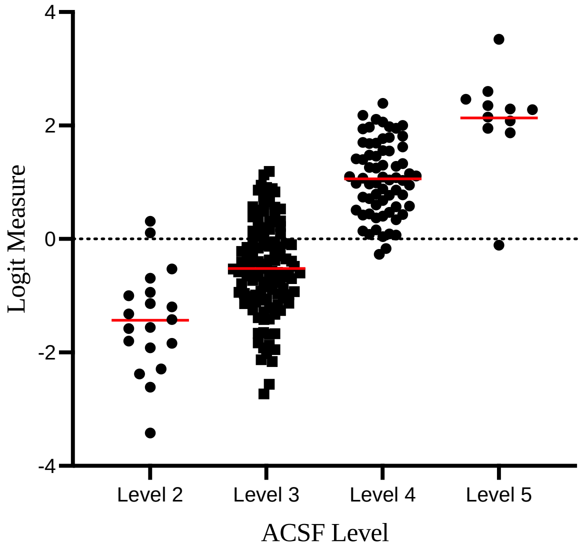


Figure 7.2.1.1 Item logit measures versus ACSF levels. The figure shows the items in ACSF Levels and the spread of their corresponding logit measures. The red lines represent the mean logit measures at each level.

In this study, Rasch measures calculated item difficulties from TES' responses to items; therefore, it must be considered that these unexpected responses may not be an accurate indication of person ability and therefore may have affected the calculations of item difficulty. Further analysis aided in determining whether these attempts should be deleted to calculate more accurate item measures.

7.2.2 Anchoring Item Measures

The attempts listed in the entire ordered Data Matrix that was produced are all attempts made on the test; however, many of these attempts are subsequent attempts made by individual persons. Further, it is reasonable to assume changes in ability measures of individual persons as they make subsequent attempts on the test. This may be due to cognitive development, knowledge and/or skill improvement due to interventions, or other reasons. It also must be considered that raw scores may even decline, especially if students are presented with more difficult items on subsequent tests; however, this does not necessarily mean that ability has declined. This research intended to track the development of students and make valid comparisons between attempts, so it was crucial to be measuring each of these attempts with one scale. Therefore, building a 'ruler' (Bond & Fox, 2015) for measurement was necessary so that all items could be marked on that ruler and all attempts could be measured against that same ruler.

As outlined in the section above, if the item difficulty estimates generated from all of the raw data were used, some things may not be taken into consideration. For example, some items may be considered easy because low-ability TES have accurately guessed them and other items may be regarded difficult because of an issue with the question, and as such may not be an accurate reflection of their abilities. Therefore, it was necessary for these unexpected item responses to be identified, explained, and considered for elimination once thoroughly examined to appropriately set item measure anchors (Bond & Fox, 2015).

7.2.2.1 Rasch Model Fit

Therefore, the next stage consisted of a further investigation into all attempts and items to ensure fit to the Rasch Model. Analysing all of the data using Rasch analysis, interpreting the results, and making necessary changes to the data were required to ensure that the data satisfactorily fit the Rasch Model to make valid observations of measurement.

The ordered Data Matrix revealed some properties about the observations that guided further data analysis for Rasch Model fit. For example, there were no items with all zeros recorded, meaning that no item was answered incorrectly on all attempts. Similarly, items did not discriminate between the observed attempts (i.e., all ones recorded, meaning every attempt on that item was correct). Therefore, for further analysis in this research, no items from this observation of the ordered Data Matrix were required to be eliminated. However, further calculations and interpretations were necessary, considering the unexpected responses that were observed.

7.2.2.1.1 Item Pathway

Firstly, to visually observe the logit measure, error estimates and fit statistics (infit) of the items in the test, an Item Pathway was developed (Figure 7.2.2.1.1). As defined by Bond and Fox (2015), Item Pathways represent the developmental acquisition of cognitive reasoning ability. A number of ideas were observed immediately from the Item Pathway (Figure 7.2.2.1.1); for example, item difficulties span almost eight complete units on the logit scale. Item 187 is the most difficult item, Item 265 is the easiest item, and many items sit near the midpoint (0 logits) on the vertical item difficulty scale. It was also observed that the easy items have the least precision (largest SE), whereas the error estimates for most of the other items were comparatively quite small.

The Item Pathway (Figure 7.2.2.1.1) also displayed the infit statistic of the items. When observing the pathway between the vertical lines at -2 and 2 (infit), it was revealed that the fit of the test to the Rasch Model’s expectation was quite good. However, locations for Items 196, 199, 217, 19, 40, 262, 10, 168, 216, 78, 17 did not seem to fit the same developmental pathways as well as the remaining items. Thus, these items were candidates for consideration for further analysis.

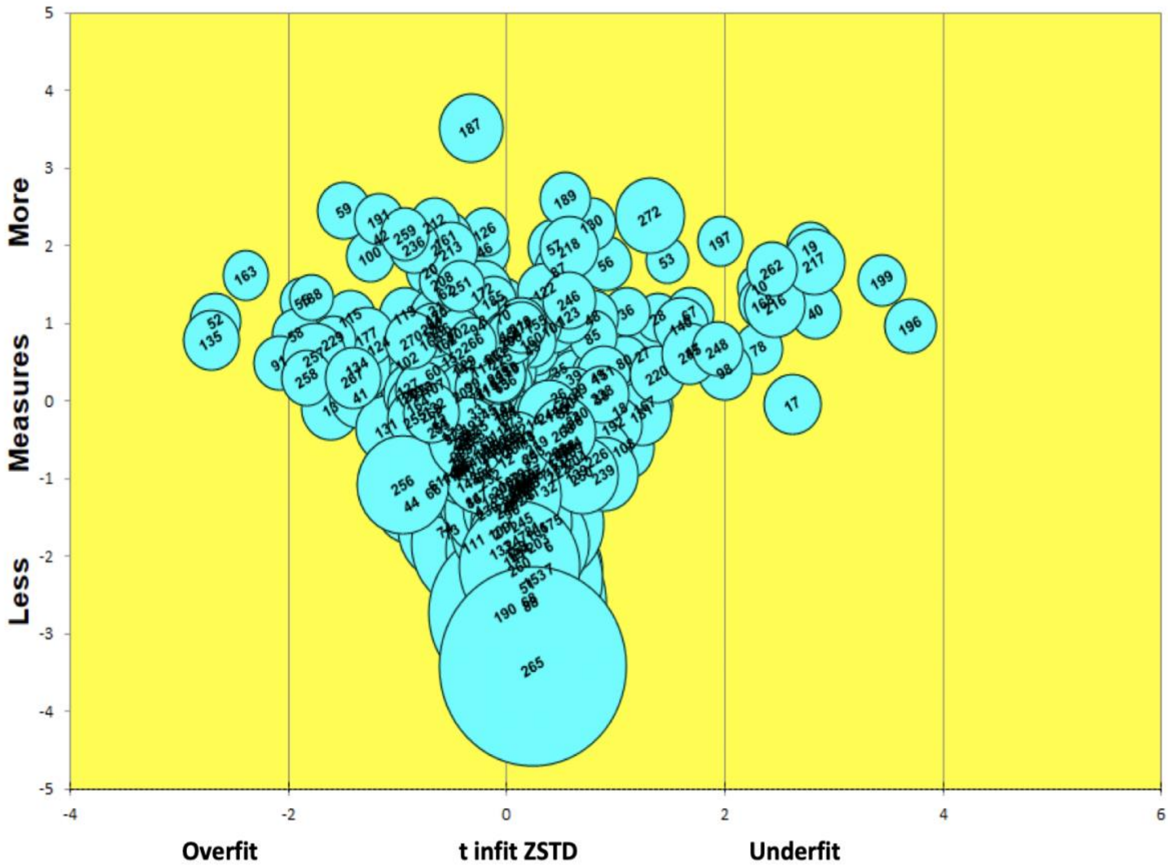


Figure 7.2.2.1.1.1 Item pathway. Easy items are located at the bottom of the map, and difficult items are located at the top. The precision of the item estimates is indicated by the vertical size of each item marker (larger markers indicate larger standard errors of measurement (SE)). Items that fit the Rasch Model are located on the pathway between the vertical parallel lines of -2 and 2.

7.2.2.1.2 Mis-fitting Items and Attempts

Further analysis was then conducted to identify and examine mis-fitting items that included both fit statistics (infit and outfit) calculations. Table 7.2.2.1.2.1 displays the item statistics of the most mis-fitting items. The score is the total number of correct responses provided on that item; the count is the total number of attempts on that item. For example, Item 174 had 100 correct responses out of 111. The measure is the logit score of that item, indicating item difficulty, and the SE is the standard error of measurement. The lower the SE, the higher the confidence in that measure. Firstly, it is observed that all items, except item 98, were easier than the average and also have low item-measure correlation. Secondly, an issue was evident with Item 174 displaying a very low item-measure correlation of -0.3. This indicated that it is not working well with the other items. All other items were considered to be acceptable in regards to the item-measure correlation; however, many items were displaying misfit for other reasons to be explored.

Table 7.2.2.1.2.1 Item statistics of most misfitting items

					Infit		Outfit		
Item	**Score**	**Count**	**Measure**	**SE**	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**	**Corr.**
174	100	111	-0.87	0.33	1.09	0.46	3.32	3.93	-0.03
239	145	162	-0.94	0.27	2.16	0.89	2.32	2.82	0.05
6	151	159	-1.87	0.37	1.09	0.39	2.26	1.95	0.04
34	121	149	-0.23	0.22	1.08	1.85	1.95	3.47	0.18
108	133	158	-0.60	0.23	1.16	1.08	1.85	2.64	0.12
8	134	147	-1.20	0.30	1.03	0.22	1.81	1.82	0.12
141	139	146	-2.02	0.40	0.99	0.08	1.77	1.28	0.17
17	103	133	-0.04	0.22	1.33	2.62	1.69	2.76	0.03
144	79	94	-0.58	0.30	1.10	0.59	1.63	1.73	0.14
98	99	136	0.39	0.21	1.21	2.00	1.54	1.53	0.31
234	122	138	-0.84	0.28	0.92	-0.37	1.54	2.55	0.17
241	137	150	-1.22	0.30	1.03	0.20	1.53	1.33	0.14

Note. Items listed in misfit order with Item, Score, Count, Measure, Standard Error, Infit and Outfit (Mean Square Residuals and Standardised z) and Partial Correlation.

In an attempt to diagnose what had gone wrong, further investigation was necessary to determine where the misfitting item responses came from, and what were the reasons for these occurrences. For example, did high ability persons provide incorrect responses to easy items?; or did low ability persons provide correct responses to difficult items? For Items 239, 6, 34, 8, 141, 234, 241, infit mean squares and infit Z were acceptable (refer to Chapter 4 for more details), indicating that the persons/attempts targeted by these items fit the Rasch Model, but the fit of persons outlying from these items is poor. This is indicated by the erratic nature of the outfit indicators. Observing the ordered Data Matrix again, Items 174, 239, and 6 are easy items; therefore the erratic indicators are likely caused by unlikely incorrect responses of some of the most able TES. For the mid-level Items; 34, 108, 17, 144, 98, the indicators are likely caused because some of the most high-ability TES responded incorrectly to these items, while some low-ability TES answered them correctly.

Considering the goal of ensuring that the data fit the demanding requirements of the Rasch Model before building the ruler and anchoring (Bond & Fox, 2015) items, further investigations were necessary. In particular, considering that unexpected responses in some attempts affected the item indicators, it was necessary to next investigate misfitting attempts, possibly eliminate them, and then reinvestigate item indicators once a second Rasch calibration had occurred. This analysis was used to determine whether removing misfitting attempts would improve the items' functioning to fit the Rasch Model's requirements better.

The attempt statistics of the most mis-fitting attempts are shown in Table 7.2.2.1.2.2 The score is the total number of correct responses provided on that attempt; the count is the total number of items attempted (i.e. 40). The measure is the logit score of that attempt (indicating person ability on that attempt), and the SE is the standard error of measurement. The lower the SE, the higher our confidence is in that measure. There were 62 misfitting attempts with ≥1.5 outfit mean square.

Table 7.2.2.1.2.2 Attempt statistics

					Infit		Outfit		
Attempt	**Total**	**Count**	**Measure**	**SE**	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**	**Corr.**
819	37	40	3.30	0.65	1.01	0.18	5.19	2.44	0.09
867	36	40	2.91	0.57	1.15	0.52	4.68	2.61	0.07
824	39	40	4.21	1.00	1.10	0.41	4.30	1.80	-0.12
586	39	40	4.20	1.03	1.13	0.44	4.03	1.73	-0.11
371	37	40	2.82	1.03	1.02	0.2	3.88	2.19	0.08
547	36	40	3.12	0.62	1.30	0.87	3.88	2.11	-0.07
223	33	40	1.87	0.56	0.96	-0.06	3.68	3.12	0.27
355	39	40	4.64	0.46	1.19	0.49	3.27	1.50	-0.06
684	39	40	4.01	1.05	1.14	0.45	3.20	1.48	-0.06
601	38	40	3.54	1.04	1.20	0.51	3.09	1.59	-0.11
273	37	40	3.10	0.75	1.11	0.39	2.84	1.64	0.06
364	33	40	1.92	0.63	1.09	0.43	2.49	2.11	0.26
885	36	40	2.86	0.46	1.09	0.36	2.48	1.43	0.15
448	33	40	1.95	0.56	1.08	0.37	2.44	2.28	0.24
655	36	40	2.75	0.46	1.16	0.52	2.37	1.45	0.15
151	36	40	2.47	-.58	1.31	0.85	2.36	1.79	-0.14
787	34	40	2.04	0.55	1.08	0.35	2.31	2.10	0.13
792	34	40	2.25	0.47	1.20	0.74	2.28	1.93	0.08

471	36	40	2.42	0.48	0.89	-0.17	2.23	1.40	0.28
834	32	40	1.77	0.56	1.09	0.45	2.20	1.83	0.23
674	37	40	2.85	0.43	1.03	0.22	2.12	1.26	0.05
362	33	40	2.00	0.62	1.22	0.9	2.11	1.74	0.09
786	33	40	1.88	0.45	1.52	1.74	2.11	1.78	-0.02
833	25	40	0.38	0.46	1.12	0.65	2.07	2.66	0.40
237	35	40	2.65	0.39	1.26	0.84	2.05	1.40	0.03
635	37	40	3.25	0.51	1.41	0.92	2.05	1.16	0.01
399	36	40	2.90	0.65	1.38	0.98	2.00	1.17	0.09
612	32	40	1.93	0.58	1.00	0.08	1.95	1.70	0.33
213	31	40	1.30	0.44	1.39	1.63	1.94	2.08	0.03
597	28	40	1.26	0.42	1.04	0.29	1.93	2.28	0.32
405	27	40	0.86	0.38	1.14	0.78	1.91	2.66	0.26
863	35	40	2.30	0.50	0.87	-0.30	1.90	1.41	0.32
582	34	40	2.17	0.48	1.33	1.13	1.87	1.32	0.09
849	34	40	2.24	0.47	1.13	0.51	1.86	1.58	0.07
573	17	40	-0.17	0.35	1.23	1.70	1.85	2.87	0.13
595	34	40	2.04	0.48	1.35	1.19	1.84	1.35	0.06
381	34	40	2.00	0.47	1.24	0.85	1.83	1.46	0.09
823	29	40	1.20	0.35	1.19	0.97	1.80	1.93	0.26
557	28	40	0.91	0.48	1.46	2.10	1.79	2.33	0.07
789	29	40	0.96	0.48	0.98	-0.01	1.78	1.98	0.39
233	32	40	1.62	0.40	1.43	1.66	1.77	1.46	-0.01
236	38	40	3.38	0.39	0.98	0.17	1.77	0.93	0.14
668	34	40	0.05	0.40	1.18	0.69	1.76	1.34	0.10
130	26	40	0.98	0.43	1.24	1.39	1.75	2.39	0.23
814	38	40	3.55	0.75	1.06	0.30	1.74	0.91	0.14
246	35	40	2.14	0.47	0.92	-0.14	1.71	1.07	0.31
848	24	40	0.76	0.37	1.29	1.71	1.71	2.50	0.23
67	17	40	-0.46	0.75	1.38	2.23	1.70	2.62	0.16
884	34	40	1.96	0.37	1.04	0.24	1.70	1.32	0.21
54	27	40	0.90	0.75	1.30	1.63	1.69	2.38	0.12
238	33	40	2.05	0.51	0.99	0.04	1.68	1.28	0.36
264	6	40	-2.04	0.37	1.21	0.82	1.68	1.05	0.16
788	37	40	2.59	0.36	1.18	0.52	1.68	0.96	0.11
205	38	40	3.21	0.47	1.24	0.57	1.67	0.87	0.02
445	14	40	-0.82	0.37	1.34	2.12	1.66	2.40	0.06
450	36	40	2.71	0.47	1.02	0.18	1.65	1.03	0.13
152	19	40	0.02	0.48	1.43	2.34	1.64	2.45	0.20
232	32	40	1.41	0.63	1.07	0.38	1.64	1.50	0.19
664	38	40	3.41	0.76	1.11	0.38	1.61	0.83	0.05
370	35	40	2.67	0.36	0.96	-0.04	1.58	0.89	0.30
102	25	40	0.59	0.55	1.14	0.95	1.57	1.87	0.31
722	29	40	0.89	0.37	1.23	1.09	1.57	1.83	0.19

Note. Attempt statistics in misfit order with Attempt Number, Score, Count, Measure, Standard Error, Infit and Outfit (Mean Square Residuals and Standardised z) and Partial Correlation.

These 62 misfitting attempts were not acceptably fitting the Rasch Model (≥1.5 outfit mean square), and therefore it was decided to eliminate them from the analysis. Once the 62 misfitting attempts were removed, the item statistics were recalculated as outlined in Table 7.2.2.1.2.3. Most items were observed to have significantly improved. This finding suggested that the mis-fitting attempts were affecting the functioning of some items. However, some items remained out of the acceptable range.

Table 7.2.2.1.2.3 Item statistics after second rasch calibrations

					Infit		Outfit		
Item	**Score**	**Count**	**Measure**	**SE**	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**	**Corr.**
98	92	128	0.43	0.21	1.22	2.15	1.61	2.97	0.16
226	91	106	-0.68	0.29	1.14	0.75	1.56	1.72	0.09
217	31	78	1.97	0.26	1.36	3.19	1.5	2.71	0.12
235	67	105	0.73	0.22	1.17	1.85	1.48	2.70	0.24
216	39	77	1.38	0.25	1.25	2.62	1.46	3.30	0.17
220	88	121	0.37	0.22	1.16	1.52	1.45	2.43	0.17
17	101	129	-0.07	0.23	1.28	2.13	1.43	1.79	0.02
62	46	109	1.78	0.21	1.24	2.72	1.43	3.32	0.19
108	127	148	-0.77	0.25	1.07	0.45	1.40	1.30	0.24
196	68	113	1.03	0.21	1.27	3.39	1.38	3.18	0.05
18	96	121	-0.05	0.24	1.15	1.16	1.34	1.47	0.12
50	83	104	-0.08	0.26	1.09	0.63	1.34	1.51	0.17
19	47	121	2.07	0.20	1.20	2.41	1.33	2.57	0.18
159	85	100	-0.59	0.30	1.11	0.65	1.3	0.86	0.24
199	59	132	1.62	0.19	1.24	3.44	1.3	3.16	0.10
167	101	128	-0.05	0.23	1.18	1.37	1.28	1.23	0.16

Note. Items listed in misfit order with Item Number, Score, Count, Measure, Standard Error, Infit and Outfit (Mean Square Residuals and Standardised z) and Partial Correlation of most mis-fitting items, when the 62 most mis-fitting attempts (Outfit Mean Square ≥1.5) have been eliminated.

7.2.2.1.3 Item Invariance

The results in Table 7.2.2.1.2.3 outline the changed statistics for items after removing 62 misfitting attempts from the data. For example, once misfitting attempts were removed, the negative point correlation of Item 174 that existed from the raw data analysis was no longer evident. This finding indicated that this item worked well with the other items when misfitting attempts were eliminated. Accordingly, it was important to further examine the results from this approach to determine and confirm that removing these attempts was appropriate in developing the measurement scale for the test. Therefore, item invariance was examined next.

The results of examining the item invariance when attempts with outfit mean square statistics ≥1.5 are eliminated are represented in Figure 7.2.2.1.3. Attempts were divided in two groups, according to ability; namely: high-ability attempts (score ≥31), and low-ability attempts (score <31). Pairs of calibrations are plotted for each item using the pair Rasch-modelled ability estimate measures. These Rasch item measures were based exactly on the totals for each subsample (high and low ability). Most items are located inside the control lines, indicating they are constant within the limits of measurement error and remain invariant despite coming from different ability subsamples of students. However, some items lie outside the control lines. For example, Item 98 lies furthest outside the control line with an item difficulty measure for the high-ability subsample of 1.39 and an item difficulty measure for the low-ability subsample of -0.52.

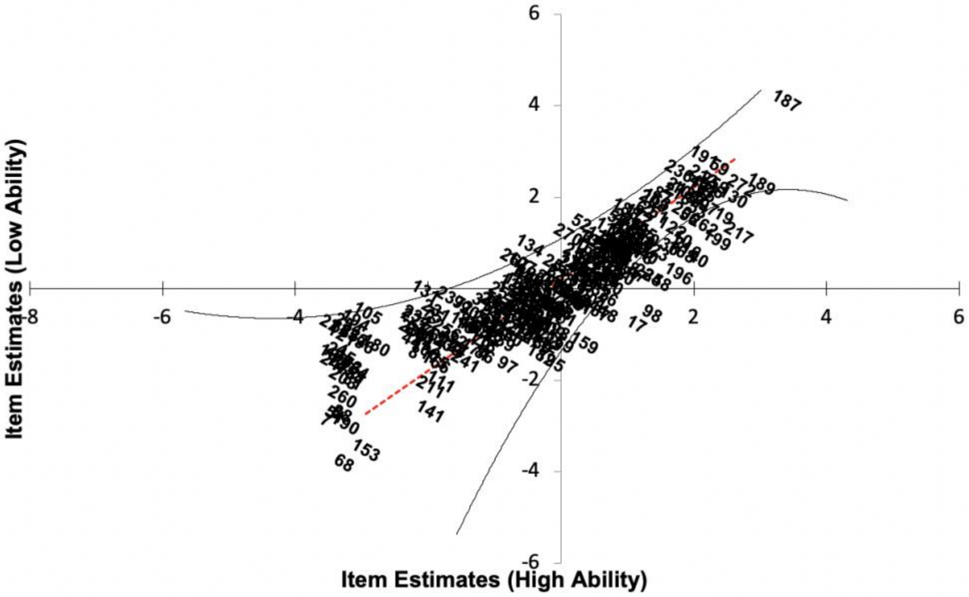


Figure 7.2.2.1.3.1 Item difficulty invariance. Data from 823 test attempts has been divided into two (397 high ability and 426 low ability) and the pair of calibrations for each item have been plotted. The red diagonal line in the figure represents calibrated mean of 0 logits. The outer curved lines represent quality-control lines to display whether the distribution of the plotted points is close enough to the modelled relationship diagonal line for the measures to be

regarded as sufficiently invariant. The 95% control lines are based on the SEs for each of the item pairs.

In order to determine if the most misfitting items (outlined in Table 7.2.2.1.2.3) were sufficiently invariant, an examination of the difference in item measures between the subsamples was conducted (Table 7.2.2.1.3). Not surprisingly, Item 98, the most mis-fitting item, also displayed the largest difference in subsample item measures, therefore not displaying sufficient invariance.

Table 7.2.2.1.3.1 Changes in item difficulty in subsamples

	Subsample A		**Subsample B**	
Item	**Item Measure**	**Standard Error**	**Item Measure**	**Standard Error**
98	1.39	0.27	-0.52	0.32
226	-0.53	0.52	-0.74	0.35
217	2.69	0.36	1.24	0.32
235	1.33	0.33	0.32	0.28
216	1.91	0.35	1.24	0.32

7.2.3 Anchored Item Measures

We have previously established that the data must fit the demanding requirement of the Rasch Model. Therefore, to most appropriately set the item anchors for measuring persons across all attempts, it was necessary to consider all of the analyses above. Firstly, it was clear that item statistics and functioning were improved when misfitting attempts were eliminated, and item invariance was satisfactory when those misfitting attempts were removed. Therefore, the decision of the items anchors was based on the results of the item statistics and item invariance when misfitting attempts ≥1.5 were eliminated. Furthermore, several aspects needed to be considered in these analyses. For example, the item statistic results indicated underfit to the model and overfit to the model; however, they each have different implications. In particular, overfitting performances suggest data that are too good to be true. However, Bond and Fox (2015) suggested that it is likely that overfit will have no implications at all and suggest that if overfitting items are eliminated, it may “rob the test of its best items” (p. 274).

Therefore, all results in the analyses above were carefully considered in order to produce the most accurate and reliable item anchors. Table 7.2.3 identifies the items considered for elimination from further analysis and provides the action and justification for each decision. These actions were taken and further calibrations produced the item anchors (Appendix A). The produced item anchors were applied to further analysis in this study, allowing for accurate comparisons between attempts.

Table 7.2.3.1 Action on misfitting items justification of action

Item	Action	Justification
98	This item and its data will be eliminated from the data set in order to set anchors. It will not be included for analysis of comparisons of attempt measures. However, it will be included when changes in item facility are investigated as changes in this item's difficulty between first and final attempts remains of interest.	This item lies furthest outside the control line. The invariance of this item is too much to be considered acceptable. Similarly, the outfit mean square and z-statistic are both considerably outside the range of acceptability. This indicates underfit to the model displaying too much unpredictability. Considering these traits displaying that the item does not fit the Rasch Model, it is justified to suggest that there is something wrong with the item.
226	This item is not considered for data elimination.	This item lies within the control lines of invariance and is therefore considered an invariant item in spite of coming from different ability subsample of students. The fit statistics for this item are close enough to an acceptable range.
217	This item will remain, although the data will be removed for the item as they will not be used for the anchors. It will remain only as an item for analysis.	This item lies outside the control line of invariance. The invariance of this item is too much to be considered acceptable. The outfit mean square value is within acceptable range but the z-statistic is outside the range of acceptability.
235	This item will remain, although the data will be removed for the item as they will not be used to for the anchors. It will remain only as an item for analysis.	This item lies slightly outside the control line of invariance. The invariance of this item is too much to be considered acceptable. The outfit mean square value is within acceptable range but the z-statistic is outside the range of acceptability.
216	This item will remain, although the data will be removed for the item as they will not be used to for the anchors. It will remain only as an item for analysis.	This item lies slightly outside the control line of invariance. The invariance of this item is too much to be considered acceptable. The outfit mean square value is within acceptable range but the z-statistic is outside the range of acceptability.
220	This item is not considered for data elimination.	This item lies within the control lines of invariance and is therefore considered an invariant item in spite of coming from different ability subsample of students. The fit statistics for this item are close enough to an acceptable range.

7.2.4 Diagnostic Test Performance

Firstly, to examine overall student performance, logit measures of all attempts were determined. Mean logit scores were calculated for all attempts at Institution A (Table 7.2.4.1), Institution B (Table 7.2.4.2), and all attempts combined (Table 7.2.4.3). The mean (±SEM) attempt measure was 1.39±0.45 at Institution A, 1.81±0.49 at Institution B, and 1.53±0.46 when combined. Institution B displayed a higher mean logit measure than Institution A. These results were consistent with the raw score results outlined in Chapter Six.

Table 7.2.4.1 Summary statistics (in logits) of institution A attempts

	Measure	**SE**
Mean	1.39	0.45
SEM	0.04	0.01
S.D.	1.05	0.15
Max.	5.82	1.85
Min.	-2.11	0.34

Note. n = 595.

Table 7.2.4.2 Summary statistics (in logits) of institution B attempts

	Measure	**SE**
Mean	1.81	0.49
SEM	0.06	0.01
S.D.	1.02	0.20
Max.	5.96	1.86
Min.	-0.17	0.35

Note. n = 290

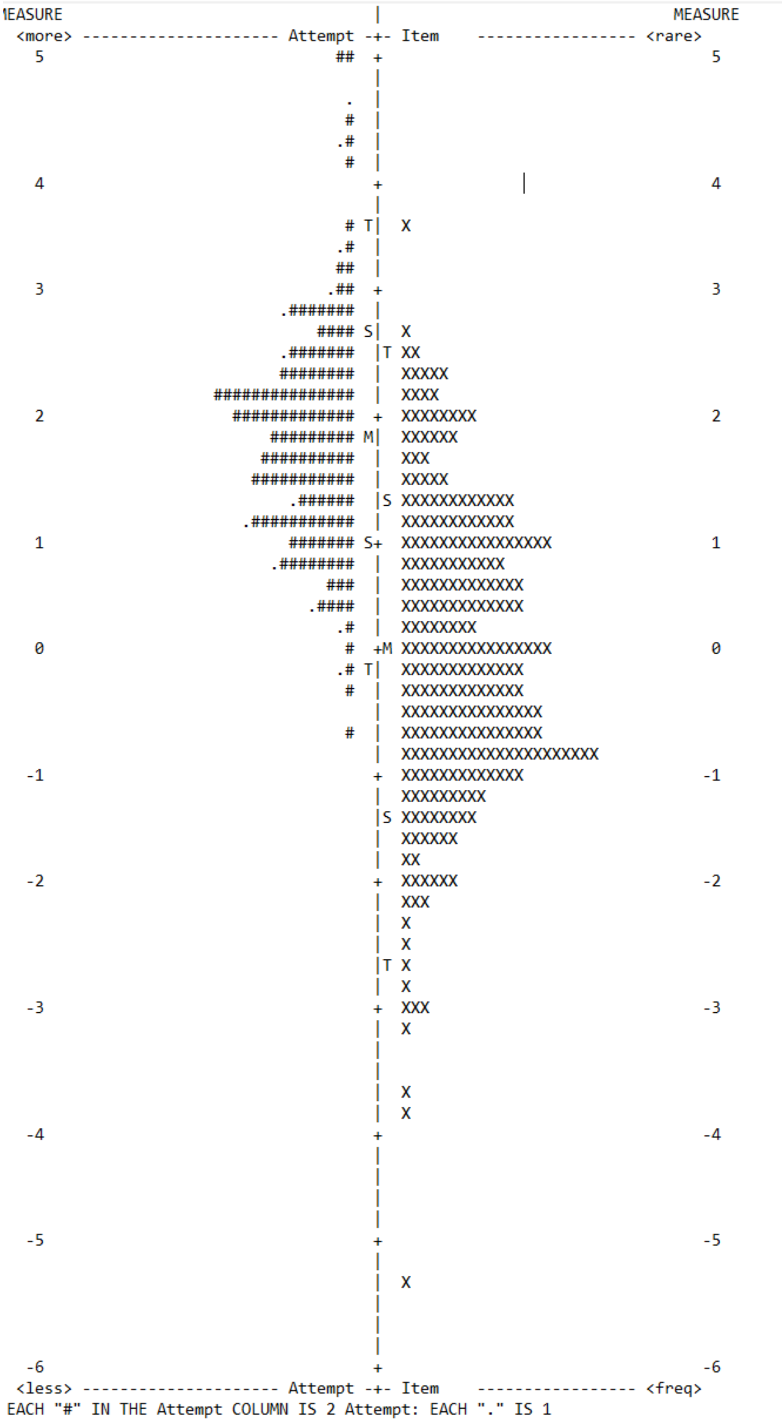
Table 7.2.4.3 Summary statistics (in logits) of all attempts

	Measure	**SE**
Mean	1.53	0.46
SEM	0.04	0.01
S.D.	1.06	0.17
Max.	5.96	1.86
Min.	-2.11	0.34

Note. n = 885.

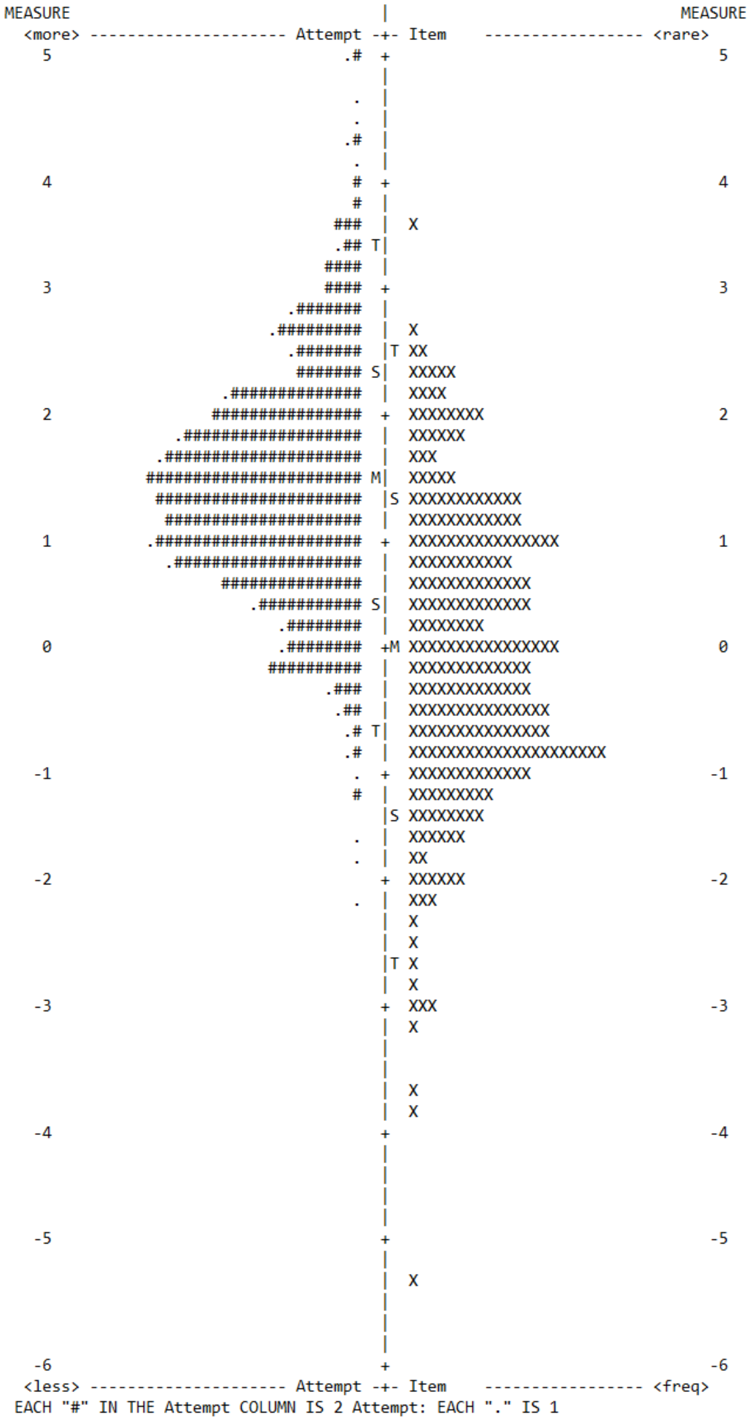
To visually observe the distribution of attempts in comparison with item measures, Item-Attempt maps, known as Wright Maps (Bond & Fox, 2015), were developed for Institution A (Figure 7.2.4A), Institution B (Figure 7.2.4B), and all attempts (Figure 7.2.4C). Many things were observed and determined from the Wright Maps. Firstly, the logit scale is displayed in the middle of the measurement unit common to both person ability (on the left) and item difficulty (on the right). Secondly, it can be observed that eleven items were anchored at the mean of the item difficulty (logit=0), and the remaining items were spread above and below 0 to represent their difficulties relative to these items' logit measures. Thirdly, when an attempt was in a horizontal line with an item, it indicated a 50% chance of success on that item. That is, the item is perfectly suited to measure that person's ability. Finally, the mean logit for attempts can be observed at 'M'. The 'S' indicates a measure that is one standard deviation higher (or lower), and the 'T' shows a measure of two standard deviations higher (or lower). Overall, the Wright Maps produced indicated that the distributions of attempts were slightly top-heavy in comparison with the item distribution. This suggests that the sample has performed quite well on the test overall.

A



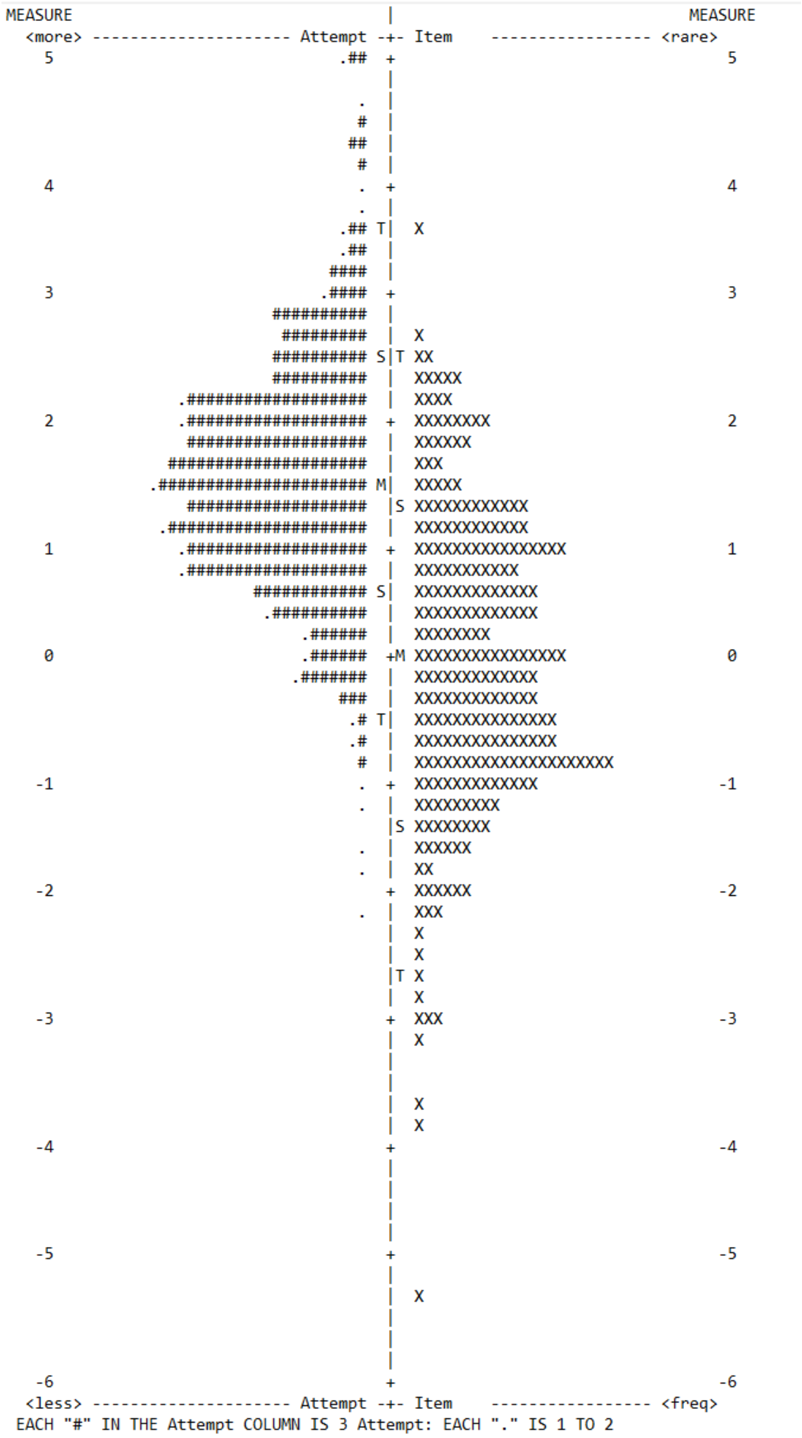
```
MEASURE                                |                               MEASURE
 <more> --------------------- Attempt -+- Item    ----------------- <rare>
   5                               ##  +                                   5
                                       |
                                    .  |
                                    #  |
                                   .#  |
                                    #  |
   4                                   +               |                   4
                                       |
                                    # T|  X
                                   .#  |
                                   ##  |
   3                              .##  +                                   3
                             .#######  |
                                 #### S|  X
                             .#######  |T XX
                             ########  |  XXXXX
                      ###############  |  XXXX
   2                    #############  +  XXXXXXXX                         2
                            ######### M|  XXXXXX
                           ##########  |  XXX
                          ###########  |  XXXXX
                              .######  |S XXXXXXXXXXXX
                         .###########  |  XXXXXXXXXXXX
   1                          ####### S+  XXXXXXXXXXXXXXXX                 1
                            .########  |  XXXXXXXXXXX
                                  ###  |  XXXXXXXXXXXXX
                                .####  |  XXXXXXXXXXXXX
                                   .#  |  XXXXXXXX
   0                                #  +M XXXXXXXXXXXXXXXX                 0
                                   .# T|  XXXXXXXXXXXXX
                                    #  |  XXXXXXXXXXXXX
                                       |  XXXXXXXXXXXXXXX
                                    #  |  XXXXXXXXXXXXXXX
                                       |  XXXXXXXXXXXXXXXXXXXXX
  -1                                   +  XXXXXXXXXXXXX                   -1
                                       |  XXXXXXXXX
                                       |S XXXXXXXX
                                       |  XXXXXX
                                       |  XX
  -2                                   +  XXXXXX                          -2
                                       |  XXX
                                       |  X
                                       |  X
                                       |T X
                                       |  X
  -3                                   +  XXX                             -3
                                       |  X
                                       |
                                       |
                                       |  X
                                       |  X
  -4                                   +                                  -4
                                       |
                                       |
                                       |
                                       |
  -5                                   +                                  -5
                                       |
                                       |  X
                                       |
                                       |
                                       |
  -6                                   +                                  -6
 <less> --------------------- Attempt -+- Item    ----------------- <freq>
EACH "#" IN THE Attempt COLUMN IS 2 Attempt: EACH "." IS 1
```

B



```
MEASURE                                 |                               MEASURE
  <more> --------------------- Attempt -+- Item    ----------------- <rare>
    5                               .#  +                                   5
                                        |
                                     .  |
                                     .  |
                                    .#  |
                                     .  |
    4                                #  +                                   4
                                     #  |
                                   ###  |  X
                                   .## T|
                                  ####  |
    3                             ####  +                                   3
                              .#######  |
                            .#########  |  X
                              .#######  |T XX
                               ####### S|  XXXXX
                       .##############  |  XXXX
    2                 ################  +  XXXXXXXX                         2
                  .###################  |  XXXXXX
                .#####################  |  XXX
               ####################### M|  XXXXX
                ######################  |S XXXXXXXXXXXX
                 #####################  |  XXXXXXXXXXXX
    1          .######################  +  XXXXXXXXXXXXXXXX                 1
                 .####################  |  XXXXXXXXXXX
                       ###############  |  XXXXXXXXXXXXX
                          .########### S|  XXXXXXXXXXXXX
                             .########  |  XXXXXXXX
    0                        .########  +M XXXXXXXXXXXXXXXX                 0
                            ##########  |  XXXXXXXXXXXXX
                                  .###  |  XXXXXXXXXXXXX
                                   .##  |  XXXXXXXXXXXXXXX
                                    .# T|  XXXXXXXXXXXXXXX
                                    .#  |  XXXXXXXXXXXXXXXXXXXXX
   -1                                .  +  XXXXXXXXXXXXX                   -1
                                     #  |  XXXXXXXXX
                                        |S XXXXXXXX
                                     .  |  XXXXXX
                                     .  |  XX
   -2                                   +  XXXXXX                          -2
                                     .  |  XXX
                                        |  X
                                        |  X
                                        |T X
                                        |  X
   -3                                   +  XXX                             -3
                                        |  X
                                        |
                                        |
                                        |  X
                                        |  X
   -4                                   +                                  -4
                                        |
                                        |
                                        |
                                        |
   -5                                   +                                  -5
                                        |
                                        |  X
                                        |
                                        |
                                        |
   -6                                   +                                  -6
  <less> --------------------- Attempt -+- Item    ----------------- <freq>
 EACH "#" IN THE Attempt COLUMN IS 2 Attempt: EACH "." IS 1
```

C



```
MEASURE                                 |                               MEASURE
  <more> --------------------- Attempt -+- Item    ----------------- <rare>
    5                              .##  +                                   5
                                        |
                                     .  |
                                     #  |
                                    ##  |
                                     #  |
    4                                .  +                                   4
                                     .  |
                                   .## T|  X
                                   .##  |
                                  ####  |
    3                            .####  +                                   3
                            ##########  |
                             #########  |  X
                            ########## S|T XX
                            ##########  |  XXXXX
                  .###################  |  XXXX
    2             .###################  +  XXXXXXXX                         2
                   ###################  |  XXXXXX
                 #####################  |  XXX
               .###################### M|  XXXXX
                   ###################  |S XXXXXXXXXXXX
                .#####################  |  XXXXXXXXXXXX
    1             .###################  +  XXXXXXXXXXXXXXXX                 1
                  .###################  |  XXXXXXXXXXX
                          ############ S|  XXXXXXXXXXXXX
                           .##########  |  XXXXXXXXXXXXX
                               .######  |  XXXXXXXX
    0                          .######  +M XXXXXXXXXXXXXXXX                 0
                              .#######  |  XXXXXXXXXXXXX
                                   ###  |  XXXXXXXXXXXXX
                                    .# T|  XXXXXXXXXXXXXXX
                                    .#  |  XXXXXXXXXXXXXXX
                                     #  |  XXXXXXXXXXXXXXXXXXXXX
   -1                                .  +  XXXXXXXXXXXXX                   -1
                                     .  |  XXXXXXXXX
                                        |S XXXXXXXX
                                     .  |  XXXXXX
                                     .  |  XX
   -2                                   +  XXXXXX                          -2
                                     .  |  XXX
                                        |  X
                                        |  X
                                        |T X
                                        |  X
   -3                                   +  XXX                             -3
                                        |  X
                                        |
                                        |
                                        |  X
                                        |  X
   -4                                   +                                  -4
                                        |
                                        |
                                        |
                                        |
   -5                                   +                                  -5
                                        |
                                        |  X
                                        |
                                        |
                                        |
   -6                                   +                                  -6
  <less> --------------------- Attempt -+- Item    ----------------- <freq>
 EACH "#" IN THE Attempt COLUMN IS 3 Attempt: EACH "." IS 1 TO 2
```

Figure 7.2.4.1 Wright Maps displaying distribution of attempts in comparison with item measures. (A) Data represents each individual attempt measure indicated on the left-hand side (‘#’ = 3 attempts) for Institution A. (B) Data represents each individual attempt measure, indicated on the left-hand side (‘#’ = 3 attempts) for Institution B. (C) Data represents each

individual attempt measure, indicated on the left-hand side ('#' = 3 attempts) for all attempts combined. Data represents anchored item measures, indicated by 'X' on the right-hand side. 'S' indicated one standard deviation from the mean and 'T' indicates two standard deviations from the mean.

Attempts were then separated into attempt numbers to determine if improvements had been made over time. Summary statistics were calculated for each attempt number from Institution A (Table 7.2.4.4), Institution B (Table 7.2.4.5), and combined attempts (Table 7.2.4.6). At Institution A, the highest mean logit score was evident on Attempt 13. The highest mean logit score was evident at Institution B on Attempt 10. When all attempts were combined, the highest mean logit score was evident on Attempt 10. In summary, the results displayed that performance improved over time.

Table 7.2.4.4 Summary statistics at each attempt number (1-13) for institution A

Attempt Number	n	Mean	SEM	SD	Max.	Min.
1	267	1.20	0.06	1.03	3.90	-2.11
2	108	1.34	0.09	0.94	3.73	-1.74
3	67	1.58	0.12	1.00	4.36	-0.41
4	38	1.68	0.16	1.00	4.28	-0.19
5	23	1.65	0.16	0.77	3.60	0.29
6	15	1.47	0.23	0.88	3.04	-0.30
7	14	1.29	0.19	0.72	2.78	0.19
8	12	1.78	0.36	1.20	4.72	-0.16
9	9	1.77	0.35	1.06	4.05	0.07
10	8	1.93	0.41	1.15	4.29	0.70
11	7	2.18	0.55	1.37	4.53	0.60
12	6	1.67	0.48	1.18	3.48	0.10
13	4	2.56	1.14	2.29	5.64	0.44

Note. Attempt Number, Number of Attempts, Mean Measure, SEM, SD, Maximum, and Minimum measure.

Table 7.2.4.5 Summary statistics at each attempt number (1-11) for institution B

Attempt Number	n	Mean	SEM	SD	Max	Min
1	118	1.79	0.10	1.09	5.96	-0.71
2	52	1.80	0.14	1.01	5.50	-0.04
3	31	1.61	0.16	0.90	4.25	0.12
4	22	1.72	0.20	0.94	4.26	-0.23
5	14	1.97	0.38	1.42	5.42	0.14
6	14	1.85	0.22	0.83	3.29	0.56
7	11	2.04	0.34	1.13	4.47	0.08
8	8	2.08	0.33	0.89	2.98	0.37
9	4	2.32	0.32	0.65	2.90	1.42
10	4	2.34	0.42	0.84	3.55	1.68
11	3	1.48	0.15	0.26	1.70	1.19

Note. Attempt Number, Number of Attempts, Mean Measure, SEM, SD, Maximum, and Minimum measure.

Table 7.2.4.6 Summary statistics of each attempt number (1-13)

Attempt Number	n	Mean	SEM	SD	Max	Min
1	385	1.41	0.06	1.12	5.96	-2.11
2	160	1.49	0.08	0.98	5.50	-1.74
3	98	1.59	0.10	0.96	2.04	-3.53
4	60	1.70	0.13	0.97	4.28	-0.23
5	37	1.77	0.17	1.04	5.42	0.14
6	29	1.66	0.16	0.85	3.29	-0.30
7	25	1.62	0.20	0.98	4.47	0.08
8	20	1.90	0.25	1.08	4.72	-0.16
9	13	1.94	0.27	0.96	1.05	0.07
10	12	2.07	0.30	1.04	4.29	0.70
11	10	1.97	0.37	1.18	4.53	0.60
12	8	1.82	0.36	1.03	3.48	-0.10
13	5	1.65	0.52	1.03	2.86	0.44

Note. Attempt Number, Number of Attempts, Mean Measure, SEM, SD, Maximum and Minimum Measure.

7.2.4.1 Identifying Change in Measures Over Time

We then sought to explore changes in performance over time by examining measures on first and final test attempts. The specific purpose of examining these differences was to investigate whether the test could be used to improve capabilities. Therefore, we combined the data and applied the method of stacking and racking the data within the Rasch Model. More specifically, stacking and racking allowed for the attempts to be measured in the same frame of reference to track the changes in persons and items between one attempt and another. This is a necessary method to allow accurate comparisons between performances (Wright, 2003).

While the mean of logit measures were reported and used to determine whether there was a statistical difference between first and final attempts, the distance between logits was not intended to be over-interpreted to inform conclusions (Bond & Fox, 2015). Therefore, the comparison between first and final attempts were quantified through scatterplots to inform many of the findings.

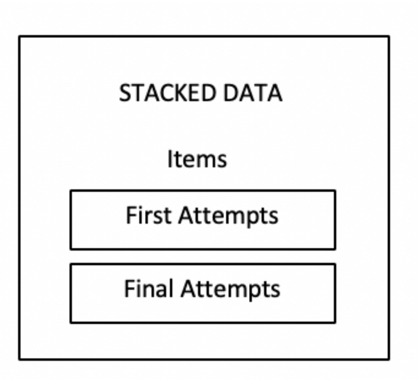


Figure 7.2.4.1.1 Conceptual visualisation of stacked data

7.2.4.1.1 Change in TES' Performance

In order to track changes in TES' performance between first and final attempts, a stacked analysis was used. This analysis allowed for an examination of changes in TES' performance over time and allowed for a determination as to whether they improved their performance from their first attempt to final attempt or if their performance declined. In other words, improvements in ability were made where student measure estimates increased from first to final attempt. Alternatively, if student measures decreased from first to final attempt, a regression in ability had occurred.

Stacking the data was performed by appending the person measures for the final attempts onto the first attempts resulting in twice as many attempts being measured (Figure 7.2.4.1.1.1). This transformed the first and final measures onto the same ruler, and the anchored item measures produced in the previous section were used to measure the attempts.

Data for all TES who attempted the test more than once (n=160) were restructured and stacked so that each TES' first and final attempt measures could be meaningfully compared. The stacked analysis produced two measures for each TES, one for first attempt and one for final attempt. The stacked data were displayed in a Wright Map to represent person and item relations (Figure 7.2.4.1.1.2A) and was then separated to visually observe the distribution of first attempts (Figure 7.2.4.1.1.2B) and final attempts (Figure 7.2.4.1.1.2C) separately. The item measures, previously anchored, are indicated by the 'X' on the right-hand side. In summary, the final attempt measures were distributed higher and more top-heavy than first attempt measures.

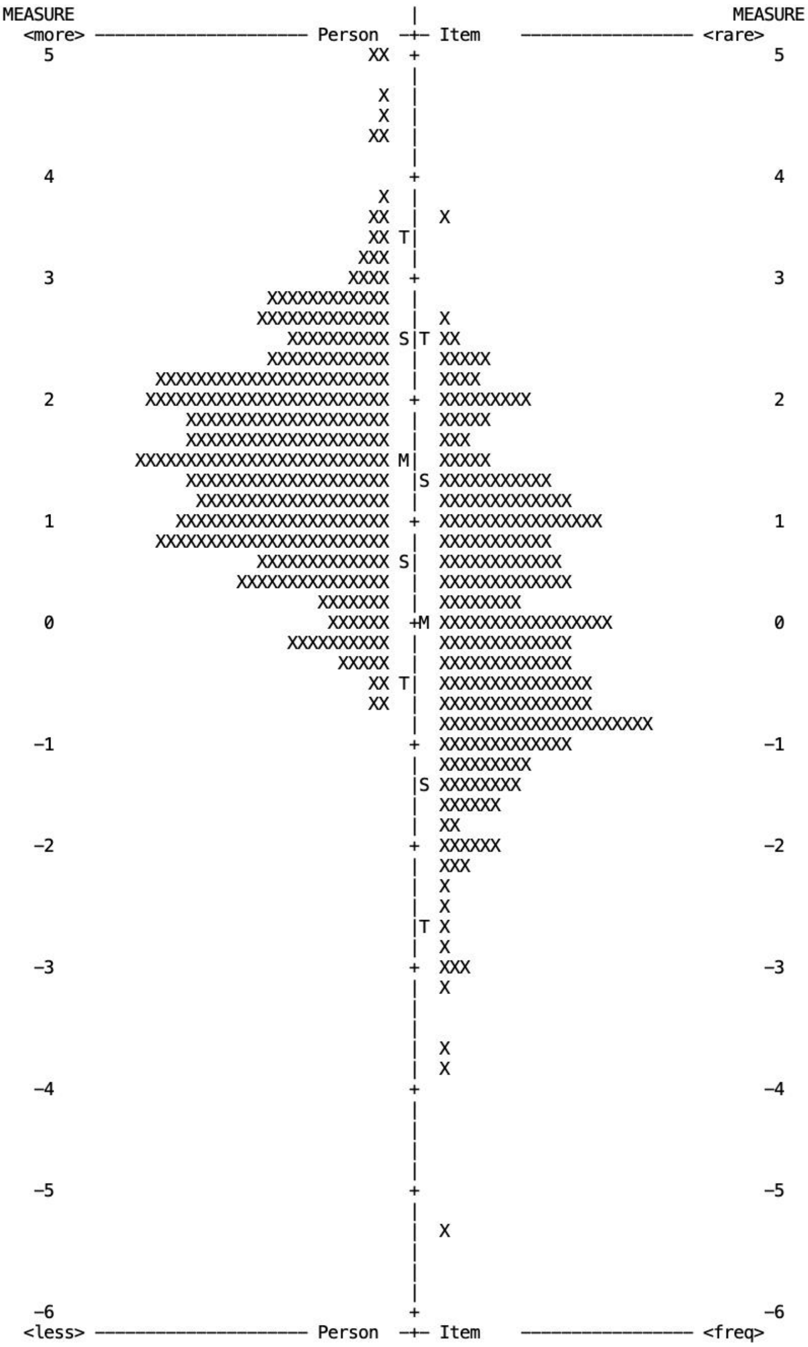
Descriptive statistics were produced to compare measures and performance on first and final attempts (Table 7.2.4.1.1). It was evident that the mean ability measures increased from first to final attempts at Institution A, Institution B, and when all results were combined. Overall improvements in performance were evident and an increase was also evident in the minimum, median, and maximum measures in all three groups. There was a statistical difference evident between first and final attempts for students at Institution A, Institution B, and all students combined (*p* <0.0001).

Table 7.2.4.1.1.1 Comparison of first and final attempt descriptive statistics for ability measures for institution a, institution b, and all students

	n	Mean	SEM	SD	Min.	Median	Max.
Institution A							
First	108	3.83	0.09	0.93	1.96	3.78	7.00
Final	108	4.47	0.10	1.04	2.15	4.43	8.73
Institution B							
First	52	4.25	0.16	1.14	1.86	4.16	8.41

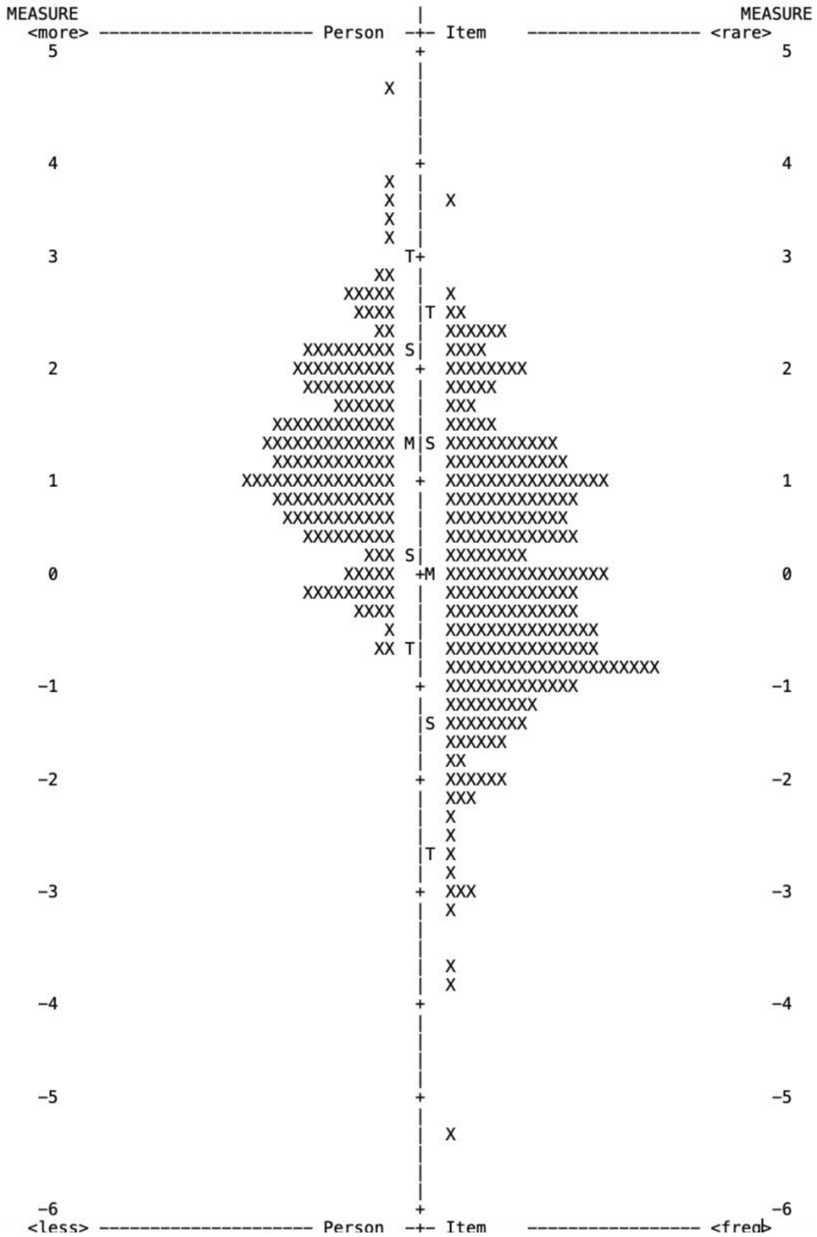
Final	52	4.86	0.15	1.12	3.31	4.82	10.11
All Students							
First	160	3.97	0.08	1.02	1.86	3.92	8.41
Final	160	4.60	0.09	1.08	2.15	4.61	10.11

A



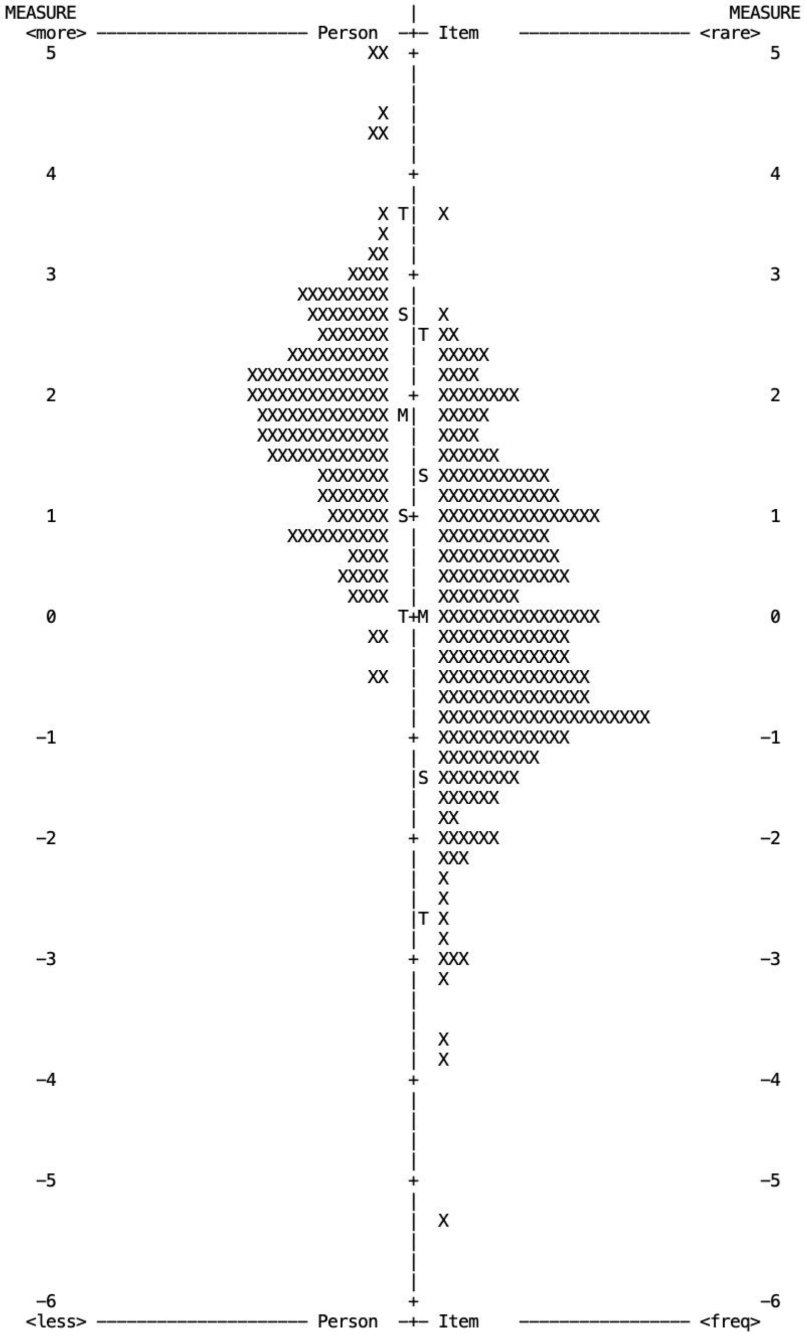
```
MEASURE                                 |                               MEASURE
  <more> --------------------- Person  -+- Item    ----------------- <rare>
    5                               XX  +                                   5
                                        |
                                     X  |
                                     X  |
                                    XX  |
                                        |
    4                                   +                                   4
                                     X  |
                                    XX  |  X
                                    XX T|
                                   XXX  |
    3                             XXXX  +                                   3
                          XXXXXXXXXXXX  |
                         XXXXXXXXXXXXX  |  X
                            XXXXXXXXXX S|T XX
                          XXXXXXXXXXXX  |  XXXXX
               XXXXXXXXXXXXXXXXXXXXXXX  |  XXXX
    2         XXXXXXXXXXXXXXXXXXXXXXXX  +  XXXXXXXXX                        2
                  XXXXXXXXXXXXXXXXXXXX  |  XXXXX
                  XXXXXXXXXXXXXXXXXXXX  |  XXX
             XXXXXXXXXXXXXXXXXXXXXXXXX M|  XXXXX
                  XXXXXXXXXXXXXXXXXXXX  |S XXXXXXXXXXX
                   XXXXXXXXXXXXXXXXXXX  |  XXXXXXXXXXXXX
    1            XXXXXXXXXXXXXXXXXXXXX  +  XXXXXXXXXXXXXXXX                 1
               XXXXXXXXXXXXXXXXXXXXXXX  |  XXXXXXXXXXX
                         XXXXXXXXXXXXX S|  XXXXXXXXXXXX
                       XXXXXXXXXXXXXXX  |  XXXXXXXXXXXXX
                               XXXXXXX  |  XXXXXXXX
    0                           XXXXXX  +M XXXXXXXXXXXXXXXXX                0
                            XXXXXXXXXX  |  XXXXXXXXXXXXX
                                 XXXXX  |  XXXXXXXXXXXXX
                                    XX T|  XXXXXXXXXXXXXXX
                                    XX  |  XXXXXXXXXXXXXXX
                                        |  XXXXXXXXXXXXXXXXXXXXX
   -1                                   +  XXXXXXXXXXXXX                   -1
                                        |  XXXXXXXXX
                                        |S XXXXXXXX
                                        |  XXXXXX
                                        |  XX
   -2                                   +  XXXXXX                          -2
                                        |  XXX
                                        |  X
                                        |  X
                                        |T X
                                        |  X
   -3                                   +  XXX                             -3
                                        |  X
                                        |
                                        |
                                        |  X
                                        |  X
   -4                                   +                                  -4
                                        |
                                        |
                                        |
                                        |
   -5                                   +                                  -5
                                        |
                                        |  X
                                        |
                                        |
                                        |
   -6                                   +                                  -6
  <less> --------------------- Person  -+- Item    ----------------- <freq>
```

B



```
MEASURE                                  |                               MEASURE
  <more> --------------------- Person   -+- Item    ----------------- <rare>
    5                                    +                                   5
                                         |
                                      X  |
                                         |
                                         |
                                         |
    4                                    +                                   4
                                      X  |
                                      X  |  X
                                      X  |
                                      X  |
    3                                   T+                                   3
                                     XX  |
                                  XXXXX  |  X
                                   XXXX  |T XX
                                     XX  |  XXXXXX
                              XXXXXXXXX S|  XXXX
    2                        XXXXXXXXXX  +  XXXXXXXX                         2
                              XXXXXXXXX  |  XXXXX
                                 XXXXXX  |  XXX
                           XXXXXXXXXXXX  |  XXXXX
                          XXXXXXXXXXXXX M|S XXXXXXXXXXX
                           XXXXXXXXXXXX  |  XXXXXXXXXXXX
    1                   XXXXXXXXXXXXXXX  +  XXXXXXXXXXXXXXXX                 1
                           XXXXXXXXXXXX  |  XXXXXXXXXXXXX
                            XXXXXXXXXXX  |  XXXXXXXXXXXX
                              XXXXXXXXX  |  XXXXXXXXXXXXX
                                    XXX S|  XXXXXXXX
    0                             XXXXX  +M XXXXXXXXXXXXXXXX                 0
                              XXXXXXXXX  |  XXXXXXXXXXXXX
                                   XXXX  |  XXXXXXXXXXXXX
                                      X  |  XXXXXXXXXXXXXXX
                                     XX T|  XXXXXXXXXXXXXXX
                                         |  XXXXXXXXXXXXXXXXXXXXX
   -1                                    +  XXXXXXXXXXXXX                   -1
                                         |  XXXXXXXXX
                                         |S XXXXXXXX
                                         |  XXXXXX
                                         |  XX
   -2                                    +  XXXXXX                          -2
                                         |  XXX
                                         |  X
                                         |  X
                                         |T X
                                         |  X
   -3                                    +  XXX                             -3
                                         |  X
                                         |
                                         |
                                         |  X
                                         |  X
   -4                                    +                                  -4
                                         |
                                         |
                                         |
                                         |
   -5                                    +                                  -5
                                         |
                                         |  X
                                         |
                                         |
                                         |
   -6                                    +                                  -6
  <less> --------------------- Person   -+- Item    ----------------- <freq>
```

C



```
MEASURE                                 |                               MEASURE
  <more> --------------------- Person  -+- Item    ----------------- <rare>
    5                               XX  +                                   5
                                        |
                                        |
                                     X  |
                                    XX  |
                                        |
    4                                   +                                   4
                                        |
                                     X T|  X
                                     X  |
                                    XX  |
    3                             XXXX  +                                   3
                             XXXXXXXXX  |
                              XXXXXXXX S|  X
                               XXXXXXX  |T XX
                            XXXXXXXXXX  |  XXXXX
                        XXXXXXXXXXXXXX  |  XXXX
    2                   XXXXXXXXXXXXXX  +  XXXXXXXX                         2
                         XXXXXXXXXXXXX M|  XXXXX
                         XXXXXXXXXXXXX  |  XXXX
                          XXXXXXXXXXXX  |  XXXXXX
                               XXXXXXX  |S XXXXXXXXXXX
                               XXXXXXX  |  XXXXXXXXXXXX
    1                           XXXXXX S+  XXXXXXXXXXXXXXXX                 1
                            XXXXXXXXXX  |  XXXXXXXXXXX
                                  XXXX  |  XXXXXXXXXXXX
                                 XXXXX  |  XXXXXXXXXXXXX
                                  XXXX  |  XXXXXXXX
    0                                  T+M XXXXXXXXXXXXXXXX                 0
                                    XX  |  XXXXXXXXXXXXX
                                        |  XXXXXXXXXXXXX
                                    XX  |  XXXXXXXXXXXXXXX
                                        |  XXXXXXXXXXXXXXX
                                        |  XXXXXXXXXXXXXXXXXXXXX
   -1                                   +  XXXXXXXXXXXXX                   -1
                                        |  XXXXXXXXXX
                                        |S XXXXXXXX
                                        |  XXXXXX
                                        |  XX
   -2                                   +  XXXXXX                          -2
                                        |  XXX
                                        |  X
                                        |  X
                                        |T X
                                        |  X
   -3                                   +  XXX                             -3
                                        |  X
                                        |
                                        |
                                        |  X
                                        |  X
   -4                                   +                                  -4
                                        |
                                        |
                                        |
                                        |
   -5                                   +                                  -5
                                        |
                                        |  X
                                        |
                                        |
                                        |
   -6                                   +                                  -6
  <less> --------------------- Person  -+- Item    ----------------- <freq>
```

Figure 7.2.4.1.1.1 Wright maps displaying distribution of stacked data. (A) Data represents distribution of all first attempt measures with anchored item measures. (B) Data represents distribution of all final attempt measures with anchored item measures. (C) Data represents distribution of first and final attempts with anchored item measures. Attempts are indicated on the left-hand side 'X' and anchored item measures indicated by 'X' on the right-hand side. 'S' indicated one standard deviation from the mean and 'T' indicates two standard deviations from the mean.

To further investigate and visualise the changes in measures, student ability measures for first and final attempts were plotted against each other (Figure 7.2.4.1.1.3). A gain in ability was evident in 126 out of 160 (79%) students. This gain in ability was seen in 84 students (78%) from Institution A and 42 students (81%) from Institution B. Interestingly, it was evident that 45 students (28%) increased their ability measure by more than one logit. In contrast, 34 students (21%) displayed a loss in ability; 7 of those students (4%) displayed a loss in ability measure of more than one logit. There were no students with the same measure on both first and final attempts. In summary, most TES were observed to have improved between first and final attempts on the Diagnostic Test.

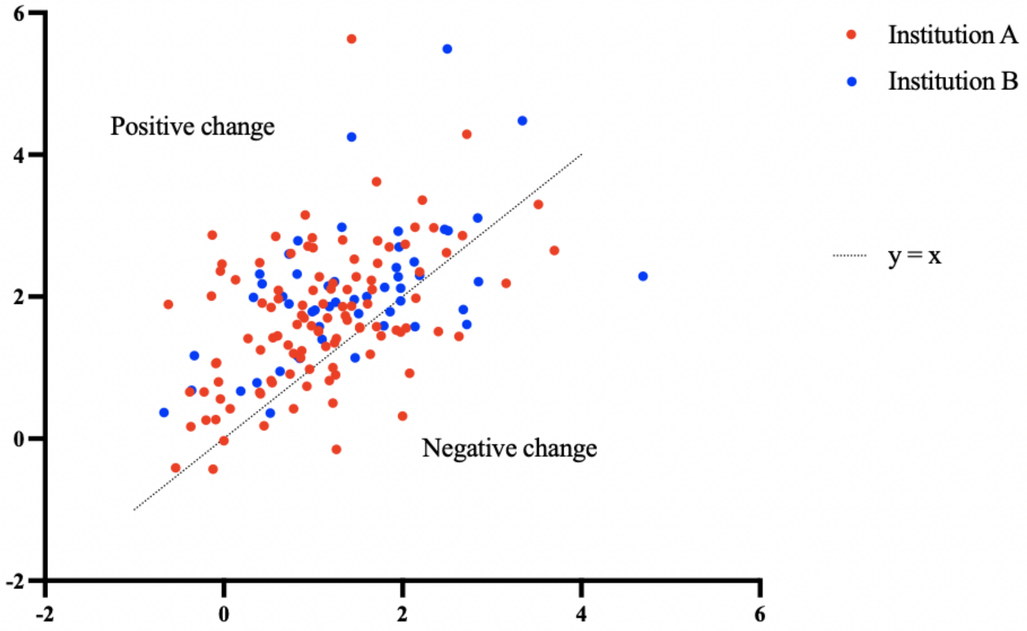


Figure 7.2.4.1.1.2 Scatterplot of first attempt versus final attempt measures. Data represents individual TES results from Institution A and Institution B. The line (y = x) indicates no change in performance (i.e., attempt measure) and the positions either side of the line indicate either a positive change (gain in ability) or a negative change (regression in performance).

7.2.4.1.2 Change in Item Difficulties

As outlined previously, an item measure quantifies the difficulty of a question. That is, the higher the item measure, the more difficult the question is quantified to be. Therefore, the expectation would be that item difficulties may change between first and final attempts. If an item measure was lower on the final attempt than the first attempt, this would indicate that the TES have become more capable of correctly completing that item. In other words, items become easier when the person's ability is higher. Therefore, we sought to explore and examine changes in the difficulty of items between first and final attempts.

A racked analysis was applied to allow for these changes in item measures to be explored. In particular, this method shifted the focus from who and how many people are improving to the changes in ability observed on the assessed content. For this reason, the results from each institution were combined for this analysis to allow for overall improvements in the assessed content to be determined. This analysis intended to identify the items that are decreasing in measures and determine whether these can be classified into any particular categories (e.g., content areas, item types, context domains, or ACSF Levels). Further, it was considered that if item estimates increased from first to final attempt, a regression in ability had been displayed. Therefore, the exploration of items with increased measures from first to final attempts was also of great interest as it was possible that misconceptions had been developed.

Racking the data was achieved by attaching the final attempt item responses onto the first attempt item responses resulting in each person taking twice as many items (Figure 7.2.4.1.2.1). This allowed for the measurement of item difficulty estimates to be made on the same ruler and therefore allowed for valid conclusions about changes in item measures from first to final attempts.

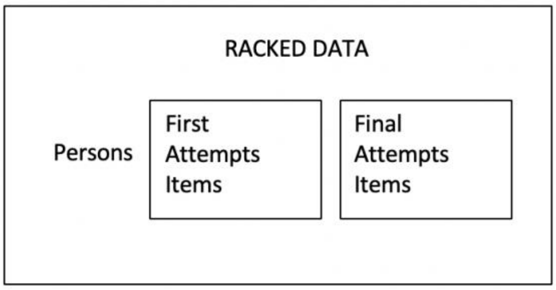


Figure 7.2.4.1.2.1 Conceptual visualisation of racking the data.

Data for the 160 students who attempted the test more than once were restructured and racked for all items (n=272). When comparing item measures on first attempts and item measures on final attempts (Table 7.2.4.1.2), it was evident that the mean item measure was lower on final attempts. In fact, the median item measure, the minimum item measure, and the maximum item measure were all lower on final attempt analysis compared to first attempt analysis. There was a statistical difference between item measures on first and final attempts (*p* <0.0001). In summary, the items were calculated to be easier on final attempt indicating results were consistent with the stacked analysis, which found that there was a gain in overall TES ability between first and final attempts.

Table 7.2.4.1.2.1 Comparison of descriptive statistics for first attempt and final attempt item measures

	First Attempt	**Final Attempt**
n	272	272
Mean	-1.27	-1.87
SEM	0.08	0.08
SD	1.38	1.38
Min.	-4.95	-5.25
Median	-1.22	-1.87
Max.	2.73	1.92

To further investigate these changes, item measures calculated on first and final attempts were plotted against each other (Figure 7.2.4.1.2.2). Overall, 191 items (70%) displayed decreased item difficulty measure, as shown by the points below the dotted line. There were no items with the exact logit measure on the first attempt and final attempt, although Item 143 showed the closest measures with a first attempt measure of -2.79 and final attempt measure of -2.78. In summary, most items displayed a decrease in measure, indicating that students became more capable of correctly answering those items. However, it was evident that some items increased their item measure; therefore, misconceptions might have been developed.

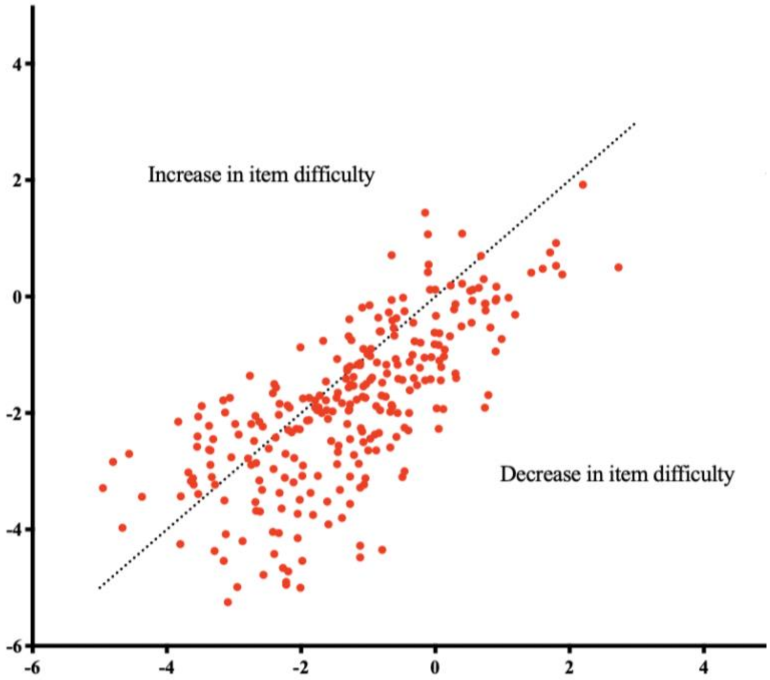


Figure 7.2.4.1.2.2 Scatterplot of first attempt versus final attempt item measures. First attempt item measures (x-axis) versus final attempt item measures (y-axis). Data represents all items (n=272). Items above the line represent items estimated to be more difficult on final attempt, while items below the line represent items were estimated to be less difficult on final attempt. The line (y = x) indicates no change in item measure.

7.2.4.1.3 Change in Item Difficulties in Categories

To determine if any links existed between changes in items of certain categories, items were separated into four categories; three calculator-allowed categories (NA, MG, & SP) and the NC category. Further analyses with the racked data were then performed, and descriptive statistics of first and final item measures were calculated (Table 7.2.4.1.3). Between the first and final attempts, the mean for NA items decreased from -1.34 to -1.86, the mean for MG items decreased from -1.28 to -1.86, the mean for SP items decreased from -1.12 to -1.61, and the mean for NC items decreased from -1.33 to -2.09. There was a statistically significant decline in measures for NA items (*p* <0.0001), MG items (*p* <0.0001), SP items (*p* <0.001), and NC items (*p* <0.0001). In summary, the results show that students became more capable of correctly answering questions in all categories on final attempt compared to first attempt.

Table 7.2.4.1.3.1 Descriptive statistics for first and final attempt item measures in each item category

	n	Mean	SEM	SD	Min.	Median	Max.
NA							
First	79	-1.34	0.15	1.36	-4.37	-1.25	1.19
Final	79	-1.86	0.16	1.40	-5.25	-1.88	1.44
MG							
First	64	-1.28	0.19	1.53	-4.95	-1.18	2.73
Final	64	-1.86	0.18	1.47	-5.00	-1.88	1.07
SP							
First	60	-1.12	0.20	1.53	-4.80	-1.09	2.20
Final	60	-1.61	0.19	1.46	-4.90	-1.52	1.92
NC							
First	69	-1.33	0.13	1.11	-4.66	-1.21	1.43
Final	69	-2.09	0.14	1.19	-4.66	-1.95	0.70

To identify individual changes in item difficulties from first to final attempts and determine whether improvements could be identified in specific categories more than others, separate scatterplots were produced. The analysis of first and final attempts for items showed variation within each item category. Specifically, 52 of the 79 NA items (66%) decreased in difficulty measure (Figure 7.2.4.1.3A), 43 out of 64 MG items (67%) decreased in difficulty measure (Figure 7.2.4.1.3B), 42 out of 60 SP items (70%) decreased in item difficulty (Figure 7.2.4.1.3C), and 54 out of 69 NC items (78%) decreased in item difficulty (Figure 7.2.4.1.3D). In summary, the NC category displayed the most items that had decreased in measure, while the other three categories showed a similar change. This suggested that ability had improved most on NC items and least on calculator-allowed NA items, albeit the difference between all calculator-allowed categories was small.

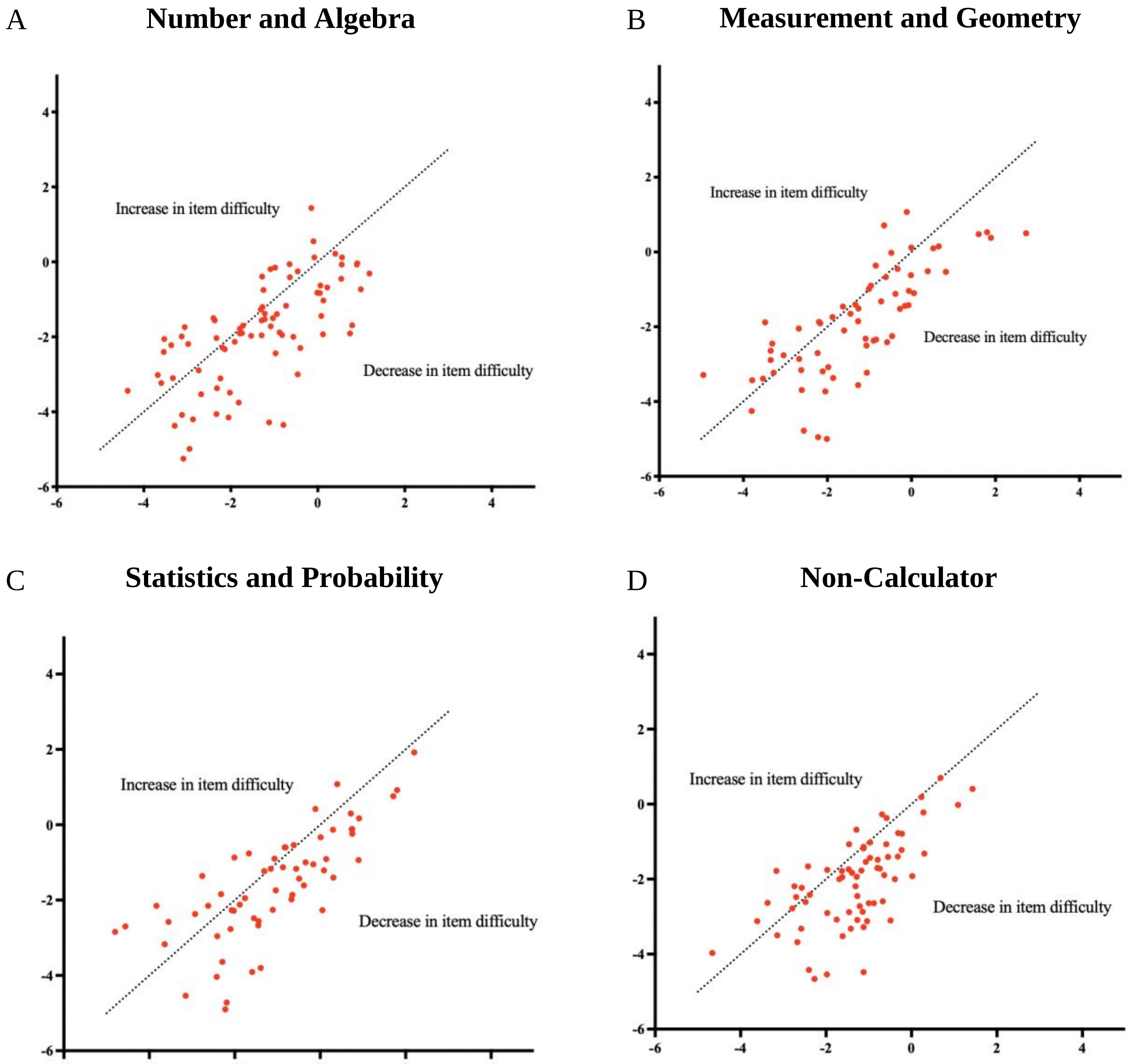


Figure 7.2.4.1.3.1 Scatterplot of first attempt versus final attempt item measures in test categories. Scatterplot of first attempt item measures (x-axis) versus final attempt item measures (y-axis) from racked analysis. (A) Data represent all NA items (n=79). (B) Data represent all MG items (n=64). (C) Data represent all SP items (n=60). (D) Data represent all NC items (n=69). Items above the line represent items estimated to be more difficult on final attempt, while items below the line represent items were estimated to be less difficult on final attempt. The line (y = x) indicates no change in item measure.

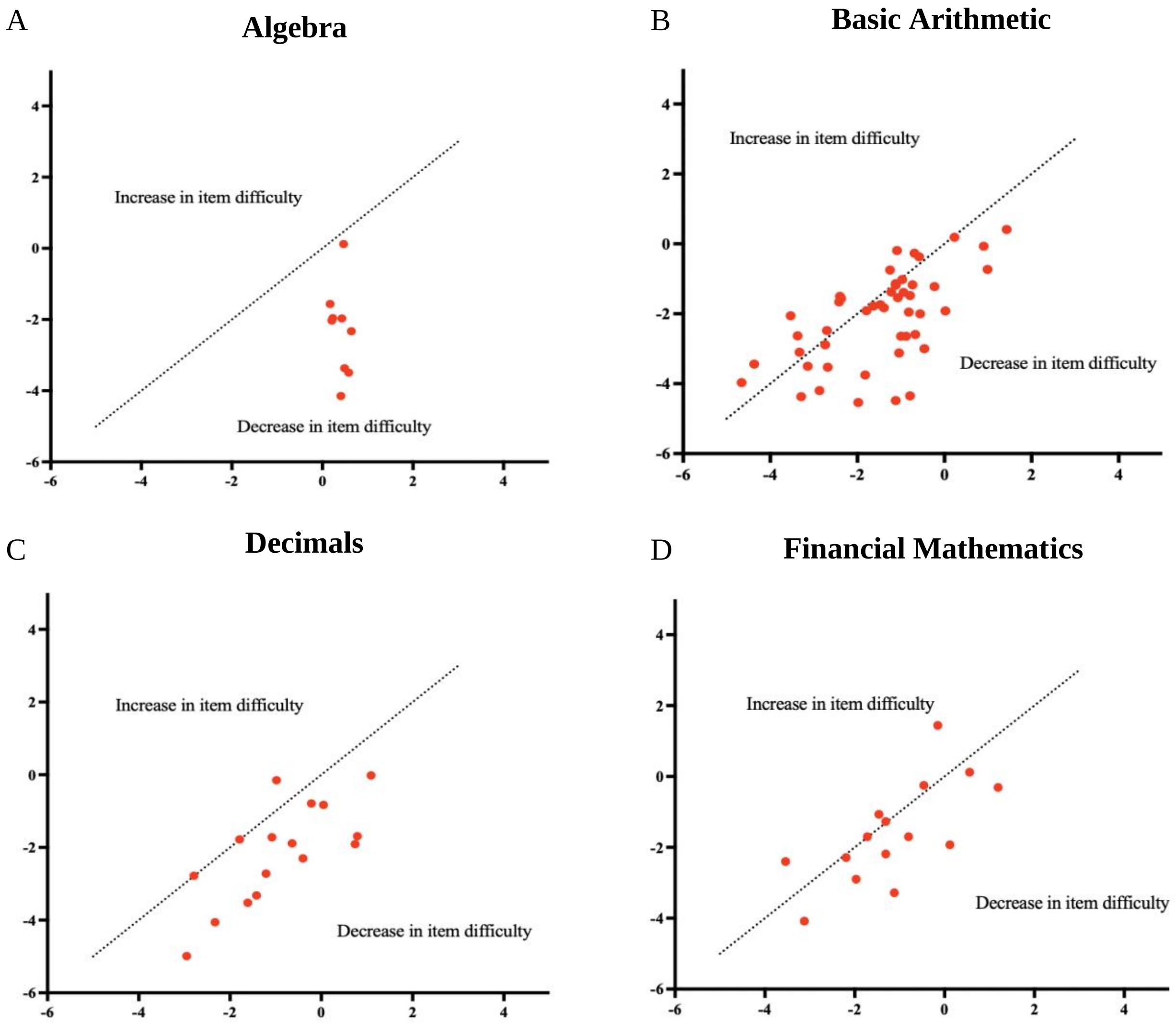
7.2.4.1.4 Change in NA Content Item Difficulties

Changes in item difficulties from first to final attempts were explored individually in the NA content area to determine whether improvements could be identified in certain NA content areas more than others. NA items in this analysis included calculator-allowed items and NC items. Descriptive statistics of first and final item measures in the NA content areas were calculated (Table 7.2.4.1.4.1) and show that the mean item measures decreased in all content areas between first and final attempts. However, a statistically significant decline was only evident for Algebra (*p* <0.0001) and Decimals (*p* <0.05). There was no statistical difference between first and final item measures for Basic Arithmetic, Financial Mathematics, Fractions, Percentages, or Rates and Ratios.

Table 7.2.4.1.4.1 Descriptive statistics for first and final attempt item measures in each NA content area

	n	Mean	SEM	SD	Min.	Median	Max.
Algebra							
First	9	0.40	0.06	0.17	0.17	0.43	0.64
Final	9	-2.30	0.42	1.26	-4.15	-2.03	0.12
Basic Arithmetic							
First	20	-1.64	0.37	1.68	-4.66	-1.23	1.43
Final	20	-2.09	0.32	1.44	-4.35	-2.03	0.41
Decimals							
First	16	-0.92	0.31	1.23	-2.95	-1.03	1.09
Final	16	-2.15	0.34	1.38	-4.99	-1.90	-0.02
Financial Mathematics							
First	15	-1.15	0.33	1.29	-3.54	-1.31	1.19
Final	15	-1.59	0.37	1.42	-4.08	-1.70	1.44
Fractions							
First	14	-1.18	0.32	1.18	-3.13	-1.20	0.30
Final	14	-1.81	0.28	1.06	-4.28	-1.62	-0.22
Percentages							
First	18	-1.79	0.28	1.17	-3.6	-1.75	0.55
Final	18	-2.36	0.3	1.28	-5.25	-2.21	-0.07
Rates & Ratios							
First	17	-0.72	0.28	1.16	-3.68	-0.64	0.91
Final	17	-1.08	0.27	1.11	-3.11	-0.68	0.55

To observe the number of items in each NA content area that decreased in measure between first and final attempts, separate scatterplots were produced and variations were evident between each topic. Interestingly, all nine Algebra items (100%) decreased in difficulty measure (Figure 7.2.4.1.4.1A). The number of items that decreased in item difficulty measure in the other NA content areas was 34 out of 47 for Basic Arithmetic (72%) (Figure 7.2.4.1.4.1B), 13 out of 16 for Decimals (81%) (Figure 7.2.4.1.4.1C), 9 out of 15 for Financial Mathematics (60%) (Figure 7.2.4.1.4.1D), 10 out of 14 for Fractions (71%) (Figure 7.2.4.1.4.1E), 13 out of 18 for Percentages (72%) (Figure 7.2.4.1.4F), and 12 out of 17 for Rates & Ratios (71%) decreased in item difficulty (Figure 7.2.4.1.4.1G). In summary, the greatest proportion of items that students became more capable of correctly answering was for Algebra, followed by Decimals. The greatest proportion of items students become less capable of correctly answering was for Financial Mathematics. All other NA content areas showed similar changes.



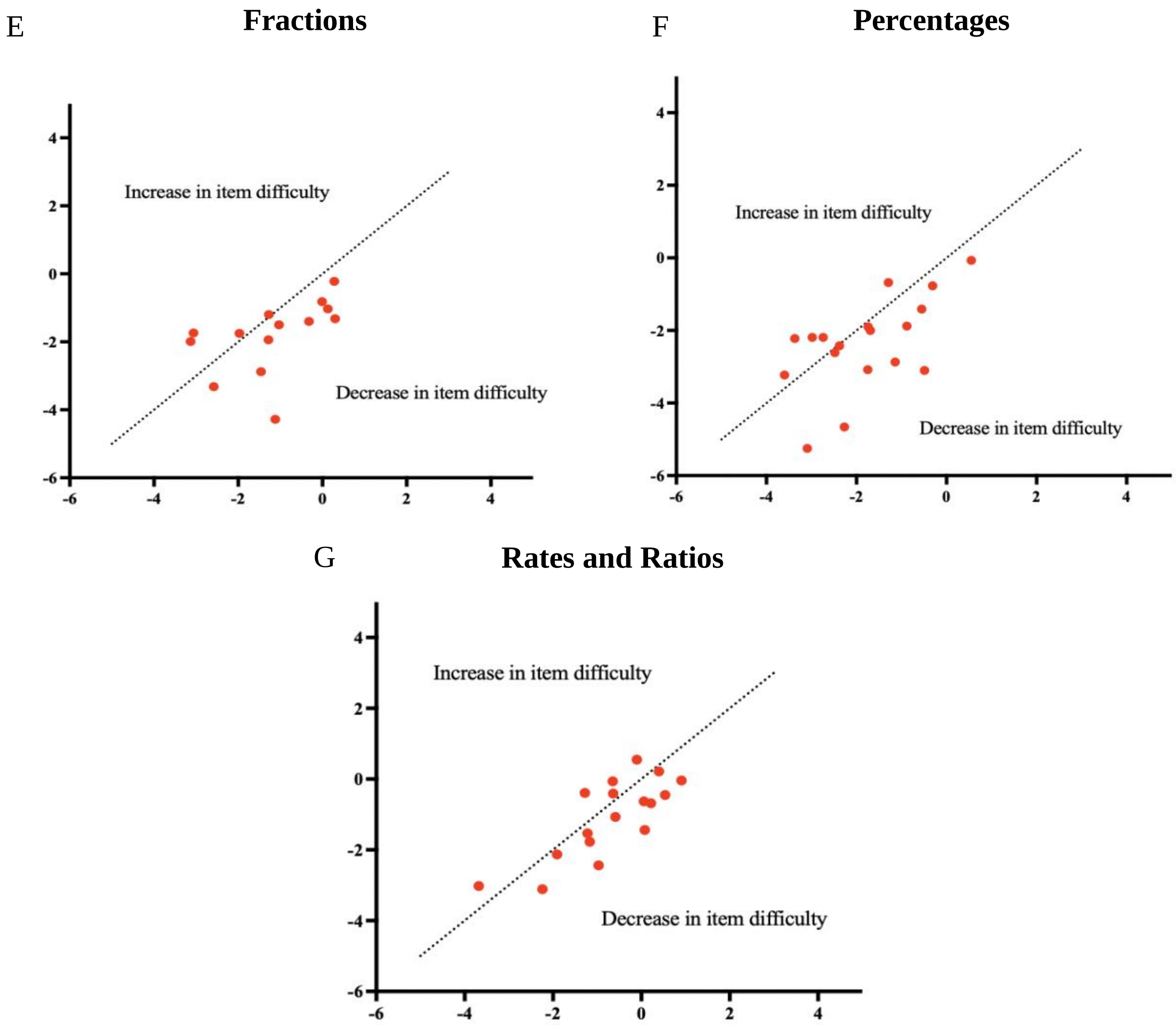


Figure 7.2.4.1.4.1 Scatterplot of first attempt versus final attempt item measures for NA content areas. Scatterplot of first attempt item measures (x-axis) versus final attempt item measures (y-axis) from racked analysis for NA content areas. (A) Data represent Algebra items (n = 9). (B) Data represent Basic Arithmetic items (n = 47). (C) Data represent Decimals items (n = 16). (D) Data represent Financial Mathematics items (n = 15). (E) Data represent Fractions items (n = 14). (F) Data represent Percentages items (n = 18). (G) Data represent Rates & Ratios items (n = 17). The line (y = x) indicates no change in item measure, items above the line represent items estimated as more difficult on final attempt while items below the line represent items estimated as less difficult on final attempt.

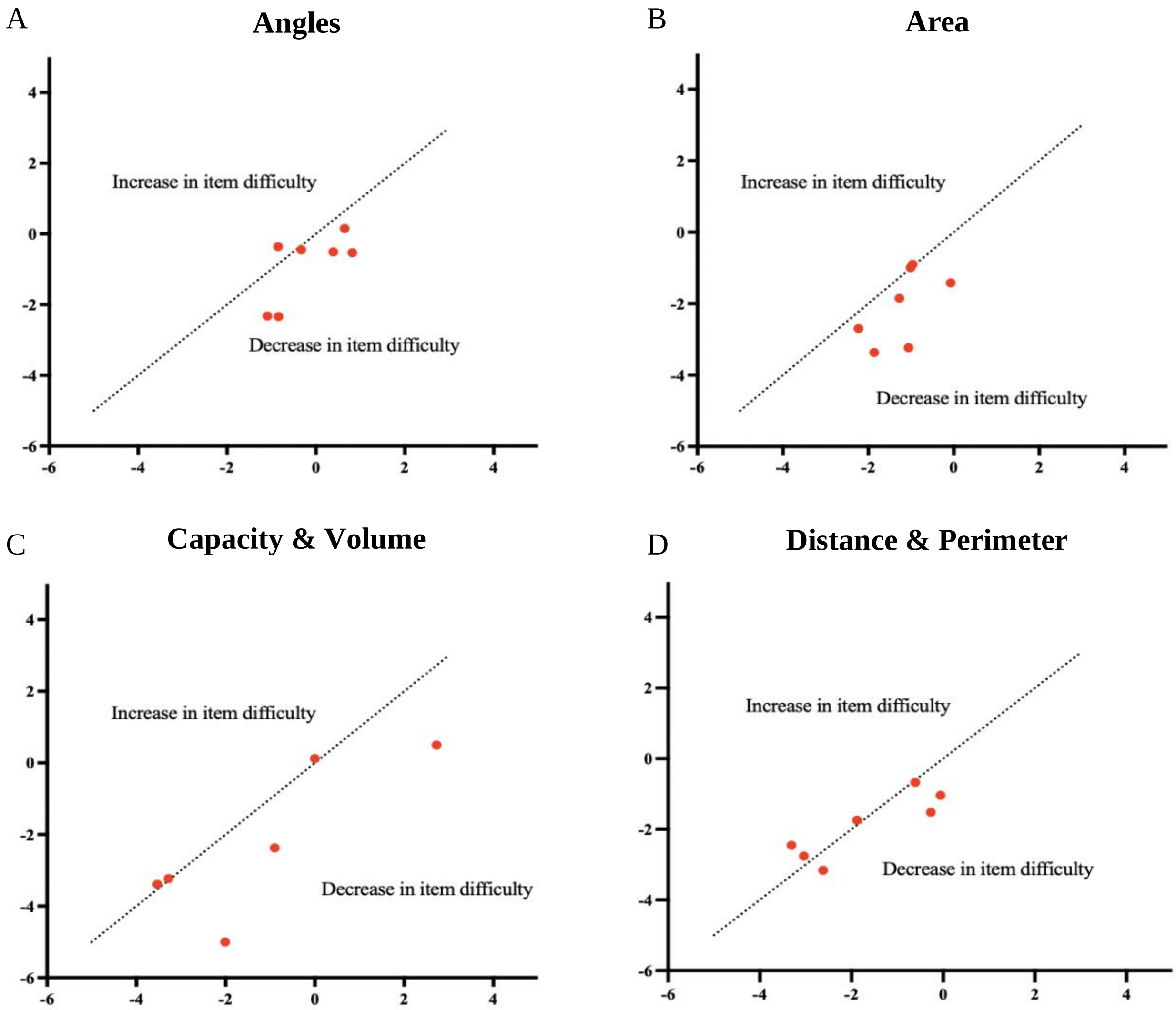
7.2.4.1.5 Change in MG Content Item Difficulties

Changes in item difficulties from first to final attempts were then explored individually in the MG content area to determine whether improvements could be identified in specific MG topics more than others. MG items in this analysis included calculator-allowed items and NC items. Descriptive statistics of first and final item measures in the MG content areas were calculated (Table 7.2.4.1.5) and show that the mean item measures decreased in all content areas between first and final attempts. However, no statistically significant decline was evident between first and final item measures for any of the MG content areas.

Table 7.2.4.1.5.1 Descriptive statistics for first and final attempt item measures in each MG content area

	n	Mean	SEM	SD	Min.	Median	Max.
Angles							
First	7	-0.18	0.30	0.79	-1.09	-0.33	0.82
Final	7	-0.91	0.38	1.00	-2.34	-0.51	0.15
Area							
First	7	-1.21	0.26	0.69	-2.23	-1.06	-0.07
Final	7	-2.07	0.39	1.04	-3.37	-1.85	-0.9
Capacity & Volume							
First	6	-1.17	0.96	2.34	-3.53	-1.46	2.73
Final	6	-2.23	0.88	2.15	-5.00	-2.80	0.50
Distance & Perimeter							
First	7	-1.68	0.52	1.37	-3.31	-1.88	-0.06
Final	7	-1.91	0.35	0.92	-3.16	-1.74	-0.67
Estimating, Reading & Converting							
First	27	-1.58	0.31	1.63	-4.95	-1.63	1.89
Final	27	-2.11	0.27	1.41	-4.95	-2.05	0.70
Space, Shapes & Symmetry							
First	7	-1.03	0.64	1.68	-3.80	-1.07	1.60
Final	7	-1.74	0.73	1.92	-4.25	-2.25	1.07
Time & Timetabling							
First	15	-1.56	0.31	1.21	-3.79	-1.45	0.52
Final	15	-1.99	0.39	1.50	-4.78	-1.87	0.71

The items in each MG content area were observed to determine the number of items that decreased in measure between first and final attempts and separate scatterplots were produced. Variations were evident between the content areas. Specifically, the number of items that decreased in item difficulty measure were 6 out of 7 for Angles (86%) (Figure 7.2.4.1.5A), 5 out of 7 for Area (71%) (Figure 7.2.4.1.5B), 3 out of 6 for Capacity and Volume (50%) (Figure 7.2.4.1.5C), 4 out of 7 for Distance and Perimeter (57%) (Figure 7.2.4.1.5D), 17 out of 27 Estimating, Reading and Converting (63%) (Figure 7.2.4.1.5E), 6 out of 7 Space, Shapes and Symmetry (86%) (Figure 7.2.4.1.5F), and 10 out of 15 for Time and Timetabling (67%) (Figure 7.2.4.1.5G). In summary, the greatest proportion of items that students became more capable of correctly answering was for Angles and Space, Shapes & Symmetry. The greatest proportion of items students became less capable of correctly answering was for Capacity and Volume, and Distance and Perimeter. All other MG content areas showed similar changes.



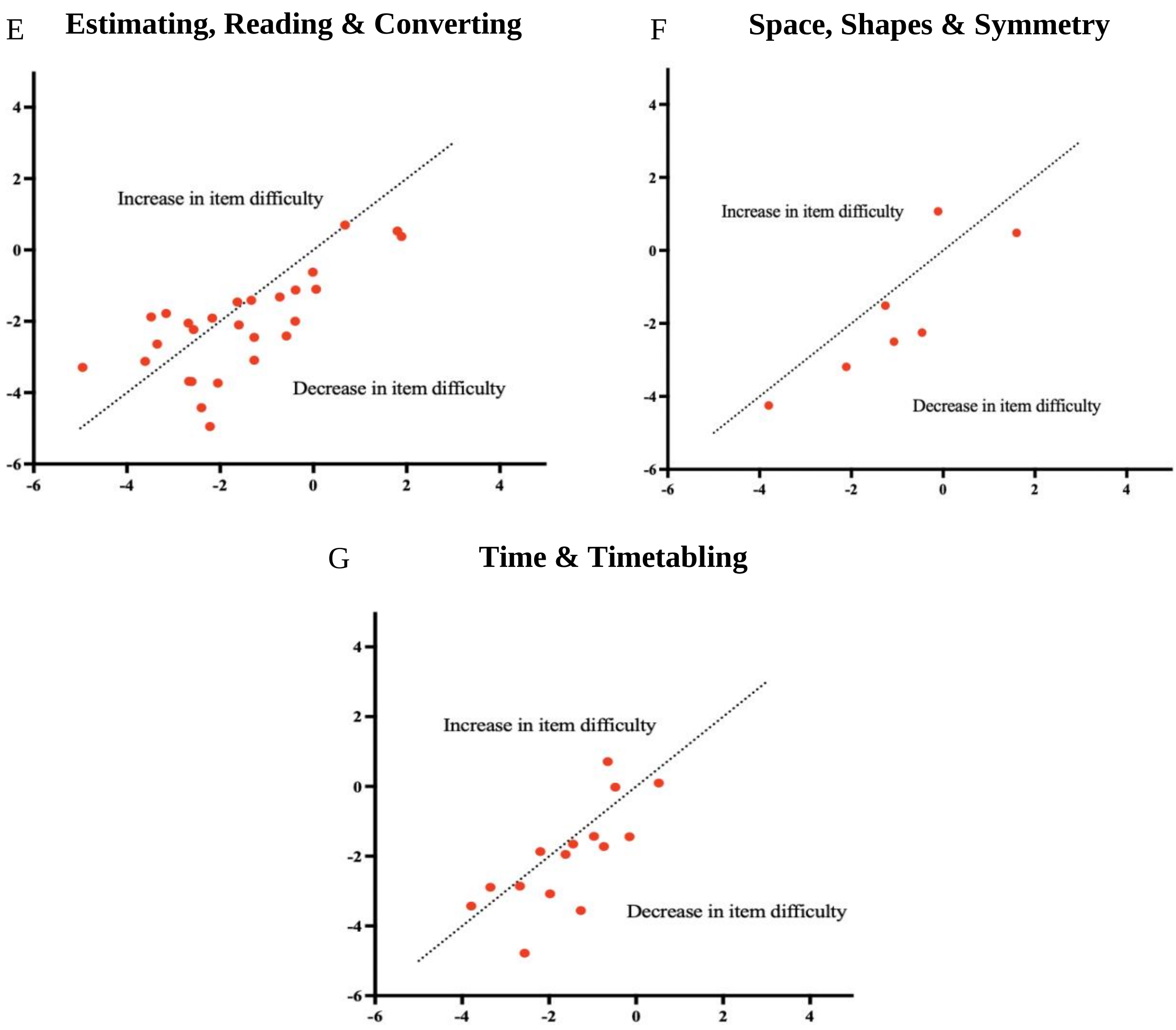


Figure 7.2.4.1.5.1 Scatterplot of first attempt versus final attempt item measures for MG content areas. Scatterplot of first attempt item measures (x-axis) versus final attempt item measures (y-axis) from racked analysis for MG content items. (A) Data represent Angles items (n = 7). (B) Data represent Area items (n = 7). (C) Data represents Capacity and Volume items (n = 6). (D) Data represent Distance and Perimeter items (n = 7). (E) Data represent Estimating, Reading and Converting items (n = 27). (F) Data represent Space, Shape and Symmetry items (n = 7). (G) Data represent Time and Timetabling items (n = 15). The line (y = x) indicates no change in item measure, items above the line represent items estimated as more difficult on final attempt while items below the line represent items estimated as less difficult on final attempt.

7.2.4.1.6 Change in SP Content Item Difficulties

Changes in item difficulties from first to final attempts were then explored individually in the SP content area to determine whether improvements could be identified in specific SP topics more than others. Descriptive statistics of first and final item measures in the SP content areas were calculated (Table 7.2.4.1.6) and show that the mean item measures decreased in all content areas between first and final attempts. However, no statistically significant decline was evident between first and final item measures for any of the SP content areas.

Table 7.2.4.1.6.1 Descriptive statistics for first and final attempt item measures in each SP content area

	n	Mean	SEM	SD	Min.	Median	Max.
Combinations							
First	6	-0.13	0.59	1.44	-0.18	-0.33	1.80
Final	6	-1.04	0.64	1.57	-0.77	-0.51	0.92
Interpreting Data							
First	22	-1.40	0.34	1.61	-1.57	-1.06	2.20
Final	22	-1.57	0.32	1.51	-1.29	-1.85	1.92
Probability							
First	12	-1.22	0.39	1.36	-1.44	-1.46	1.71
Final	12	-2.33	0.46	1.60	-2.05	-2.80	0.76
Statistics							
First	13	-1.37	0.45	1.64	-1.31	-1.88	0.75
Final	13	-1.55	0.37	1.33	-1.40	-1.74	1.08

Separate scatterplots were produced, and the analysis of the first and final attempt for SP content showed variation between each topic. Specifically, 5 of the 6 Combinations items (83%) decreased in item difficulty (Figure 7.2.4.1.6A), 18 of the 29 Interpreting Data items (62%) decreased in item difficulty (Figure 7.2.4.1.6B), all 12 Probability items (100%) decreased in item difficulty (Figure 7.2.4.1.6C), and only 7 of the 13 (54%) Statistics items decreased in item difficulty (Figure 7.2.4.1.6D). In summary, there was variation between item difficulty changes from first and final attempts in the SP content areas. The greatest proportion of items that students became more capable of correctly answering was for Probability, followed by Combinations, then Interpreting Data. The greatest proportion of items students become less capable of correctly answering was Statistics, indicating that misconceptions may have been developed most in this content area.

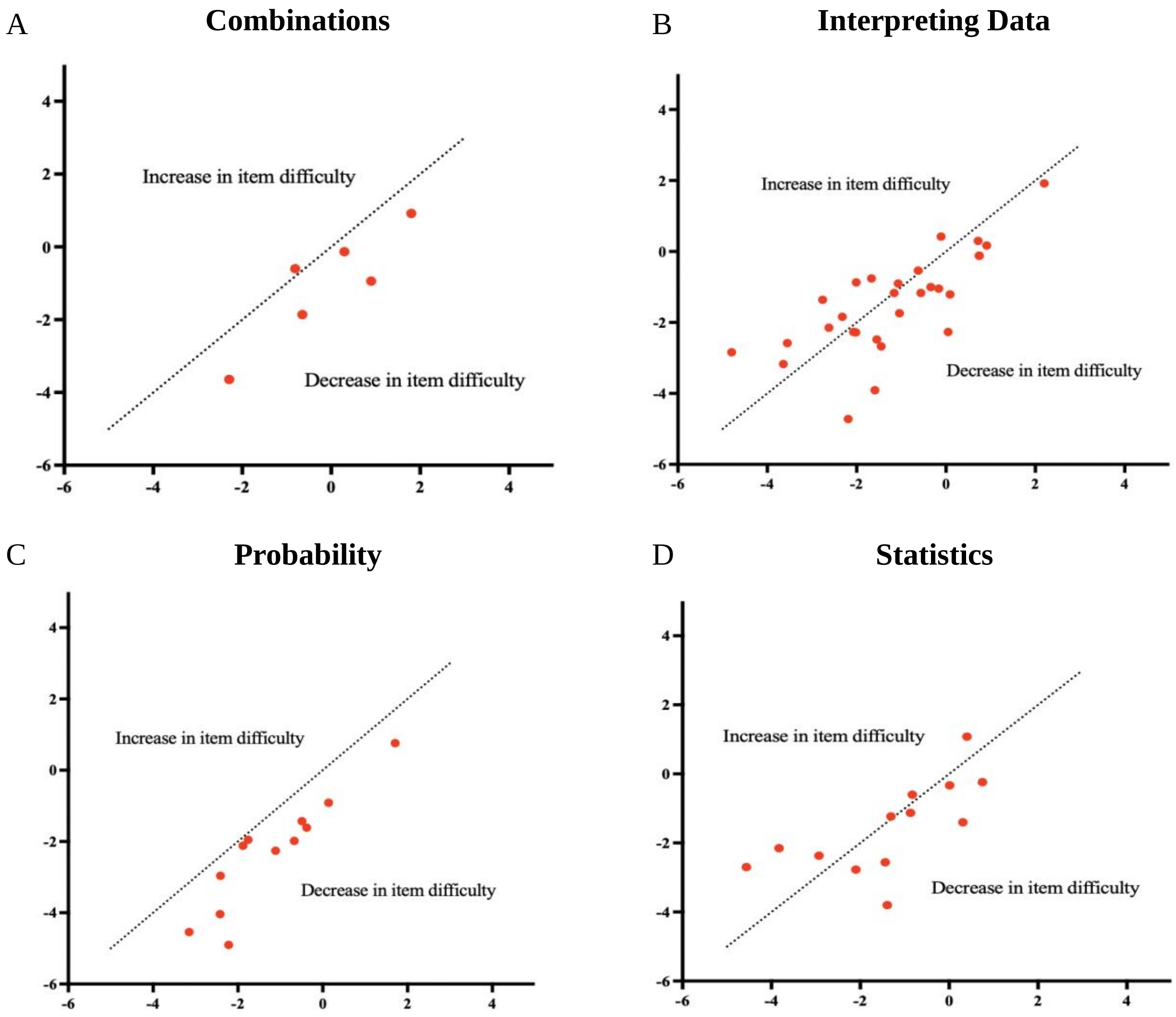


Figure 7.2.4.1.6.1 Scatterplot of first attempt versus final attempt item measures for SP content areas. Scatterplot of first attempt item measures (x-axis) versus final attempt item measures (y-axis) from racked analysis for SP content items (A) Data represent Combinations items (n = 6). (B) Data represent Interpreting Data items (n = 29). (C) Data represent Probability items (n = 12). (D) Data represent Statistics items (n = 13). The line (y = x) indicates no change in item measure, items above the line represent items estimated as more difficult on final attempt while items below the line represent items estimated as less difficult on final attempt.

7.2.4.1.7 Change in Item-Type Item Difficulties

Further analysis was then performed to determine if any links existed between item difficulty changes in the different item types (Fill-in-the-Blank, Multiple-Choice, and True/False). Table 7.4.2.1.7 displays the descriptive statistics for first and final attempts item measures for each item type. There was a statistically significant decline in all item types (p<0.0001 for Fill-in-the-Blank and Multiple-Choice items, and p <0.05 for True/False items). Overall, these results show that the items in all item types were found to be considerably easier on the final attempt than on the first attempt.

Table 7.2.4.1.7.1 Descriptive statistics for first and final attempt item measures in each item type

	n	Mean	SEM	SD	Min.	Median	Max.
Fill-in-the-Blank							
First	74	-0.83	0.16	1.34	-3.83	-0.81	2.20
Final	74	-1.52	0.17	1.44	-4.95	-1.41	1.92
Multiple-Choice							
First	183	-1.45	0.10	1.37	-4.95	-1.29	2.73
Final	183	-1.99	0.10	1.31	-5.25	-1.96	1.44
True/False							
First	17	-1.29	0.30	1.22	0.91	-1.23	0.91
Final	17	-1.96	0.41	1.68	1.08	-1.91	1.80

Changes in item difficulties from first to final attempts were then explored individually and separated into item types to determine whether improvements could be identified in certain item types more than others. Separate scatterplots were produced, and variations between item types were evident. Specifically, 57 of the 74 Fill-in-the-Blank items (77%) decreased in difficulty measure (Figure 7.2.4.1.7A), 122 out of 183 Multiple-Choice items (66%) decreased in difficulty measure (Figure 7.2.4.1.7B), and 13 out of 17 True/False items (76%) decreased in difficulty measure (Figure 7.2.4.1.7C). These results suggest that the greatest proportion of items that students became more capable of correctly answering was for Fill-in-the-Blank items and the greatest proportion of items students become less capable of correctly answering was for Multiple-Choice items.

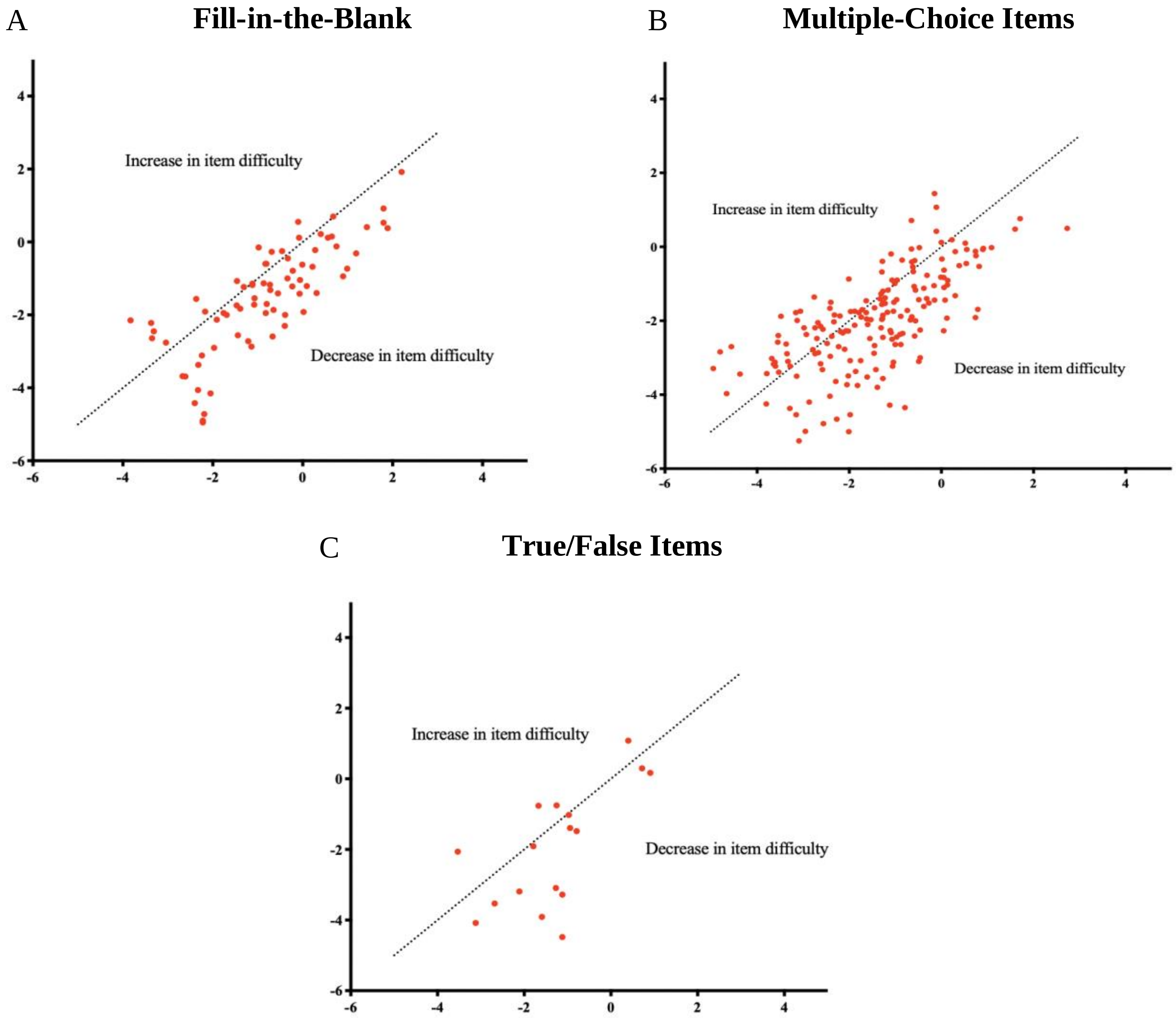


Figure 7.2.4.1.7.1 Scatterplot of first attempt versus final attempt item measures for item type. Scatterplot of first attempt item measures (x-axis) versus final attempt item measures (y-axis) from racked analysis in item types. (A) Data represent Fill-in-the-Blank items (n=74). (B) Data represent Multiple-Choice items (n=183). (C) Data represent True/False items (n=17). The line (y = x) indicates no change in item measure, items above the line represent items estimated as more difficult on final attempt while items below the line represent items estimated as less difficult on final attempt.

7.2.4.1.8 Change in Context-Domain Item Difficulties

To determine whether improvements could be identified in certain context domains more than others, descriptive statistics were calculated and displayed in Table 7.2.4.1.8 for first and final attempt item measures for Personal and Community items, Workplace and Employment items, and Education and Training items. Between the first and final attempts, the mean item difficulty measure decreased for all context domains. There was a statistically significant decline for Personal and Community items, and Workplace and Employment items (*p* <0.0001) but no statistically significant decline for Education and Training items.

Table 7.2.4.1.8.1 Descriptive statistics for first and final attempt item measures in each context domain

	n	Mean	SEM	SD	Min.	Median	Max.
Personal & Community							
First	184	-1.29	0.10	1.41	-4.95	-1.24	2.73
Final	184	-1.88	0.10	1.40	-5.00	-1.88	6.08
Workplace & Employment							
First	68	-1.28	0.15	1.24	-4.37	-1.22	1.19
Final	68	-1.85	0.16	1.34	-5.25	-1.93	1.44
Education & Training							
First	20	-1.11	0.36	1.60	-4.80	-0.99	2.20
Final	20	-1.40	0.29	1.29	-3.53	-1.38	1.92

Changes in item difficulties from first to final attempts were then explored individually and separated into context domains to determine whether improvements could be identified in certain context domains more than others. Separate scatterplots were then produced, and little variation was evident between each context domain. There was a decrease in difficulty measure for 129 out of 184 Personal and Community items (70%) (Figure 7.2.4.1.8A), 48 out of 68 Workplace and Employment items (71%) (Figure 7.2.4.1.8B), and 14 out of 20 Education and Training items (70%) (Figure 7.2.4.1.8C). In summary, all context areas showed a similar number of items that students became more capable of correctly answering.

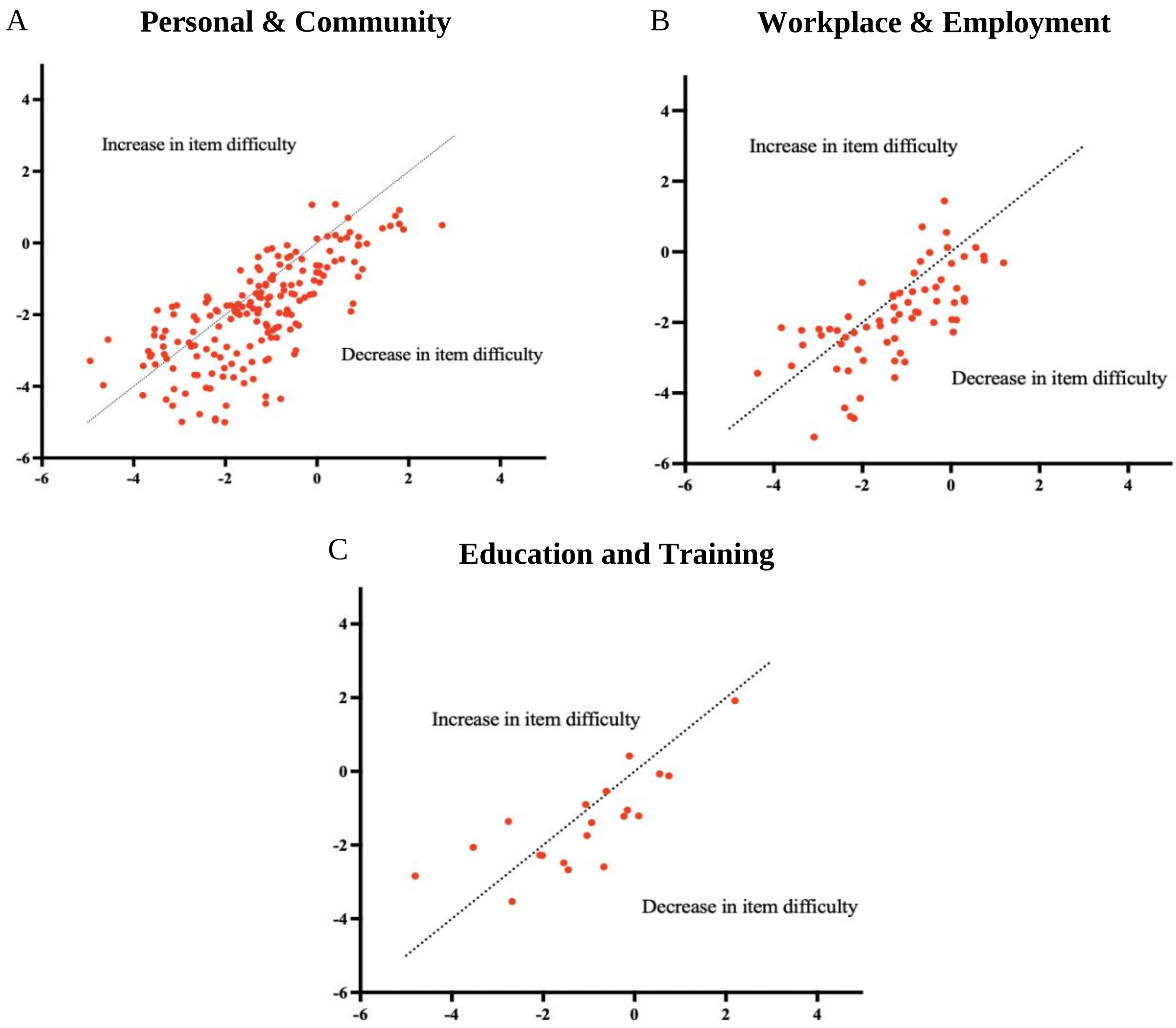


Figure 7.2.4.1.8.1 Scatterplot of first attempt versus final attempt item measures for context domains. Scatterplot of first attempt item measures (x-axis) versus final attempt item measures (y-axis) from racked analysis for items in context domains. (A) Data represent Personal & Community items (n=184). (B) Data represent Workplace and Employment items (n=68). (C) Data represent Education and Training items (n=20). The line (y = x) indicates no change in item measure, items above the line represent items estimated as more difficult on final attempt while items below the line represent items estimated as less difficult on final attempt.

7.2.4.1.9 Change in ACSF-Level Item Difficulties

To determine whether improvements could be identified in certain ACSF Levels more than others, descriptive statistics were calculated and displayed in Table 7.2.4.1.9 for first and final attempt item measures for Level 2, Level 3, Level 4, and Level 5 items. Between the first and final attempts, the mean item difficulty measure decreased for all ACSF Levels. However, a statistically significant decline was only evident for Level 3 and Level 4 items (*p* <0.0001). There was no statistically significant decline evident for Level 2 and Level 5 items.

Table 7.2.4.1.9.1 Descriptive statistics for first and final attempt item measures in each ACSF level

	n	Mean	SEM	SD	Min.	Median	Max.
Level 2							
First	19	-2.69	0.27	1.19	-4.95	-2.95	-0.46
Final	19	-3.28	0.25	1.11	-4.99	-3.23	-1.43
Level 3							
First	157	-1.84	0.09	1.07	-4.80	-1.79	0.74
Final	157	-2.45	0.08	1.03	-5.25	-2.32	-0.22
Level 4							
First	86	-0.22	0.08	0.76	-1.76	-0.27	1.80
Final	86	-0.75	0.10	0.91	-3.56	-0.65	1.44
Level 5							
First	10	1.22	0.42	1.31	-1.79	1.66	2.73
Final	10	0.44	0.31	0.99	-1.91	0.49	1.92

Changes in item difficulties from first to final attempts were then explored individually to determine whether improvements could be identified in certain levels more than others. Separate scatterplots were produced, and variations were evident between the levels. There was a decrease in difficulty measure for 11 out of 19 Level 2 items (58%) (Figure 7.2.4.1.9A), 110 out of 157 Level 3 items (70%) (Figure 7.2.4.1.9B), 61 out of 86 Level 4 items (71%) (Figure 7.2.4.1.9C), and 9 out of 10 Level 5 items (90%) (Figure 7.2.4.1.9D). In summary, the greatest proportion of items that students become more capable of correctly answering was at Level 5, followed by Level 4, then Level 3. The greatest proportion of items students become less capable of correctly answering was at Level 2.

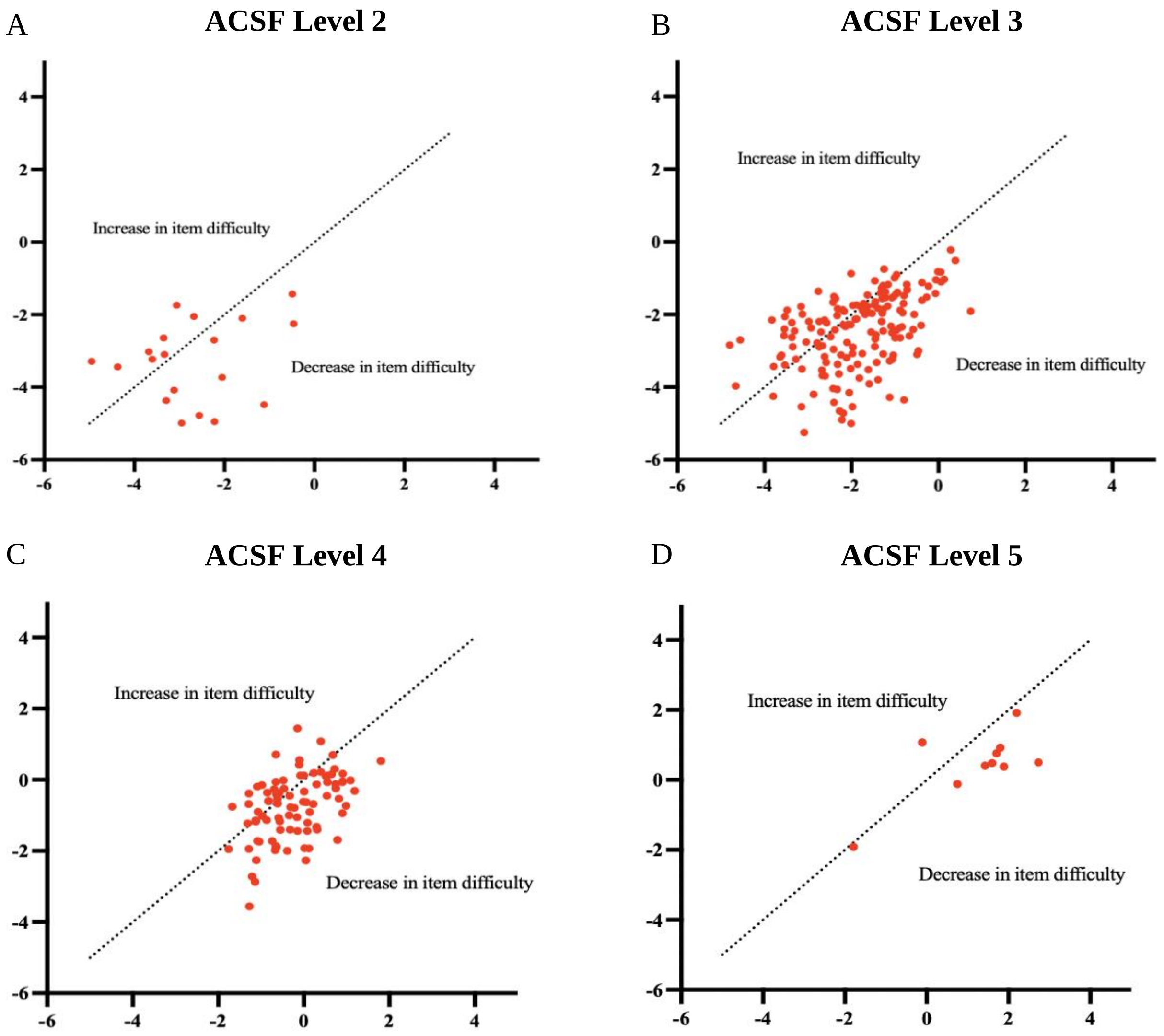


Figure 7.2.4.1.9.1 Scatterplot of first attempt versus final attempt item measures for ACSF levels. Scatterplot of first attempt item measures (x-axis) versus final attempt item measures (y-axis) from racked analysis for items in ACSF Levels. (A) Data represent Level 2 items (n=19). (B) Data represent Level 3 items (n=157). (C) Data represent Level 4 items (n=86). (D) Data represent Level 5 items (n=10). The line (y = x) indicates no change in item measure, items above the line represent items estimated as more difficult on final attempt while items below the line represent items estimated as less difficult on final attempt.

7.3 Discussion

In the previous chapter, we assessed raw scores to evaluate TES' performance and explored trends in their strengths and weaknesses. However, the test was specifically designed to support students to practise and learn mathematics and numeracy skills. Indeed, a large number of students repeated the test as anticipated. When all attempts were combined from both institutions, 43% of students were seen to have attempted the test more than once, 25% of students attempted the test three or more times, 10% attempted the test five or more times, and 3% attempted the test 10 or more times. Therefore, further evaluation of TES' numeracy capabilities was examined to determine changes in performance over time. This included assessing performance between attempts made by TES and evaluating the individual changes in performance between first and final attempts in mathematical strands, content areas, item types, context domains, and ACSF Levels. More specifically, the changes in these capabilities between first and final test attempts were explored to study any common areas of improvement or areas that appeared to be problematic.

In order to compare performance between attempts, normalising scores was necessary. This was especially crucial because the randomness of the test meant that each test had a different level of difficulty. In addition, students may have been presented with the same questions on repeated attempts and therefore may have remembered the answers or the processes to reach some solutions. In fact, these elements also display a positive effect of the test considering the promotion of learning involved. For example, if a student were to remember the steps to solve problems and apply newly-acquired skills when re-attempting the test, this would be considered best practice in mathematics. However, it is essential to note the unlikeliness of students being presented with the same questions on repeated attempts, as the repeated questions would likely only comprise a very small proportion of the overall test. This is due to the large pool of 272 questions drawn upon for each 40-question test.

Nevertheless, the possibility that students were presented with repeated questions, combined with the unequal difficulties across all test items, required the results to be normalised. More specifically, Rasch analysis allowed raw test scores to be expressed in terms of performance on a linear scale that accounted for the unequal difficulties across all test items and the possibility of repeated items. Using this Rasch Model allowed for an estimate of the person’s ability based on the items they attempted by making comparisons to other persons and items (Bond & Fox, 2015). This chapter specifically explored changes in performance by applying the Rasch Measurement Model that allowed for estimates to be made of TES’ ability based on the items they attempted so that comparisons could be made between test attempts. Similarly, item measures were calculated based on the performance of all attempts on that item, which allowed us to determine item difficulties and explore changes in difficulty measures over time.

Firstly, in our assessment of the performance of all test attempts from initial Rasch calculations, TES displayed a mean (±SEM) logit measure of 1.50±0.03. The minimum attempt measure was -2.04, and the maximum attempt measure was 5.87. When item measures were calculated to estimate the difficulty of each question, the minimum item measure was -3.42 logits, and the maximum item measure was 3.52 logits. These results indicate that abilities were more spread than the difficulty of the questions.

Furthermore, our findings indicate that many TES were motivated to practise and improve their numeracy skills. This is evident through the repeated attempts made by many TES. Specifically, more than 50% of the students attempted the test more than once, more than 25% attempted the test three times, and approximately 10% of students attempted the test five or more times. A possible explanation for the motivation to re-attempt the test is that the feedback provided by worked solutions helped TES learn the content. This feedback improved TES’ knowledge and confidence, thus motivating them to try again. Overall, these findings suggest that TES benefited from the opportunities provided to learn, practise, and demonstrate their understanding through repeated test attempts.

On examination of TES’ performance on repeated test attempts, overall improvement was evident at both institutions. The highest mean logit score at Institution A was evident at Attempt Number 13 (n=4). At Institution B, the highest mean logit score was apparent at Attempt Number 10 (n=4). When all results were combined, the highest mean logit score was evident at Attempt Number 10 (n=12). These results suggest that students’ ability improved on subsequent test attempts, and the similarities between the institutions indicates that the results might be generalisable. These findings support the research of Bangert-Drowns et al. (1991a) who reported that performance positively correlated with assessment frequency. In other words, performance improved as the frequency of assessments increased. However, Bangert-Drowns et al. (1991a) noted that improvements were incrementally smaller with each test added which differed from our results which saw the greatest improvement in performance between Attempt 7 and Attempt 8. Many possible reasons explain the improvements with more repeated attempts we saw on our Diagnostic Test. For example, it is possible that the TES practised their skills between attempts, increased their confidence each time they attempted the test, or learned from the feedback provided. Interestingly, in addition to improving performance, Bangert-Drowns et al. (1991a) also found that testing more frequently also made students' attitudes towards learning more positive. This reasoning may also explain the performance improvements we saw on repeated attempts, as there may have also been an increase in motivation for the students to learn.

Performance improvements were also evident when comparing between TES' first and final test attempts. It is important to note that a final attempt may have been the second attempt in some cases, and in other instances, a final attempt was Attempt Number 3, 4, or even 13. In making this comparison, our findings showed an increase in mean measures and a significant difference between first and final attempts at both institutions (p<0.0001). More specifically, there was a positive ability gain of 0.64 logits at Institution A, 0.61 logits at Institution B, and 0.63 when results were combined. There was also an increase evident in both institutions' minimum, maximum, and median measures. Overall, improvements were observed for 79% of all students (78% from Institution A and 81% from Institution B). We also saw that 28% of all students displayed a significant improvement of more than one logit. These results suggest that TES' numeracy abilities considerably improved between their first and final test attempts. This finding is in line with the findings of Afamasaga-Fuata'i et al. (2007), who reported in their longitudinal study that TES' mathematics and numeracy competency can be improved with dedicated numeracy training.

An explanation for TES' improvement between their first and final attempt could be due to the test acting as a numeracy intervention, which supported TES' skill development. It is possible that the intervention of our Diagnostic Test may have encouraged students to source specific learning resources or tutoring which assisted in developing their skills. Alternatively, it must be considered that the approach of using the test to build skills may have been a successful intervention strategy alone. In fact, interventions have previously been shown to improve students’ numeracy performance (Sellings et al., 2018). Sellings et al. (2018) showed that performance between two numeracy tests increased after an intervention had been made.

It is important to acknowledge that our findings also showed that 21% of TES displayed a regression in performance between first and final test attempts. This finding is similar to the study of Afamasaga-Fuata’i et al. (2007), who reported that 26% of TES displayed a regression in performance. Furthermore, our findings are also in line with the work of Sellings et al. (2018), who found that 6% of TES displayed lower numeracy capabilities on subsequent tests even after specific numeracy interventions. It has been suggested that the regression in performance could be due to the testing conditions, which might have caused test anxiety or mathematics anxiety (Boaler, 2014; Finlayson, 2014; Sellings et al., 2018). However, this is unlikely to have been the primary cause in our study due to the test's voluntary nature. TES were free to take the test whenever and wherever, there were no grades associated with the test, and the test was not linked to any coursework. Instead, it is possible that TES who demonstrated regression did not take a final test attempt seriously; however, considering our purposive sampling this is also not likely the case. Therefore, it must be considered that TES may have developed misconceptions between first and final attempts in some areas.

To explore this possibility of the development of misconception that could explain regression in numeracy performance seen for some TES, we explored the changes in individual item difficulty between first and final attempts. This method allowed for observations of changes in item difficulty measures, to identify changes in ability on those items. We intended to provide some possible explanations for poor numeracy performance that may have contributed to the decline in capabilities that were evident and also identify specific areas that appeared to be most problematic.

Firstly, item difficulty measures decreased for 70% of the items between first and final attempts, meaning that students became more capable of correctly answering most items. For the remaining 30% of items, students became less capable of correctly answering these items. Therefore, misconceptions may have been developed in some areas. To qualify this notion, the items were initially separated into their test sections (in mathematical strands for calculator-allowed and NC) for further exploration. Findings showed a similar number of items that students became more capable of correctly completing across all mathematical strands in the calculator-allowed section. This suggested a similar positive ability change in each strand. For example, in NA, 66% of items decreased in measure, 67% of MG items decreased in measure, and 70% of SP items decreased in measure. These findings indicated no particular strand in the calculator-allowed section displayed most improvements or suggested the most noticeable development of misconceptions. However, the greatest proportion of items that students became more capable of correctly completing was evident in the NC category, where 78% of the NC items decreased in item difficulty measures. Overall, these findings suggested that TES became more capable, mostly for NC items.

Changes in item measures were then explored in specific content areas to determine if improvements or regression in ability could be identified within certain content areas. We noted an interesting finding that all Algebra items displayed a decrease in item measure in the NA strand. This indicates that TES became more capable of correctly completing all Algebra items. This finding is in line with the study of Afamasaga-Fuata'i et al. (2007), who found common improvements in algebraic properties. Similarly, our findings also support the findings of Wilkie (2016), who found that teachers were readily able to improve their algebra content knowledge. Considering that Wilkie (2016) research specifically explored changes in primary teachers' algebra mathematical knowledge, it is interesting that similar results emerged in our research, which examined primary and secondary TES.

Another NA content area displaying a large number of items decreasing in item measure was Decimals. There were 13 out of the 16 (81%) Decimals items that showed a decrease in item measure, suggesting that TES became more capable of correctly answering most Decimals items. This finding was particularly encouraging because performance in the Decimals content area was one of the lowest in the NA strand in the analysis outlined in the previous chapter. Thus, our findings suggest that although Decimals may be a TES weakness, there is the potential for improvements to be made.

The greatest proportion of items that TES became less capable of correctly completing in the NA strand was in Financial Mathematics. More specifically, we found that 6 out of 15 items (40%) displayed an increase in item difficulty measure, suggesting that several misconceptions may have been developed in this area. This finding is particularly concerning considering the relevance of Financial Mathematics in everyday life. Applications of these skills are necessary for calculating costs, calculating price increases and discounts, and managing personal finances, such as calculating interest. Furthermore, Financial Mathematics is an essential numeracy skill to be implemented across the curriculum. For example, understanding money is a valuable skill that enables adults to make decisions about spending, budgeting, and saving money. Considering the necessity to carry out everyday tasks involving Financial Mathematics and the relevance of teaching school students these skills, it is concerning that our findings suggest that misconceptions can easily be developed in these concepts. To avoid these misconceptions being developed, strategies should be put in place to target teaching of this content area to enhance understanding.

In the MG strand, the greatest proportion of items that students became more capable of correctly completing was evident in Angles and Space, and Shapes and Symmetry. This is consistent with the findings from Afamasaga-Fuata'i et al. (2007), who also saw improvements in these areas. Our finding is also particularly significant considering the weaknesses we identified in these areas in the previous chapter. Further, these findings are also significant considering overall weaknesses that have been identified in other research in spatial reasoning skills that involve angles and shapes (Ozdemir & Goktepe Yildiz, 2015). Although overall spatial reasoning skills have been found to be weak, our findings in this chapter suggest that improvements can be made. Interestingly, the greatest positive ability change was evident in a question requiring the calculation of an angle on a straight line. This suggests that many TES may have misunderstood angle questions or do not remember basic angle facts, such as a straight line adds to 180 degrees.

In the MG strand, topics of Capacity and Volume and Distance and Perimeter displayed the greatest proportion of items that TES became less capable of correctly completing between first and final test attempts. There were 50% of the Volume and Capacity items that increased in difficulty measure and 43% of the Distance and Perimeter items. This is in line with the findings from Afamasaga-Fuata'i et al. (2007), who also found that students demonstrated persistent misconceptions in content areas involving measurement concepts. However, our results showed that the increased difficulty measure in Capacity and Volume items only displayed a very small increase (0.05-0.14 logits). Therefore, we did not consider that misconceptions in this content area needed to be explored further. For Distance and Perimeter, there were two items with considerable negative changes in ability displayed. Interestingly, these items were presented in a straightforward context with a very simple image presented. More specifically, the questions asked TES to find the perimeter of a rectangle or the missing side of a rectangle where they were given the perimeter. All other items in this content area included story contexts or more detailed graphical images. In fact, the greatest improvements were seen in items where both story contexts and visual maps were presented. This finding supports the research from Vappula and Clausen-May (2006), who explored the impacts of graphical images and diagrams on numeracy performance and found that diagrams improved performance in certain numeracy areas. In addition, the authors suggested that the results they found indicated that pictures provided a more powerful model to think with than real-life contexts.

For the SP strand, all items in Probability decreased in difficulty measure. This finding indicates that students became more capable of correctly completing all Probability items. Considering that previous studies have identified that TES struggle with probability concepts (Afamasaga-Fuata'i et al., 2007; Karatoprak et al., 2015), our result demonstrates that these skills can be improved. This is especially important considering that probability skills are required to identify, predict, and make decisions in life, and these skills must be implemented across the curriculum. For example, these skills are necessary for analysing risk-taking behaviours in Health and Physical Education and making predictions based on experimental evidence in science.

Upon further exploration of the Probability items that students were found to become more capable of answering correctly, it was evident that the most considerable change was evident on a question requiring an understanding of experimental probability. The difficulty measure of this item decreased from -2.22 to -4.90 logits. One possible explanation for this considerable improvement seen in this concept is that students can easily learn an appropriate approach. That is, this concept does not require any difficult processes or formulae to memorise. Instead, students may only need to practice one example of this type of question, which is enough for them to develop an appropriate approach. Considering the requirement of TES to teach this concept across the curriculum, and for their own students to understand the likeliness of something to occur based on experimental evidence, our finding is particularly encouraging. Our finding suggests that TES can improve their skills in this area to adequately teach these skills.

In the SP strand, there was a clear difference between the number of items decreasing in difficulty measure for Statistics and Interpreting Data compared to Probability and Combinations. More specifically, a much smaller number of decreased item difficulty measures were displayed for Statistics (54%) and Interpreting Data (62%) compared to Probability (100%) and Combinations (83%). This finding indicates that students became more capable of correctly answering questions involving concepts of chance compared to questions involving data. Again, this is likely due to the ease of learning the concepts involved with probability compared to learning concepts of statistical analyses and data interpretation. These findings are in contrast to the results reported by Afamasaga-Fuata'i et al. (2007), who identified Interpreting Data as a content area displaying considerable improvements made by TES. To explain this observation, Afamasaga-Fuata'i et al. (2007)highlighted that TES improved most on easier items than more difficult items in their study. However, the difficulty level of the Interpreting Data questions in our Diagnostic Test ranged from ACSF Level 3 to Level 5 and did not include Level 2 questions. The latter would have been considered easy questions. Therefore, the difference in our finding may be due to the difficulty of the questions used in our test.

The number of items that students became less capable of correctly answering between first and final attempts for Statistics (46%) and Interpreting Data (38%) present a concern because it suggests that misconceptions may have been developed in these concepts. This is a particular concern because all teachers are required to teach Statistics and Interpreting Data skills in all subject areas across the curriculum. For example, in the science curriculum, applications of mathematical data analyses are outlined in the content descriptions as part of the F-10 Curriculum. Additionally, skills involving data are outlined as part of the Information and Communication Technology (ICT) general capability, where selecting and evaluating data and information is outlined as a skill within the element of Investigating with ICT (ACARA, n.d.). More specifically, ACARA (n.d.) describes that by the end of Year 10, students should be able to evaluate the quality, suitability, and credibility of located data or information and sources. These requirements suggest that it is essential for TES to have adequate skills to interpret and analyse data that will allow them to teach these skills successfully. If TES are graduating with insufficient knowledge in these areas, their potential to be adequately teaching these skills will be limited.

Furthermore, the specific weaknesses we have identified in data analysis (Statistics) and Interpreting Data capabilities are important to address. This is particularly important since skills related to these concepts have been reported as a weakness in school students. More specifically, it has been reported that students struggle with skills to represent data, such as creating tables, charts, graphs, and sorting data using ICT (Phillips, 2015). In addition, Phillips reported that only 55% of Year 6 students achieved expected standards, and 52% of Year 10 students achieved expected standards in these skills. One explanation for these findings, provided by the author, was that teachers are not equipped with the skills they need to teach. Therefore, it is reasonable to suggest that teachers and TES need to improve their personal skills involving all data concepts to ensure they have adequate skills to implement these concepts into their teaching and improve students' abilities.

We also explored changes in item measures in the different item types and observed a statistically significant improvement in TES' ability in all item types. We observed that the least improved item type was Multiple-Choice. Only 66% of Multiple-Choice items decreased in item difficulty, which was less than Fill-in-the-Blank (77%) and True/False (76%). This finding suggested that TES were the least likely to improve in Multiple-Choice items compared to other item types. Although other studies have found that students performed lower on Multiple-Choice items compared to other item types (Abida et al., 2011), in contrast, our results in Chapter Six found that overall performance on Multiple-Choice items was best. Therefore, our research found that although Multiple-Choice was the best-performed item type, when errors were made, they were more likely to continue to be problematic.

It is reasonable to suggest that the misconceptions in the Multiple-Choice item type were likely due to the misconceptions developed in the content they assessed, as discussed above, and not due to the actual format of the questions. For example, one of the Multiple-Choice items displaying a considerable increase in item difficulty measure between first and final attempts was for a question in the Interpreting Data content area. Considering we identified this content area as an area displaying considerable decline, this may be a possible explanation. However, it must be considered that there may also be an item type explanation for the declining performance. For example, it is possible that TES may increasingly over-think

Multiple-Choice questions in repeated attempts. Understanding the cause of this observation requires further research.

When we explored changes in item measures in the different context domains, there was no difference observed in the number of items that students became more capable of correctly answering in the different context domains. In other words, similar improvements were seen across all context domains. This finding is consistent with the research of Vappula and Clausen-May (2006), who found that story context did not affect performance. However, as previously outlined, Vappula and Clausen-May also found that diagrams as context improved student performance. We also identified this as a possible factor when we explored changes in performance in Distance and Perimeter questions. Therefore, how question context is presented (either as a written story or diagram) and its influence on learning and assessment would be interesting to explore in future work.

When we explored the changes in item measures in the different ACSF Levels, we found a statistically significant decline in item measures only for Level 3 and Level 4 items ($p<0.0001$), but not for Level 2 or Level 5. However, when observing changes in individual item measures, we saw the greatest proportion of item measures decreased for Level 5 items, and the least number was seen for Level 2 items. These results suggest that TES became more capable of correctly answering difficult questions (Level 5) and showed the least improvement for easier questions (Level 2). In contrast to previous studies that reported improvement made by TES are only possible for easy items (Afamasaga-Fuata'i et al. (2007), our findings suggest that improvements can be made on the most difficult items. Further, our results specifically showed a large number of items that students became more capable of correctly answering for Level 3 and Level 4. In Chapter Six, we identified that TES numeracy capabilities are required to be in the range of Level 3 and Level 4. We made this determination because of the proportion of these items that are assessed in the LANTITE. Therefore, we considered these levels to be a representation of the skills expected. Thus, our findings suggest that TES can make considerable improvements at the levels they are most required to possess.

In summary, the results in this chapter support the need to evaluate TES' numeracy capabilities and explore where improvement can be made using an online Diagnostic Test. Firstly, it was evident that overall, students performed better on the test with repeated use.

Secondly, we saw the TES improve in most areas as improvements were made in all test categories (NA, MG, SP, NC), content areas, item types, context domains, and ACSF Levels. Overall, our findings suggest that TES benefit from being be given opportunities to learn, practise, and demonstrate their understanding of skills through repeated use of the test.

Chapter Eight: Discussion and Conclusion

8.1 Introduction

This chapter addresses the research questions and provides possible explanations for the results obtained. The implications of the research findings, highlighting the utility of an online diagnostic test as a form of A*f*L to support TES' numeracy development is also discussed. Finally, recommendations are provided with suggestions for future initiatives and research to ensure the successful development of TES numeracy skills. This is especially important considering it is imperative that TES have adequate numeracy capabilities before they become teachers.

8.2 Addressing the Research Questions

This study sought to answer the research question: *To what extent can TES' numeracy skills be evaluated and improved using an online diagnostic test?* To address this research question in detail, three sub-questions were explored. These sub-questions are:

1. To what extent can an online test be used to diagnose TES' numeracy capabilities?
2. What are TES' mathematical strengths and weaknesses?
3. To what extent can an online diagnostic test be used to improve TES' numeracy skills?

These research questions are individually addressed below.

8.2.1 To What Extent can an Online Test be used to Diagnose TES' Numeracy Capabilities?

To address this research question, the level of detail that an online diagnostic test was able to garner about TES' numeracy capabilities was assessed. Findings revealed that TES' numeracy capabilities could be examined by mathematical categories, topics, item types, contexts domains, and ACSF levels through the assessment of raw scores. Furthermore, findings revealed that capabilities could be diagnosed on individual test attempts, including separately on first and final attempts to allow for accurate comparisons.

Careful and thoughtful development of the test items, which included several iterative processes conducted by the researcher and research assistants, and the specific design of the test, played a significant role in what we could diagnose about TES' numeracy skills. It was initially sort to identify overall capabilities and capabilities in test sections, mathematical strands, and content areas, and these agenda were successfully answered and skills were evaluated.

More specifically, our raw score analysis produced an accurate diagnosis of capabilities in sections (Calculator-allowed and NC), in strands (NA, MG, and SP), in NA content areas (Algebra, Basic Arithmetic, Decimals, Financial Mathematics, Fractions, Percentages, and Rates & Ratios), in MG content areas (Angles, Area, Capacity & Volume, Distance & Perimeter, Estimating, Reading & Converting, Space, Shapes & Symmetry, and Time & Timetabling), and in SP content areas (Combinations, Interpreting Data, Probability, and Statistics). Additionally, the raw score analysis diagnosed performance under different question conditions; in item types (Fill-in-the-Blank, Multiple-Choice, and True/False), in context domains (Education & Training, Personal & Community, and Workplace & Employment), and ACSF Levels (Level 2, Level 3, Level 4, and Level 5). The Rasch analysis produced an accurate diagnosis of capabilities across different test attempts and identified capabilities separately on first and final attempts in test categories (NA, MG, SP, and NC), NA content areas, MG content areas, SP content areas, and also in the different item types, context domains, and ACSF levels.

Although the Diagnostic Test allowed for a thorough evaluation of skills, as outlined above, it is important to acknowledge that the specific test design may have produced results that are not a true indication of the capabilities some TES possess. For example, some students may have used a calculator in the NC section; therefore, abilities in the NC section may be lower than what was reported. Furthermore, TES may not have attempted the Diagnostic Test alone and may have had help. Therefore, again, some results may not be a true indication of an individual's capabilities. However, considering the large sample involved in the study and the voluntary nature of the test, it is most likely that TES completed the test as intended, and the results provided an accurate indication of capabilities.

Furthermore, the voluntary nature of the test suggests a limitation to the results we obtained. It is possible that attempts were only made by TES who felt confident to show their capabilities, and low-ability TES may have avoided attempting the test altogether. If this were the case, the extent of the information gathered by the test would have been impacted. However, despite their capabilities, many TES would likely have used to test in preparation for the LANTITE to inform their readiness or to use as practice. Therefore, we considered that our findings shone a light on a multitude of TES's numeracy capabilities representing all TES.

In summary, the Diagnostic Test allowed diagnosis of skills and identification of differences in capabilities displayed on first and final attempts. When combining the findings from both analyses, areas of strength and weakness could be determined, and areas of improvement and decline could be identified.

8.2.2 What are TES' Mathematical Strengths and Weaknesses?

Firstly, TES' capabilities were explored to identify strengths and weaknesses in each section of the test (Calculator-allowed and NC) from our raw score analysis. TES at both institutions performed best in the NC section compared to the Calculator-allowed section. When applying the Rasch Measurement Model to determine capabilities, results showed that students were most capable of correctly answering items in the NC section compared to the Calculator-allowed section, and this was seen on both first and final attempts.

Next, capabilities were explored within the strands of each test section. Within the Calculator-allowed section, performance in the NA and MG content strand were similar. However, students' performed marginally lower in the SP content strand at both institutions. We observed the same trend in our Rasch analysis, where students were estimated to be less capable of correctly answering SP items on both first and final attempts. In the NC section of the test, a statistically significant difference was evident between the strands (NA and MG) at both institutions; however, a conflict between the mean and median made it challenging to determine NC strengths or weaknesses.

Interestingly, the findings showed the best performance in the NC section, considering other research has found performance to be lowest in the NC section of the LANTITE (Hall &

Zmood, 2019). Further, the NC section was at the end of the Diagnostic Test, and it was expected for students' attention to decrease throughout the test. However, results in this section suggest that TES could sustain their attention and demonstrate their abilities. Furthermore, the test did not have timed conditions; unlike the LANTITE, therefore, the TES were given the opportunity to demonstrate their skills on all questions without a time limit. Accordingly, the findings suggest that TES' NC skills are quite good, and the test allowed them to demonstrate their skills accurately. This finding will be beneficial for other institutions planning on developing a numeracy diagnostic test. The findings suggest that it is essential for test developers to consider the intention of their test. For example, to diagnose capabilities, non-timed conditions may produce more accurate results. This is especially important if the purpose is also to improve skills. Alternatively, if the intention is to mirror conditions of the LANTITE, to predict performance on that test, timed conditions would be appropriate, despite the possibility of skills not being diagnosed accurately.

An explanation of the similarities found between performance in the NA and MG strands could be due the relationship between some of the skills assessed in the content areas in each of these strands. That is, there are some interconnections between many concepts in NA and MG. For example, NA Basic Arithmetic skills, applying the four operations, are needed to calculate Distance & Perimeter, Area, and Capacity & Volume in MG. Similarly, knowledge of Fractions and Decimals (NA) is necessary for Estimating, Reading & Converting (MG). Therefore, a strength or weaknesses in some NA skills may imply a strength or weakness in some MG skills. Alternatively, a possible explanation for the marginally lower performance in the SP strand may have been due to a significant weakness in a single SP content area that affected the overall strand performance. A single significant weaknesses would have affected the overall strand performance considering only four content SP areas were assessed compared to seven content areas that were assessed in each strand of NA and MG. This is most likely the explanation as a significant weakness was found in Combinations in the SP strand.

These explanations further emphasise that if other institutions were to develop a test, the specific test intentions must be carefully considered. For example, if the main purpose is to compare performance in strands, it may be beneficial to include more content areas in the SP strand to match the other strands of NA and MG. For example, additional SP content areas could be developed, such as Displaying Data, Collecting Data, and Analysing Data. Although it was sought to explore overall performance in strands, it was also an aim to more explicitly evaluate strengths and weaknesses in mathematical content areas within each strand.

In the NA content assessed, the raw score analysis found that overall, TES at both institutions displayed a strength for Algebra and Percentages and a weakness for Rates & Ratios. However, the Rasch analysis found that students displayed a strength for Algebra only on final attempts. In fact, on first attempts, Algebra was the most obvious weakness. Rates & Ratios was seen to be a weakness on both first attempts and final attempts. In the MG content areas, a shared strength at both institutions was for Distance & Perimeter. The Rasch analysis also found Distance & Perimeter a strength on first attempt and final attempts. However, on final attempts, students were even more capable of correctly answering Capacity & Volume items. A common weakness from the raw scores was seen in Angles and Space, Shapes & Symmetry, which was in line with our Rasch analysis results on both first and final attempts. In the SP content, a common strength from our raw scores was seen for Probability. However, this strength was only seen in the Rasch analysis on final attempts. It was noticed that students were most capable of correctly answering Statistics and Interpreting Data items on first attempts. The raw score analysis found a weakness for Combinations which we also saw as a weakness on first and final attempts from the Rasch analysis.

When comparing performance across all strands, overall TES mathematical strengths were identified for Algebra and Percentages. Explanations for these areas to be overall mathematical strengths of TES may be due to various strategies that can be used to approach questions in these content areas. For example, in Algebra, inspection skills and formal algebraic skills can be equally successful in solving algebraic equations and applying formulae. For Percentages, skills of fractions, decimals, and estimations can often all be applied. Therefore, considering that students have a range of strategies to draw upon, they are more likely to answer these concepts correctly. Furthermore, considering the application of Percentage skills required in the life of a TES (e.g., applying discounts and working out a percentage from a test result), it is not surprising that this area was identified as a strength.

Interestingly, the Rasch analysis identified similar overall strengths when exploring capabilities on first and final attempts separately. The analysis also found Percentages to be a strength on both first and final attempts; however, Algebra was only a strength on the final attempt. In fact, Algebra was found to be the most apparent weakness on first attempts. These results suggest that considerable improvements were made in this area that will be discussed in the next section.

When comparing performance at both institutions across all strands, overall mathematical weaknesses were demonstrated for Angles, Combinations, and Rates & Ratios. This was also consistent with our Rasch analysis findings that students were least capable of correctly answering items in these content areas on both first and final attempts. A possible explanation for these weaknesses is the specialised knowledge or rules required to understand and correctly answer questions on these concepts. For example, in Angles, knowledge of angle sums of each shape and other geometrical relationships is necessary to answer these questions accurately. If these are not known to the TES, there is no way of guessing an answer or working them out intuitively. Another example is in Combinations, where one method of working out the number of possible combinations is to write out every possible outcome and then count the number of outcomes written. However, this method easily results in missing one or more outcomes and therefore answering incorrectly. Alternatively, if a student knew and applied multiplication strategies, an accurate answer is much more likely to be reached. Considering the weaknesses found in these topics, it is justified to suggest that these specific rules and methods are unknown, and therefore TES' knowledge in these areas requires reinforcement.

For this research question, TES strengths and weaknesses were also explored when presented with different question types or questions in different context domains. Across both institutions, students performed best on Multiple-Choice items and worst on Fill-in-the-Blank items, consistent with results identified separately on first and final attempts. A possible explanation for this is the element of chance involved in correctly guessing Multiple-Choice items. Despite the higher probability of correctly guessing a True/False item, TES scored lower in these questions types than Multiple-Choice. Therefore, there may be other reasons why students performed best on Multiple-Choice items, possibly related to the content and not the item type.

Interestingly, there were no trends apparent in the context domains of the questions considering only minor variations between them were evident at each institution. This suggests that the context of numeracy questions does not affect TES performance. Similarly, our Rasch analysis showed minimal differences in mean item logit measures between the context domains on both first and final attempts. In our exploration of performance on items in each of the ACSF levels, unsurprisingly, a strength at both institutions was evident on Level 2 items, and a weakness was apparent on Level 5 items. This was consistent with the Rasch analysis that displayed that students were most capable of correctly answering Level 2 items and least capable on Level 5 on both first and final attempts.

8.2.3 To What Extent can an Online Diagnostic Test be used to Improve TES' Numeracy Skills?

Considering the benefits of online learning (Alcoholado et al., 2016), the accuracy online diagnostic tests provide in diagnosing skills (Linsell & Anakin, 2012), and the ability they have to be able to track learning and progress (Snekalatha et al., 2021), we sought to determine whether the Diagnostic Test used in this study can be used as a learning tool to improve TES' numeracy capabilities. The online Diagnostic Test was specifically designed using the A*f*L theory of Black and Wiliam (1998) with consideration of the elements reported to enhance learning through formative assessments. In particular, the test was developed to encourage self-assessment, response and reception, goal orientation, and self-perception. Additionally, the benefits of frequent assessments were considered, the TES were allowed to attempt the test multiple times, and TES were provided effective instant feedback after each test. Whether this is a viable way to improve TES' numeracy skills has not previously been studied, and we sought to address this gap.

This research question was specifically addressed by applying the Rasch Measurement Model to the data to convert raw scores to logit measures that account for the unequal difficulties across all test items and test attempts. Scores were analysed from TES who attempted the Diagnostic Test multiple times to evaluate performance trends. The generated item anchors were applied to compare performance on first and subsequent test attempts. Overall, TES' at both institutions improved ability on subsequent test attempts. Overall, results showed that improvements in numeracy ability had been made over time.

The Rasch method of stacking was then applied to compare first and final attempts only. Overall, there was a positive ability change at both institutions. The mean logit measures increased from 3.83 to 4.47 at Institution A, 4.25 to 4.86 at Institution B, and 3.97 to 4.60 when results were combined. Additionally, the comparison of TES' first and final attempts showed that approximately 80% of students displayed improvements between their first and final attempts (78% from Institution A and 81% from Institution B).

Furthermore, the Rasch method of racking produced item difficulty measures for first and final attempts, allowing us to compare the estimated ability changes on the items between attempts. This focused on the likelihood of anyone correctly responding to a question rather than focusing on individual students' performance. Thus, allowing for examination of overall areas of improvement or decline. Findings showed that the TES became more capable of answering 70% of the items, confirming that students displayed general gains in performance overall. This was also confirmed by the number of items displaying decreased item difficulty measures between first and final attempts in each category; NA (66%), MG (67%), SP (70%), and NC (78%). Interestingly, the greatest proportion of items students became more capable of correctly completing displayed in the NC category (78%), suggesting that improvements are most easily made on NC items. When comparing these results with the raw score analysis, our findings indicate that further improvements are possible despite overall skills already being highest in the NC category.

Furthermore, all content areas in each strand displayed several items that students became more capable of correctly completing; therefore, improvements were apparent in all content areas. However, to highlight the areas where students had improved their capabilities the most, the number of items that students became more capable of correctly answering in each content area were identified. The areas showing the most considerable improvements between first and final attempts were apparent for Algebra (100%), Probability (100%), Angles (86%), Space, Shapes & Symmetry (86%), and Combinations (83%). Overall, a possible explanation for the considerable improvements in some of these content areas extends upon the hypothesis provided earlier in this chapter. Once the specific rules required for these topics have been identified, learned, and understood, skills are easily transferred and applied to answer questions more accurately. For example, once TES identify angle sums of shapes, properties of shapes, or methods of accurately calculating the number of possible combinations, questions requiring applications of these skills become much more manageable.

In particular, two areas were seen to have all items decrease in item difficulty measures: Algebra and Probability. This suggests that students were estimated to have become more capable of correctly answering all items in these content areas. Therefore, it is likely that TES most successfully improved skills involved in these topics between test attempts. However, it must also be acknowledged that some concepts in these content areas can be quite complex, and a variety of skills are required to be successful on questions within these topics. Therefore, a reasonable explanation for the considerable improvements in these areas is that students may have undertaken targeted study between test attempts. This suggests that the Diagnostic Test allowed TES to self-assess their capabilities and self-identify their errors to make appropriate decisions about improving their skills.

The analysis also identified areas that had the least number of items that students became more capable of correctly answering; Capacity & Volume (50%), Statistics (54%), Distance & Perimeter (57%), Financial Mathematics (60%), and Interpreting Data (62%). Therefore, these content areas were found to have the most number of items that students became less capable of correctly answering. Although these results still suggest that improvements were apparent across these areas, there is also a suggestion that some regression in ability may have occurred. Reasons for this could be varied; however, it is possible that in an attempt for TES to improve skills, misconceptions were instead developed. This finding is extremely significant as identifying the areas where TES are developing misconceptions highlights the need for these areas to be appropriately addressed.

In addressing this research question, changes in item measures in each item type and context domain were also explored; and overall improvements were apparent. However, some differences in the number of items that students became more capable of correctly answering differed. For example, the greatest proportion of items that students became more capable of correctly answering was seen for Fill-in-the-Blank items (77%), and the least number was seen for Multiple-Choice items (66%). This is interesting considering that Multiple Choice was the best-performed item type in the raw score analysis. A rationale for this finding is that the least improvements are made in areas already well performed. This could also suggest that there are limits in the amount of progress that can be made in some areas. Vice versa, considerable improvements can be made in areas identified as weakest (e.g., Fill-in-the-Blank items).

When exploring improvements in the context domains, there were no obvious differences observed in the number of items that students became more capable of correctly answering in each context domain; Personal and Community (70%), Workplace and Employment (71%), Education and Training (70%). Therefore, similar improvements were seen across all context domains. This extends upon the finding in the raw score analysis, suggesting that TES performance is not affected by the context of the question and indicates that the context of the question also does not influence improvements. When exploring improvements in ACSF levels, improvements were evident at all levels. Interestingly, the most number of items that students became more capable of correctly answering was seen for Level 5 (90%) and the least number for Level 2 (58%). Additionally, findings showed a considerable number of items that students became more capable of correctly answering at Level 3 (70%) and Level 4 (71%). Considering Level 3 and Level 4 questions were developed at the standard expected of TES, the findings present some promising findings suggesting that TES can make considerable improvements at the levels they are most required to possess.

Overall, the analysis used to address this research question produced results that informed two main conclusions. Firstly, it was identified that some areas displayed many items that TES become less capable of correctly answering between first and final attempts. This conclusion is especially meaningful as it has highlighted the specific need to address this issue and target skills in certain areas to improve TES' numeracy capabilities. Secondly, since improvements were observed across all categories, content areas, item types, context domains, and ACSF levels, it can be concluded that the Diagnostic Test can be used as a successful learning tool and can be considered an effective A*f*L.

8.3 Implications of Research Findings and Recommendations

Although a variety of initiatives have already been implemented and evaluated by institutions in Australia (Callingham et al., 2015; Forgasz & Hall, 2019; Sellings et al., 2018)**,** the findings of this research support that a Diagnostic Test can be used as a valid A*f*L tool to successfully improve numeracy skills. Therefore, diagnostic tests could be implemented and used by other institutions as an initiative to enhance numeracy capabilities.

It is important to highlight and discuss the elements that were implemented into the test design that could help inform the development of a similar test. More specifically, elements of online learning models, and elements considered determinants of effective A*f*L according to the theory of Black and Wiliam (1998), were adopted. Firstly, the design of the test considered the overlapping elements identified between online learning models of TAM, ADDIE, and the e-learning systems framework, and Black and Wiliam's (1998) A*f*L theory. That is, the design carefully considered the users and participants, the intention of the test/technology used, and the services that needed to be involved. Secondly, the design specifically adopted the ADDIE Model phases in the development of the test to ensure best learning design practice.

Finally, the elements in Black and Wiliam's (1998) A*f*L theory were embedded to encourage learning through the use of the test. Firstly, the test was designed to encourage self-assessment by presenting TES with their overall score and correct or incorrect answers on individual items. This allowed for a self-evaluation of areas of weakness and strengths. A positive response and reception was also encouraged by displaying a 'Correct. Well done!" message and worked solutions for inaccurate answers. It was anticipated that this would motivate students to learn the correct methods or seek assistance. Thirdly, it was sought to encourage positive attitudes and goal orientation by naming the test; "Practice Test" and not making information about other students' or average results available. The frequency of assessments was also considered, and we allowed TES to make unlimited test attempts. Flexibility was allowed for TES to sit the test at any time as there were no restrictions (day or time), and this was expected to enhance motivation and learning. Most importantly, instant feedback was implemented into the test design through worked solutions presented at the end of a test attempt for incorrect answers. The intention was to allow TES to promptly identify the gap between their current achievement level and the standard required to master a particular skill.

This feedback element was of particular interest in this research, as it was the only intervention we provided to students. More specifically, automatic feedback was presented to students through the online Diagnostic Test in two different ways. Firstly, students were given feedback that included their overall mark out of 40, and correct or incorrect feedback was provided for each question. This allowed students to gain an overall view of how they performed and an indication of their success for each question. In addition, this feedback would have allowed students to notice their strengths and weaknesses; therefore, informing them of what they need to work on to improve. Secondly, feedback was provided in the way of worked solutions for questions that were answered incorrectly. This feedback was presented immediately after a test sitting and allowed students to examine methods of approaching the questions they answered incorrectly. Therefore, they were given the opportunity to gain new knowledge through examples of working.

A major aspect of the feedback provided was that it was presented immediately. Specifically, the immediacy of task feedback has been found to improve mathematical learning (Alcoholado et al., 2016) and timely feedback has been considered especially important in tests (Wiggins, 1993). Therefore our findings suggest that the instant feedback we provided may have played a significant role in the mathematical improvements we saw between first and final attempts. Furthermore, considering the improvements that were seen through the use of the test, it is reasonable to suggest that the overall method of feedback was effective and likely a main determinant of the effectiveness of the A*f*L tool. Firstly, the feedback allowed students to examine their strengths and weaknesses and explore correct methods of answering questions to gain new knowledge. Secondly, the students were given the opportunity to act on the feedback immediately. This finding supports the literature that widely discusses feedback as a main determinant of the effectiveness of formative assessment (Yorke, 2003).

There were many ways that TES may have acted on the feedback provided in the Diagnostic Test to improve their numeracy capabilities. For example, TES may have noted the topics of questions they answered incorrectly and sought external help in accessing learning resources between test attempts (e.g., private tutoring). Another possibility is that students may have simply read the worked solutions, learned the required methods, and were then able to appropriately transfer and apply their new skills in other questions on subsequent test attempts. In this case, it is possible that the worked solutions were enough to assist in developing skills. Therefore, the Diagnostic Test alone provided adequate learning opportunities to enhance learning.

It is important to note that the factors that led to the TES' improvements; were not directed assessed, therefore, it cannot be confirmed or denied that there was any specific influence of any of the factors. However, it can be concluded that the overall incorporation of the elements from Black and Wiliam's (1998) theory successfully encouraged student learning through the use of the test. Therefore, in the future development of a similar test, it would be recommended for elements to be incorporated similarly. However, there are modifications that may be considered in the development of a future A*f*L diagnostic test to promote further improvements. For example, Black and Wiliam (1998) explored Butler's (1988) work and found that giving grades could impair the help of positive feedback comments. Thus, it could be reasonable to consider not providing students with their overall marks. This could encourage the TES to pay more attention to the other, more useful feedback they receive.

Furthermore, through the exploration of the studies of Cameron and Pierce (1994) and Kluger and DeNisi (1996), Black and Wiliam (1998) also concluded that providing feedback that draws attention towards self-esteem, such as giving praise, can have a negative effect. Thus, another consideration in developing a future diagnostic test could be to display worked solutions for all questions instead of presenting students with a message of praise for correct responses. It is possible that through our approach, we may not have seen the extent to which the students could have improved because worked solution feedback was only given on incorrect responses. This could have been an issue, in particular, if students correctly guessed an answer. For these items, they would have only been given the "Correct, Well Done!" message of praise and would not have been prompted by the worked solution feedback to self-evaluate their work. Therefore, the feedback provided in the way of praise may not have been effective. In summary, it could be possible to observe further overall TES improvement by providing effective feedback on all answers, correct or incorrect.

According to Tan's (2013) triangulated model theory, three elements: task-dependency, time-dependency, and feedback, must be balanced to support student learning. For our work, the test section, strand, topic, type, context, and ACSF level can be considered as the task-dependent axis of the model. That is, factors that influence the level of difficulty that TES experience. Time spent on each attempt and time between attempts can be considered the model's time-dependent axis. These axes are bridged from feedback in the Diagnostic Test in two ways: feedback from test scores and feedback provided in worked solutions for incorrect responses. Through this lens, students that demonstrated an overall improvement can be considered as having an appropriate balance of all three elements. In contrast, students who showed no improvement or a decline in performance can be viewed as having an imbalance in one or more axes. For example, a deficit in the task-dependency axis may indicate a weakness in a specific test section, strand, topic, item type, context, or ACSF level and could be addressed by providing targeted interventions to improve skills in the identified weakness area. A deficit in the time-dependency axis could mean that a student did not allow adequate time to enhance knowledge in between attempts and could be addressed by incorporating a restriction of when a student may sit a subsequent attempt, ensuring adequate time for learning. A deficit in the feedback axis would mean that a student may not have acted appropriately on the feedback provided. They may have only looked at their overall results, thus, disregarding the more specific feedback they received. Alternatively, they may not have understood the worked solutions they were presented with.

There are implications to each of these items, precisely due to the online nature of the Diagnostic Test. More specifically, considering there was no researcher-student interaction, it was challenging to determine whether the students were able to accurately identify the areas they needed to improve upon, to determine what motivated them to re-attempt the test, or to determine whether they fully understood the specific feedback that was presented to them. To address this gap in our knowledge, it could be beneficial to conduct consultations with students to gather information about the knowledge they gained from the test, their understanding of the feedback provided to them, and what determined their decision to re-attempt the test.

Furthermore, given the two possible outcomes produced in this research when exploring comparisons between first and final attempts (i.e., improvement in ability or no improvement in ability), this study showed that it might be possible to view students' learning through the angle of inclination. In particular, the data from this research could be used to determine ideal proportions of the vertical and horizontal axes, displaying learning gaps and the time required to act on closing those gaps, respectively, in a triangulated model that extends upon Tan's (2013) model. More specifically, the design of the Diagnostic Test and the results produced in this research could inform definitions of the axes in a new triangulated model (Figure 8.3.1) where the horizontal axis represents the time between a student's first attempt and subsequent attempt (time-dependency) and the vertical axis represents the learning gaps that are identified (task-dependency). The third component depicts the action taken on feedback and is represented by the trajectory incline. In addition, and determined by the proportions of the axes, the angle of inclination could be determined.

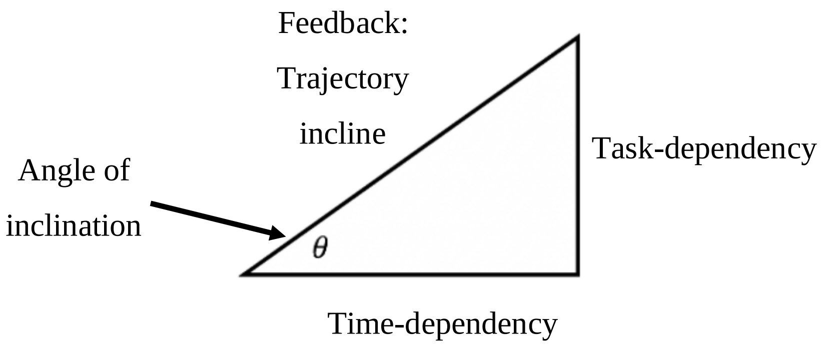


Figure 8.3.1 Extension to Tan's (2013) triangulated model of AfL

The angle of inclination would be of particular interest in this model. Considering the results of this research show that some students improved performance while others did not, it would be beneficial to observe the angles of inclination produced in each case using the proposed extended triangulated model.

For example, consider a student who identified many gaps in their learning on the first attempt, took a long time to act on the feedback provided before attempting the test again, and then improved their performance on the final attempt. A triangle could be constructed displaying the proportions of these components with an angle of inclination that is sufficient considering that improvements were made (Figure 8.3.2A). Alternately, consider a student who was to self-identify minimal learning gaps on the first attempt, did not require much time to act on the feedback provided, and attempted the test again quite soon. If this student also displayed improvements, the angle of inclination produced in the triangle would also be considered sufficient (Figure 8.3.2B). Therefore, although different proportions of the axes would be produced in these two examples, the angles of inclination observed would, in fact, be similar.

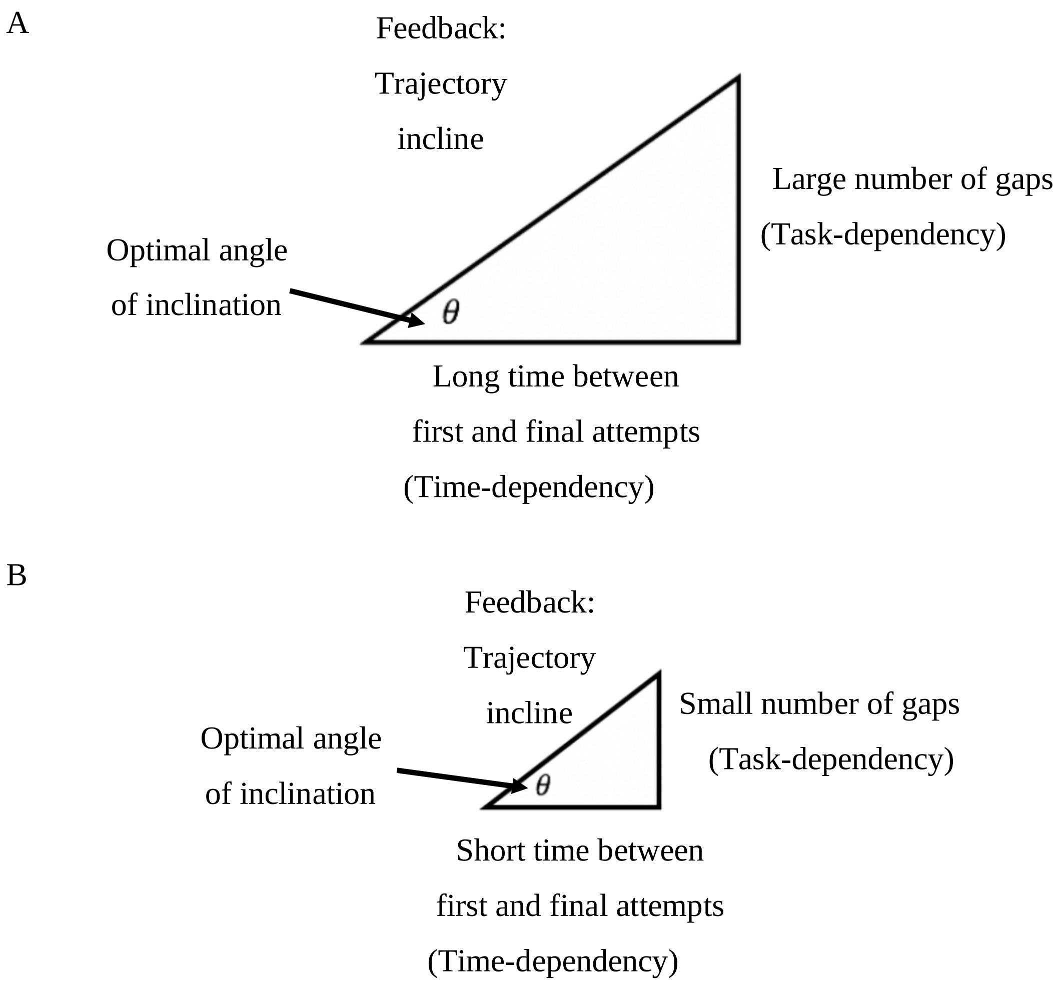


Figure 8.3.2 Triangulated model displaying ideal proportions and optimal angle of inclinations. (A) Model displaying a large number of learning gaps and a long time taken to act on feedback. (B) Model displaying a small number of learning gaps and a short time taken to act on feedback

Equally important, angles of inclination could be observed for students who did not improve performance between first and final attempts and may be determined sub-optimal (undesirable). For example, if a student was to identify many gaps in their learning and then was too ambitious acting on the feedback provided and did not allow much time between first and final attempts, this student would likely not display improvements (Figure 8.3.3A). Similarly, if a student was to identify minimal learning gaps in their first attempt and then wait too long between first and final attempts, skills may have been lost or misconceptions may have been developed in this time; thus, improvements may not have been made (Figure 8.3.3B).

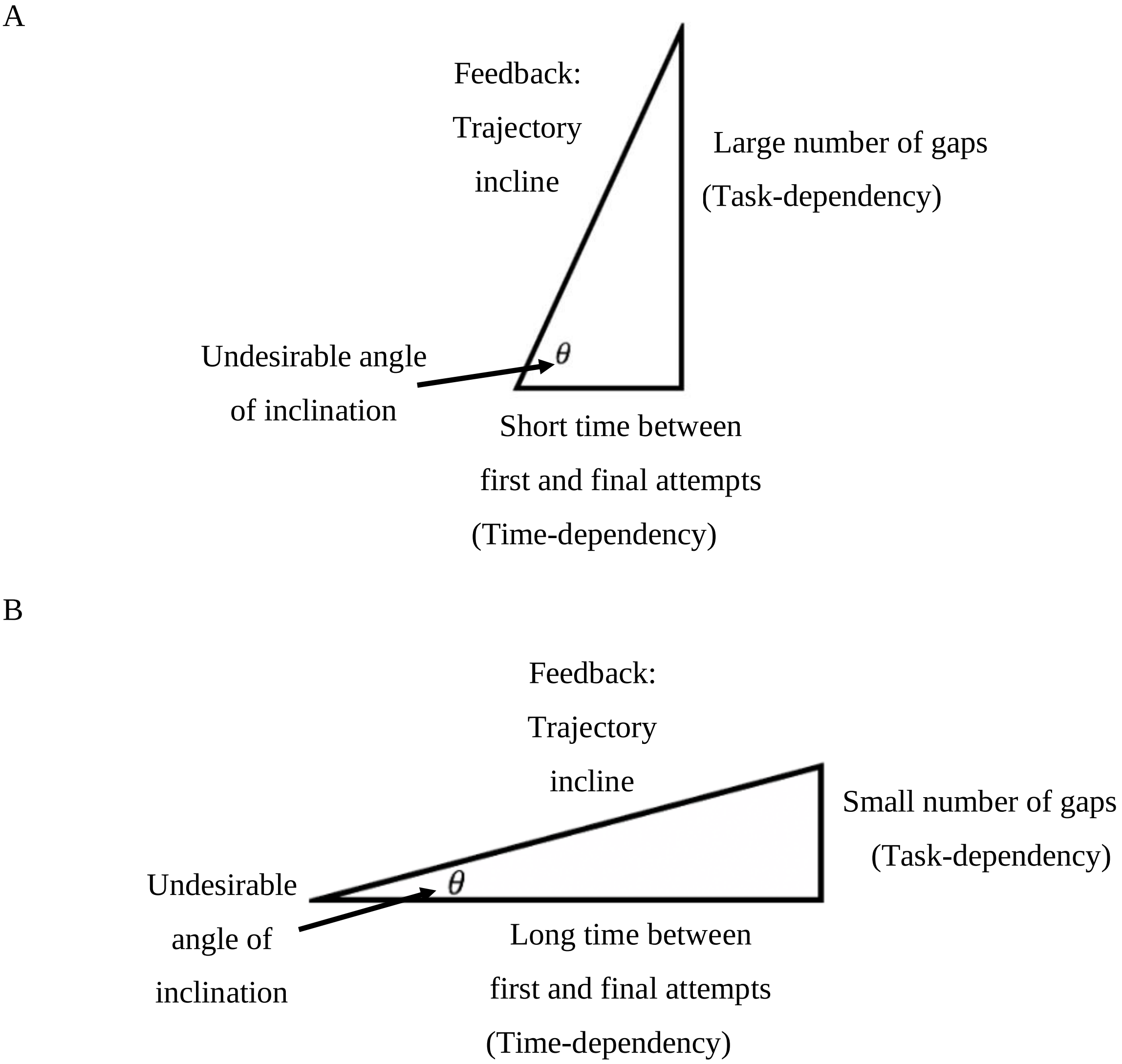


Figure 8.3.3 Triangulated model displaying undesirable proportions and sub-optimal angle of inclinations. (A) Model displaying a large number of learning gaps and not enough time taken to act on feedback. (B) Model displaying a small number of learning gaps and a long time taken to act on feedback

The overall objective of exploring this proposed model would be to view students' learning through the angle of inclination. This could be achieved using the results from this research and further existing data collected from the online system (i.e., the time between first and final attempts). This investigation is an area for future work, and if validated, could provide early warnings to institutions on how to support TES' numeracy development.

8.4 Future Direction

An area for future research is to explore how TES used the feedback provided to them through the online Diagnostic Test. Considering that most students displayed improvements and the fact that many students were displaying considerable improvements, it would be insightful to understand how TES acted on this feedback. For example, it would be beneficial for researchers to interview students who demonstrated improvements and compare/contrast with students who showed no improvement or decreased performance. Information on TES' perception of the Diagnostic Test, their experience using the test (from an A*f*L perspective), whether they acted on the feedback and how they acted on the feedback (including the sources used, e.g., private tutoring, textbooks, online resources, etc.), are all important factors to explore. Future research could also explore whether students found the worked solutions alone to be enough to improve their learning.

It is imperative that deficiencies and misconceptions identified through the Diagnostic Test be remediated to improve TES' numeracy capabilities. Therefore, future work could focus on studying approaches to improve TES' areas of weakness which we identified (Angles, Combinations, and Rates & Ratios). Given that the findings show that these areas were sufficiently improved upon through repeated use of the test, mechanics, and strategies that encourage TES to actively utilise the Diagnostic Test warrants exploration. Additionally, it is important to explore topics that we identified that TES declined in performance between first and final attempts. Lastly, it is necessary to understand the cause of this decline in performance, which could be due to misconceptions and/or other non-academic factors (e.g., psychological or emotional factors).

8.5 Conclusion

In order to address the main research question and the other questions posed, it was crucial to collect data that allowed for a thorough evaluation of TES' numeracy skills. Therefore, the development of the data collection instrument was a major component of the research. For this reason, it is important to acknowledge the importance of utilising a suitable methodology, especially the design and development of the testing instrument used to collect the data. In particular, it is important to highlight the necessity to develop, implement, and evaluate a Pilot Test that then informed the design and development of the main Diagnostic Test. Furthermore, the development of the items, and the iterative processes undertaken, ensured the quality design of the test that allowed thorough exploration of the research questions to address the gap in the literature.

In summary, findings from this study describe the extent that TES' numeracy skills can be evaluated and improved using an online Diagnostic Test. Findings showed that skills could be evaluated in test sections, mathematical strands, content areas, item types, context domains, and ACSF levels. Capabilities were seen to be best in the NC test section, Algebra and Percentages content areas, Multiple-Choice items, and Level 2 items. Capabilities were lowest overall in the SP strand, in the Angles, Combinations and Rates & Ratios content areas, Fill-in-the-Blank items, and Level 5 items. Furthermore, findings showed that that TES' overall ability improved on subsequent test attempts, and individual improvements were evident for most TES between first and final attempts. Improvements were seen in all test sections, mathematical strands, content areas, item types, context domains, and ACSF levels. However, TES displayed the most gains for NC items, Algebra, Probability, Angles, and Space, Shapes & Symmetry, Multiple-Choice items, and the most difficult (Level 5) items. The least gains were seen in the content areas of Capacity & Volume, Statistics, Distance & Perimeter, Financial Mathematics, and Interpreting Data, and also for Fill-in-the-Blank items, and the easiest (Level 2) items. Although improvements were seen across all areas in the test, the most significant finding was that TES could make considerable improvements at the levels they are most required to possess (Level 3 and 4).

Since all Australian TES are expected to teach numeracy regardless of the year that they teach or specialised teaching area, it is crucial that TES have adequate numeracy capabilities before they graduate and become teachers. The findings show that through the use of an online Diagnostic Test, it is possible to identify TES' numeracy capabilities across multiple dimensions (mathematical categories and content areas, question types, context domains, and ACSF levels). Importantly, it was evident that this was a viable way to help TES improve their numeracy skills. Knowledge gained from this research will enable initial teacher education providers that seek to better understand and help improve their TES' numeracy capability via a sustainable, relatively low-cost approach. Additionally, this approach provides information on individual TES' and the cohort's strengths and weaknesses, which will allow for further, more targeted teaching and learning strategies to be implemented at each institution.

Finally, it is clear that Australian institutions acknowledge the need for TES to improve their numeracy capabilities, and attempts have been made to implement successful initiatives to improve skills (Callingham et al., 2015; Forgasz & Hall, 2019; Sellings et al., 2018). Considering the limited number of studies that have examined TES' numeracy capabilities in Australia and internationally, this research shows for the very first time a method of assessing, tracking, and improving TES' numeracy skills that can be implemented in all initial teacher education programmes. In the long term, this will benefit schools by having increasingly more numeracy-competent teachers educating students.

Reference List

Abida, K., Azeem, M., & Bashir Gondal, M. (2011). Assessing students' maths proficiency using multiple-choice and short constructed response item formats. *The International Journal of Technology, Knowledge, and Society*, *7*(3), 135-149.

Afamasaga-Fuata'i, K., Meyer, P., Falo, N., & Sufia, P. (2007). Future teachers' developing numeracy and mathematical competence as assessed by two diagnostic tests. Australian Association for Research in Education 2006 International Education Research Conference,

Alcoholado, C., Diaz, A., Tagle, A., Nussbaum, M., & Infante, C. (2016). Comparing the use of the interpersonal computer, personal computer and pen-and-paper when solving arithmetic exercises. *British Journal of Educational Technology*, *47*(1), 91-105. https://doi.org/10.1111/bjet.12216

Allan, B. (2016). *Emerging strategies for supporting student learning: A practical guide for librarians and educators*

Andrich, D., & Marais, I. (2019). *A course in Rasch Measurement Model: Measuring in the educational, social and health sciences*. Spinger Test in Education. https://doi.org/10.1007/978-981-13-7496-8_1

Angrist, J., & Guryan, J. (2008). Does teacher testing raise teacher quality? Evidence from the state certification requirements. *Economics of Education Review*, *27*(5), 483-503. https://doi.ord/10.1016/j.econedurev.2007.03.002

Aparicio, M., Bacao, F., & Oliveira, T. (2016). An e-learning theoretical framework. *Educational Technology & Society*, *19*(1), 292-307.

Australian Bureau of Statistics. (n.d.). *Sample size calculator*. https://www.abs.gov.au/websitedbs/d3310114.nsf/home/sample+size+calculator

Australian Council for Educational Research. (2014). *Improve literacy and numeracy skills for Australia's future* http://www.acer/.edu.au/filesMR_Improve_literacy_and_numeracy_skills_for_Australias_future.pdf

Australian Council for Educational Research. (2017). How to intepret the statement of test results for the literacy and numeracy test for intitial teacher education students. https://teacheredtest.acer.edu.au/files/How_to_interpret_the_statement_of_results_2017.pdf

Australian Council for Educational Research. (2018b). *PISA*. Australian Council for Educational Research. Retrieved April 25 from https://www.acer.org/ozpisa

Australian Council for Educational Research. (2018c). *Literacy and numeracy test for initial teacher education students*. Retrieved April 22 from https://teacheredtest.acer.edu.au

Australian Council for Educational Research (ACER). (2017). *Literacy and numeracy test for initial teacher education students assessment framework*. Retrieved 5 June from https://teacheredtest.acer.edu.au/files/Literacy-and-Numeracy-Test-for-Initial-Teacher-Education-Students-Assessment-Framework.pdf

Australian Council for Educational Research (ACER). (2018a). *Trends in international mathematics and science study (TIMSS)*. Retrieved 20 May from https://www.acer.org/timss/australian-results-timss-2015-dec-2016

Australian Curriculum and Reporting Authority (ACARA). (n.d.). Australian curriculum. https://acara.edu.au/curriculum

Australian Curriculum Assessment and Reporting Authority. (2012). Reliability and validity of NAPLAN. Retrieved 20 August 2014, from

Australian Institute for Teaching and School Leadership. (2011). Australian professional standards for teachers. https://www.aitsl.edu.au/docs/default-source/national-policy-framework/australian-professional-standards-for-teachers.pdf

Baker, J. (2019). Alam bells: Australian students record worse result in global tests. *Sydney Morning Herald*. https://www.smh.com.au/education/alarm-bells-australian-students-record-worst-result-in-global-tests-20191203-p53gie.html

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications*, *10*(3798). https://doi.org/10.1038/s41467-019-11691-y

Ball, D. L. (1990). The mathematical understandings that prospective teachers bring to teacher education. *The Elementary School Journal*, *90*(4), 449-466.

Bangert-Drowns, R. L., Kulik, C. L. C., & Kulik, J. A. (1991a). Effects of frequent classroom testing. *Journal of Educational Research*, *85*, 88-89.

Bangert-Drowns, R. L., Kulik, C. L. C., & Kulik, J. A. (1991b). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*, 213-238.

Barnes, M., & Cross, R. (2018). Why we need to review how we test for teacher quality. *The Conversation*. https://theconversation.com/why-we-need-to-review-how-we-test-for-teacher-quality-95074

Barry, S. (2017). New teachers score 95 percent in skills test. https://www.school-news.com.au/news/new-teachers-score-95-percent-in-skills-test/

Becker, W., Greene, W., & Rosen, S. (1990). Research on high school economic education. *The Jounral of Economic Education*, *21*(3), 231-245.

Berry, R. (2008). *Assessment for learning*. Hong Kong University Press.

Berry, R., & Kennedy, K. J. (2008). *Assessment for learning*. Hong Kong Univeristy Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7-74.

Blanco, M., Estela, M. R., Ginovart, M., & Saà, J. (2009). Computer assisted assessment through moodle quizzes for calculus in an engineering undergraduate course. *CIEAEM61*. https://core.ac.uk/download/pdf/41758161.pdf

Boaler, J. (2014). Research suggests that timed tests cause math anxiety. *Teaching Children Mathematics*, *20*(8), 469-474. https://doi.org/10.5951/teacchilmath.20.8.0469

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Boud, D. (2000). Sustrainable assessment: rethinking assessment for the learning society. *Studies in Continuing Education*, *22*(2), 151-167. https://doi.org/10.1080/713695728

Brown, B., & Baker, S. (2007). *Philosophies of research into higher education*. Bloomsbury Publishing Plc.

Butler, A. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning. *Journal of Applied Research in Memory and Cognition*, *7*, 323-331.

Bynner, J., McIntosh, S., Vignole, A., Dearden, L., Reed, H., & Van Reenan, J. (2001). *Improving adult basic skills: Benefits to the individual and society* (251). http://hdl.voced.edu.au/10707/33312

Callingham, R., Beswick, K., & Ferme, E. (2015). An initial exploration of teachers' numeracy in the context of professional capital. *ZDM Mathematics Education*, *47*, 549-560. https://doi.org/https://doi-org.ipacez.nd.edu.au/10.1007/s11858-015-0666-7

Campus Morning Mail. (2018). *Where teaching graduates are best taught literacy and numeracy*. Retrieved April 25 from http://campusmorningmail.com.au/news/where-teaching-graduates-are-best-taught-literacy-and-numeracy/?utm_campaign=website&utm_source=sendgrid.com&utm_medium=email

Carmody, G., Godfrey, S., & Wood, L. (2006). Diagnostic tests in a first year mathematics subject. *UniServe Science: Proceedings of the assessment in science teaching and learning symposium*, 24-30.

Castro, M., & Tumibay, G. (2019). A literature review: efficacy of online learning courses for higher education institution using meta-analysis. *Education and Information Technologies*, *26*(2), 1367-1385.

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Education Psychology Review*, *24*, 205-249. https://doi.org/10.1007/s10648-011-9191-6

Cockcroft, W. (1982). *Mathematics counts*. HMSO.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7 ed.). Routledge.

Cohen, L., Manion, L., & Morrison, K. (2017). *Research methods in education* (8 ed.)

Commission on Teacher Credentialing. (2017). *California basic educational skills test (CBEST): Test specifications*. http://www.ctcexams.nesinc.com/content/docs/CBESTUpdatedTestSpecs.pdf

Conyers, S., & Scott, A. (2012). *Teachers want funding focused on schools not NAPLAN*. Retrieved August 22 from http://www.couriermail.com.au/questnews/teachers-want-funding-focused-on-schools-not-on-naplan-tests/story-fn8ygho7-1226369291428

Corbetta, P. (2003). *Social research: Theory, methods and techniques*. Sage Publications.

Crotty, M. (2020). *The foundation of social research: Meaning and persepctive in the research process*. Routledge.

Dabbagh, N. (2005). Pedagogical models for e-learning: A theroy-based design framework. *International Journal of Technology in Teaching and Learning*, *1*(1), 25-44.

Dann, R. (2014). Assessment as learning: blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy, and Practice*, *21*(2), 149-166.

Darling-Hammond, L. (2000). Teacher quality and student achievement: a review of state policy evidence. *Education Policy Analysis Archives*, *8*, 1-44.

DeCuir-Gunby, J. T. (2008). Mixed methods research in the social sciences. In J. Osborne (Ed.), *Best practices in quantitative methods*. SAGE Publications. https://doi.org/ https://dx-doi-org.ipacez.nd.edu.au/10.4135/9781412995627

Department for Education. (2011). *Training our next generation of outstanding teachers: Implementation plan*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/181154/DFE-00083-2011.pdf

Department for Education. (2015). *Professional skills test for prospective teachers: Numeracy test format and content*. http://sta.education.gov.uk/system/resources/W1siZiIsIjIwMTUvMDQvMjkvMTBfMjlfMzFfODE5X051bWVyYWN5X3Rlc3Rfc3BlY2lmaWNhdGlvbl9mb3JfY2FuZGlkYXRlc19BcHJpbF8yMDE1LnBkZiJdXQ/Numeracy%20test%20specification%20for%20candidates%20April%202015.pdf

Department for Education. (2018a). *Professional skills test*. Department for Education. Retrieved April 23 from http://sta.education.gov.uk

Department for Education. (2018b). *Numeracy skills test*. Department of Education. Retrieved April 23 from http://sta.education.gov.uk/professional-skills-tests/numeracy-skills-tests

Department for Education. (2018c). *Skills test statistics*. Department for Education. Retrieved April 23 from http://sta.education.gov.uk/professional-skills-tests/skills-tests-statistics

Department for Education. (2020a). *Changes to professional skills test*. http://sta.education.gov.uk/.

Department for Education. (2020b). *Teacher recruitment and retention strategy*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/786856/DFE_Teacher_Retention_Strategy_Report.pdf

Department of Employment, Skills, Small and Family Business,. (2015). Australian core skills framework. https://www.dese.gov.au/skills-information-training-providers/resources/australian-core-skills-framework

Education Quality and Accountability Office. (2020). *The Ontario mathematics proficiency test (MPT): Assessment blueprint*. https://s3.ca-central-1.amazonaws.com/authoring.mathproficiencytest.ca/user_uploads/6526/authoring/MPT_Assessment_Blueprint-EN/1605733279372/MPT_Assessment_Blueprint-EN.pdf

Education Testing Service. (2018a). *PRAXIS*. Retrieved May 10 from https://www.ets.org/praxis

Fask, A., Englander, F., & Wang, Z. (2015). On the integrity of online testing for introductory statistics course: A latent variable approach. *Practical Assessment, Research and Evaluation*, *20*(10). https://web-a-ebscohost-com.ipacez.nd.edu.au/ehost/pdfviewer/pdfviewer?vid=1&sid=957b1d04-446f-43a1-9718-3a9a63e536c2%40sdc-v-sessmgr03

Ferme, E. (2014). What can other areas teach us about numeracy? *Australian Mathematics Teacher*, *70*(4).

Finlayson, M. (2014). Addressing maths anxiety in the classroom. *Improving Schools*, *17*(1), 99-115. https://doi.org/10.1177/1365480214521457

Fitzmaurice, O., Walsh, R., & Burke, K. (2021). The 'mathematics problem' and preservice post primary mathematics teachers - analysing 17 years of diagnostic test data. *International Journal of Mathematical Education in Science and Technology*, *52*(2), 259-281. https://doi.org/10.1080/0020739X.2019.1682700

Forgasz, H., & Hall, J. (2019). Leaning about numeracy: The impact of a compulsory unit on pre-service teachers' understanding and beliefs. *Australian Journal of Teacher Education*, *44*(2). https://doi.org/10.14221/ajte.2018v44n2.2

Fujita, T., & Jones, K. (2006). Primary trainee teachers' understanding of basic geometrical figures in scotland. *PME*, *30*(129-136).

Galligan, L., & Hobohm, C. (2015). Investigating students' academic numeracy in 1st level university courses. *Mathematics Education Research Journal*, *27*, 129-145.

Gipps, C. (1994). Development in educational assessment: what makes a good test? *Assessment in Education: Principles, Policy, and Practice*, *1*(3), 283-292. https://doi.org/10.1080/0969594940010304

Gipps, C. (2011). *Beyond testing: Towards a theory of educational assessment* Taylor & Francis Group.

Girvan, C., & Savage, T. (2012). Ethical considerations for education research in a virtual world. *Interactive Learning Environments*, *20*(3), 239-251. https://doi.org/10.1080/10494820.2011.641678

Glidden, P. L. (2008). Prospective Elementary Teachers' Understanding of Order of Operations. *School Science and Mathematics*, *108*(4), 130-136. https://doi.org/10.1111/j.1949-8594.2008.tb17819.x

Golsteyn, B., Vermeulen, S., & de Wolf, I. (2016). Teacher literacy and numeracy skills: International evidence from PIAAC and ALL. *De Economist*(164), 365-389. https://doi.org/10.1007/s10645-016-9284-1

Graham, A. (2013). Black teacher education candidates' performance on PRAXIS I: What the results do not tell us. *Negro Educational Review*, *64*(1), 9-35.

Gudmundsdottir, S., & Shulman, L. (1987). Pedagogical content knowledge in social studies. *Scandinavian Journal of Educational Research*, *31*(2), 59-70. https://doi.org/10.1080/0031383870310201

Guler, M., & Celik, D. (2018). Uncovering the relation between CK and PCK: An investigation of preservice elementary mathematics teachers' algebra teaching knowledge. *REDIMAT - Journal of Research in Mathematics Education*, *7*(2), 162-194.

Gustafson, K., & Branch, R. (2002). What is instructional design? *Trends and Issues in Instructional Design and Technology*, 10-16.

Hacisalihaglu Karadeniz, M., Baran Kaya, T., & Bozkus, F. (2017). Explanations of prospective middle school mathematics teachers for potential misconceptions on the concept of symmetry. *International Electronic Journal of Elementary Education*, *10*(1), 71-82.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*(4), 999-1010.

Hall, J., & Zmood, S. (2019). Australia's literacy and numeracy test for initital teacher education students: Trend in numeracy for low- and high-achieving students. *Australian Journal of Teacher Education*, *44*(10). https://doi.org/10.14221/ajte.2019v44n10.1

Handal, B., Novak, A., Watson, K., Maher, M., MacNish, J., & Eddles-Hirsch, K. (2014). Numeracy education through mobile apps. *Middle Years of Schooling Asiciation 14*(1), 28-37.

Harlen, W. (2007). *Assessment of learning*. SAGE Publications.

Hartley, R., & Horne, J. (2006). Researching literacy and numeracy costs and benefits: What is possible. *Literacy and Numeracy Studies*, *15*(1), 5-22.

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement* Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112. https://doi.org/10.3102/003465430298487

Havnes, A., Smith, K., Dysthe, O., & Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. *Studies in Educational Evaluation*, *38*(1), 21-27.

Holm, J. (2018). MB4T (mathematics by and for teachers): Relating area and perimeter. *Ontario Mathematics Gazette*, *56*(3), 31-33.

Hoy, W. (2010). *Quantitative research in education: A primer*. Sage Publications Ltd.

Iramaneerat, C., Smith Jr, E., & Smith, R. (2008). An introduction to rasch measurement. In J. Osborne (Ed.), *Best practices in quantitative methods*. SAGE Publications. https://doi.org/10.4135/9781412995627.d6

Isik, C., & Kar, T. (2012). The analysis of the problems the pre-service teachers experience in posing problems about equations. *Australian Journal of Teacher Education*, *37*(9).

Izsák, A., Orrill, Chandra H., Cohen, Allan S., & Brown, Rachael E. (2010). Measuring middle grades teachers' understanding of rational numbers with the mixture rasch model. *The Elementary School Journal*, *110*(3), 279-300. https://doi.org/10.1086/648979

Karatoprak, R., Akar, G. K., & Börkan, B. (2015). Prospective elementary and secondary school mathematics teachers' statistical reasoning. *International Electronic Journal of Elementary Education*, *7*(2), 107-124.

Ladyshewsky, R. (2015). Post-graduate student performance in 'supervised in-class' vs 'unsupervised online' mulitple choice tests: implications for cheating and test security. *Assessment & Evaluation in Higher Education*, *40*(7), 883-897. https://doi.org/ doi:10.1080/02602938.2014.956683

Lefever, S., Dal, M., & Matthiasdottir, A. (2007). Online data collection in academic research: advantages and limitations. *British Journal of Educational Technology*, *38*(4), 574-582. https://doi.org/10.1111/j.1467-8535.2006.00638.x

Lindberg, S. M., Hyde, J. S., Peterson, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*(6), 1123-1135.

Linsell, C., & Anakin, M. (2012). Diagnostic assessment of pre-service teachers' mathematical content knowledge. *Mathematics Teacher Education and Development*, *14*(2), 4-27.

Livy, S., Muir, T., & Maher, N. (2012). How do they measure up? Primary pre-service teachers' mathematical knowledge of area and perimeter. *Mathematics Teacher Education and Development*, *14*(2), 91-112.

Lo, J.-J., & Luo, F. (2012). Prospective elementary teachers' knowledge of fraction division. *Journal of Mathematics Teacher Education*, *15*(6), 481-500. https://doi.org/10.1007/s10857-012-9221-4

Lowrie, T., Logan, T., & Scriven, B. (2012). Perspectives on geometry and measurement in the Australian curriculum: Mathematics. In *Engaging the Australian Curriculum: Mathematics - perspectives from the field* (pp. 71-88). MERGA.

Lui, O., & Wilson, M. (2009). Gender differences in large-sclae math assessments: PISA trends 2000 and 2003. *Applied Measurement in Education*, *22*, 164-184. https://doi.org/10.1080/08957340902754635

Mason, R., & Rennie, F. (2006). *Elearning: The key concepts*. Routledge.

Matthews, M., & Ding, M. (2011). Common mathematical errors of pre-service elementary school teachers in an undergraduate course. *Mathematics and Computer Education*, *45*(3), 186.

McGaw, B., Louden, W., & Wyatt-Smith, C. (2020). *NAPLAN review final report*. https://qed.qld.gov.au/programsinitiatives/education/Documents/naplan-review-final-report.pdf

Metz, A. M. (2008). The effect of access time on online quiz performance in large biology lecture course. *Biochemistry and Molecular Biology Education*, *36*(3), 196-202.

Ministry of Education Singapore. (2018). *Entrance proficiency tests*. Retrieved May 10 from https://www.moe.gov.sg/careers/teach/entrance-proficiency-test

Muijs, D. (2010). *Doing quantitative research in education using SPSS*. SAGE Publicaitons.

Mulhern, G., & Wylie, J. (2006). Mathematical prerequisites for learning statistics in psychology: assessing core skills of numeracy and mathematical reasoning among undergraduates. *Psychology Learning and Teaching*, *5*(2), 119-132.

New South Wales Council of Deans of Education. (2015). *NSW council of deans of education (NSWCDE) welcomes new literacy and numeracy tests for initial teacher education* https://www.acde.edu.au/?wpdmact=process&did=OTUuaG90bGluaw

Ngu, B. H. (2019). Solution representation of percentage change problems: the pre-service primary teachers' mathematical thinking and reasoning. *International Journal of Mathematical Education in Science and Technology*, *50*(2), 260-276. https://doi.org/10.1080/0020739X.2018.1494860

Nichols Hess, A., & Greer, K. (2016). Designing for engagement: Using the ADDIE model to integrate high-impact practices into an online information literacy course. *Communications in Information Literacy*, *10*(2).

Norton, S. (2019). The relationship between mathematical content knowledge and mathematical pedagogical content knowledge of prospective primary teachers. *Journal of Mathematics Teacher Education*, *22*, 489-514.

Nortvedt, G. A., & Siqveland, A. (2019). Are beginning calculus and engineering students adequately prepared for higher education? an assessment of students' basic mathematical knowledge. *International Journal of Mathematical Education in Science and Technology*, *50*(3), 325-343. https://doi.org/10.1080/0020739X.2018.1501826

NSW Education Standards Authority. (2017). *NSW supplementary documentation: elaborations in priority areas*. https://educationstandards.nsw.edu.au/wps/wcm/connect/15b74dc4-462b-4fad-9e10-74c8a2124c19/elaboration-in-priority-areas.pdf?MOD=AJPERES&CVID

NSW Education Standards Authority. (2018). *Literacy and numeracy tests*. Retrieved April 18 from http://educationstandards.nsw.edu.au/wps/portal/nesa/teacher-accreditation/how-accreditation-works/your-accreditation/future-teachers/literacy-numeracy-tests

NSW Education Standards Authority. (2020). *Studying teaching*. https://educationstandards.nsw.edu.au/wps/portal/nesa/teacher-accreditation/teaching-qualifications/studying-teaching

NSW Education Standards Authority (NESA). (2017). *English language proficiency of teachers for provisional or conditional accreditation policy*. Retrieved May 10 from https://educationstandards.nsw.edu.au/wps/wcm/connect/cf6101b7-b4e5-4560-b488-b2127ce9ac80/English+Language+Proficiency+Policy.pdf?MOD=AJPERES&CVID

NSW Government. (2021). *Assessment of learning*. https://education.nsw.gov.au/teaching-and-learning/professional-learning/teacher-quality-and-accreditation/strong-start-great-teachers/refining-practice/aspects-of-assessment/assessment-of-learning

Ontario College of Teachers. (2020). *Mathematics Proficiency Test*. https://www.oct.ca/becoming-a-teacher/requirements/mathematics-test

Organisation for Economic Co-operation and Development. (2013). *Survey of adult skills first results*. Retrieved 9 July from http://www.oecd.org/skills/piaac/Country%20note%20-%20Australia_final.pdf

Ozdemir, A. S., & Goktepe Yildiz, S. (2015). The analysis of elementary mathematics preservice teachers' spatial orientation skills with SOLO model. *Eurasian Journal of Educational Research*(61), 217-236.

Paparistodemou, E., Potari, D., & Pitta-Pantazi, D. (2014). Prospective teachers' attention on geometrical tasks. *Educational Studies in Mathematics*, *86*(1), 1-18. https://doi.org/10.1007/s10649-013-9518-y

Park, C.-G. (2010). Mathematics or numeracy? An ongoing debate. *Teaching Mathematics*, *35*(1), 14-19.

Pearson Education Inc. (2018b). *California educator credentialing examinations*. Retrieved May 10 from https://www.ctcexams.nesinc.com/PageView.aspx?f=GEN_Tests.html

Pearson Education Inc. (2018d). *National Evaluation Series*. Retrieved May 10 from https://www.nestest.com

Perry, B., Lowrie, T., & Logan, T. (2012). *Research in mathematics education*. Sense Publishers.

Perso, T. (2006). Teachers of mathematics of numeracy? *Australian Mathematics Teacher*, *62*(2), 36-40.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*(5), 407-413. https://doi.org/10.1111/j.1467-9280.2006.01720.x

Phillips, M. (2015). ICT is falling in schools - here's why. *The Conversation*. https://theconversation.com/ict-is-falling-in-schools-heres-why-50890

Ranellucci, J., Rosenburg, J., & Poitras, E. (2020). Exploring pre-service teachers' use of technology: The technology acceptance model and expectancy-value theory. *Journal of Computer Assisted Learning*, 810-824.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin*, *135*(6), 943-973. https://doi.org/10.1037/a0017327

Riley, D. (2007). The paradox of positivism. *Social Science History*, *31*(1), 115-126. (Cambridge University Press)

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119-144.

Şahin, O., & Başgül, M. (2020). Pre-service primary school teachers' pedagogical content knowledge on quadrilaterals. *Acta Didactica Napocensia*, *13*(2), 284-305.

Salloum, S. A., Mohammad Alhamad, A. Q., Al-Emran, M., Adbdel Monem, A., & Shaalan, K. (2019). Exploring students' acceptance of e-learning through the development of a comprehensive technology acceptance model. *IEEE Access*, *7*, 128445-128462.

School News. (2020). Gov confirms 90% of test students met literacy and numeracy standards benchmark in 2019. https://www.school-news.com.au/news/gov-confirms-90-of-teaching-students-met-literacy-and-numeracy-standard-benchmarks-in-2019/

Schwartz, L., Woloshin, S., Black, W., & Welch, G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, *127*(11).

Sekulich, K. M. (2020). Learning through formative feedback: A review of the literature. *Delta Kappa Gamma Bulletin*, *86*(3), 51-59.

Sellings, P., Felstead, K., & Goriss-Hunter, A. (2018). Developing pre-service teachers: The impact of an embedded framework in literacy and numeracy. *Australian Journal of Teacher Education*, *43*(4). https://doi.org/10.14221/ajte.2018v43n4.1

Shirvani, H. (2015). Pre-service elementary teachers' mathematics content knowledge: A predictor of sixth graders' mathematics performance. *International Journal of Instruction*, *8*(1), 133-142.

Shomos, A. (2010). Links between literacy and numeracy skills and labour market outcomes. *Productivity Commission Staff Working Paper*. https://ssrn.com/abstract=1802872

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1-22.

Siegfried, C., & Wuttke, E. (2019). Are multiple-choice items unfair? and if so, for whom? *Citizenship, Social and Economic Education*, *18*(3), 198-217. https://doi.org/10.1177/2047173419892525

Snekalatha, S., Mohamed Marzuk, S., Meshram, S., Uma Maheswari, K., Sugapriya, G., & Sivasharan, K. (2021). Medical students' perception of the reliability, usefulness and feasibility of unproctored online formative assessment tests. *Advances in Physiology Education*, *45*(1). https://doi.org/10.1152/advan.00178.2020

Somekh, B. (2006). *Action research: a methodology for change and development* (P. Sikes, Ed.). McGraw-Hill Education.

Son, J.-W., & Lee, J.-E. (2016). Pre-service teachers' understanding of fraction multiplication, representational knowledge, and computational skills. *Mathematics Teacher Education and Development*, *18*(2), 5-28.

Song, Z., Cheung, M., & Prud'Homme, S. (2017). Theoretical frameworks and research methods in the study of MOOC/e-learning behaviours: A theoretical and emperical review. *New Ecology for Education - Communication X Learning*, 47-65. https://doi.org/10.1007/978-981-10-4346-8_5

Stables, A., Martin, S., & Arnhold, G. (2004). Student teachers' concepts of literacy and numeracy. *Research Papers in Education*, *19*(3), 345-364. https://doi.org/10.1080/0267152042000248007

Starr, K. (2014). The influences and implications of PISA: an Australian perspective. *AASA Journal of Scholarship & Practice*, *10*(4), 19.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, *62*, 339-355. http://dx.doi.org/10.1177/0022487111404241.

Tan, K. (2013). A framework for assessment for learning: Implications for feedback practices within and beyond the gap. *ISRN Education*, *2013*, Article 640608. https://www.hindawi.com/journals/isrn/2013/640609/

Taras, M. (2005). Assessment - summative and formative - some theoretical reflections. *British Journal of Educational Studies*, *53*(4), 466-478.

Taras, M. (2008). Summative and formative assessment perceptions and reality. *Active learning in higher education*, *9*(2), 172-192.

Taras, M. (2010). Assessment for learning: assessing the theory and evidence. *Procedia Social and Behavioral Sciences*, *2*(2010), 3015-3022.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics test with mixed item formats. *Applied Measurement in Education*, *25*(3), 246-280.

Tchoshanov, M., Cruz, M. D., Huereca, K., Shakirova, K., Shakirova, L., & Ibragimova, E. N. (2017). Examination of lower secondary mathematics teachers' content knowledge and its connection to students' performance. *International Journal of Science and Mathematics Education*, *15*(4), 683-702. https://doi.org/10.1007/s10763-015-9703-9

Teo, T., Lee, C. B., & Chai, C. S. (2008). Understanding pre-service teachers' computer attitudes: Applying and extending the technology acceptance model. *Journal of Computer Assisted Learning*, *24*(2), 128-143.

The Government of the Hong Kong Special Administrative Region. (2018). *Language proficiency assessments for teachers*. Retrieved May 10 from https://www.edb.gov.hk/en/teacher/qualification-training-development/qualification/language-proficiency-requirement/lpat.html

Thomson, S., De Bortoli, L., Underwood, C., & Schmid, M. (2019). PISA 2018: Reporting Australia's results. Volume I student performance. https://research.acer.edu.au/ozpisa/35/

Tobias, J. M. (2013). Prospective elementary teachers' development of fraction language for defining the whole. *Journal of Mathematics Teacher Education*, *16*(2), 85-103. https://doi.org/10.1007/s10857-012-9212-5

Trust, T., & Pektas, E. (2018). Using the ADDIE model and universal design for learning principles to develop an open online course for teacher professional development. *Journal of Digital Learning in Teacher Education*, *34*(4), 219-233.

Vappula, H., & Clausen-May, T. (2006). Context in maths test questions: Does it make a difference. *Research in Mathematics Education*, *8*(1), 99-115. https://doi.org/10.1080/14794800008520161

Vinner, S. (1991). The role of definitions in the teaching and and learning of mathematics. *Advanced Mathematical Thinking*, *11*.

Walstad, W., & Robson, D. (1997). Differential item functioning and male-female differences on multiple choice tests in economics. *The Journal of Economic Education*, *28*(2), 155-171.

Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. Jossey-Bass.

Wilkie, K. (2016). Learning to teach upper primary school algebra: changes to teachers' mathematical knowledge for teaching functional thinking. *Mathematics Education Research Journal*(28), 245-275.

Willis, S. (1998). Which numeracy? *Unicorn*, *24*(2), 32-42.

Wilson, R. (2013). Makes maths mandatory and we'll improve our international education rankings. *The Conversation*. theconversation.com/make-maths-mandatory-and-well-improve-our-international-education-rankings

World Economic Forum. (2016). *Which countries have the best literacy and numeracy rates?* Retrieved November 19 from https://www.weforum.org/agenda/2016/02/which-countries-have-the-best-literacy-and-numeracy-rates/

Wright, B. D. (2003). Rack and stack: Time 1 vs. time 2. *Rasch measurement transactions*, *17*(1), 905-906. http://www.rasch.org/rmt/rmt171a.htm

Yigit, M. (2014). An examination of pre-service secondary mathematics teachers' conceptions of angles. *The Mathematics Enthusiast*, *11*(3), 707-736, Article 13. https://scholarworks.umt.edu/tme/vol11/iss3/13

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education*, *45*(4), 477-501. https://doi.org/10.1023/A:1023967026413

Zembat, I. O. (2010). Prospective elementary teachers' conceptions of volume. *Procedia - Social and Behavioral Sciences*, *2*(2), 2111-2115. https://doi.org/10.1016/j.sbspro.2010.03.290

Zevenbergen, R. (2004). Technologizing numeracy: Intergenerational differences in working mathematically in new times. *Educational Studies in Mathematics*, *56*(971), 17.

Appendix

Appendix A. Item Anchor Measures

Item Number	Anchored Measure	Item Number	Anchored Measure	Item Number	Anchored Measure	Item Number	Anchored Measure	Item Number	Anchored Measure
1	-0.9	56	1.76	111	-1.95	166	-1.64	221	0.28
2	2.06	57	1.99	112	-0.8	167	-0.05	222	-0.85
3	1.07	58	0.91	113	0.64	168	1.32	223	-0.77
4	-1.58	59	2.53	114	-0.37	169	-0.72	224	-1.28
5	-0.5	60	0.4	115	1.09	170	1.22	225	0.58
6	-2.87	61	-1.12	116	-0.61	171	-0.42	226	-0.68
7	-3.04	62	1.54	117	0.48	172	1.42	227	-0.88
8	-1.51	63	0.21	118	-1.13	173	-0.19	228	0.23
9	0.56	64	1.14	119	1.13	174	-1.31	229	0.86
10	1.47	65	-0.48	120	0.13	175	-1.58	230	-1.34
11	0	66	-1.19	121	-1.24	176	-0.16	231	-0.64
12	-0.8	67	1.19	122	1.53	177	0.94	232	-1.11
13	-0.05	68	-3.95	123	1.06	178	0.95	233	-1.11
14	-0.03	69	-0.45	124	0.73	179	-0.46	234	-0.86
15	0.92	70	1.2	125	-0.29	180	-1.41	235	
16	-0.92	71	0.09	126	2.26	181	-0.12	236	2.08
17	-0.07	72	1.28	127	0.23	182	-1.04	237	-0.39
18	-0.03	73	-1.87	128	-0.46	183	-0.48	238	0.17
19	2.07	74	-1.85	129	0.48	184	-0.03	239	-1.04
20	1.78	75	-0.77	130	2.35	185	1.27	240	-0.66
21	1.29	76	-1.11	131	-0.34	186	-0.28	241	-1.53
22	-1.09	77	-0.89	132	-0.08	187	3.6	242	-1.32
23	-1.19	78	0.73	133	-1.87	188	1.41	243	-0.64
24	1.02	79	-1.16	134	0.51	189	2.78	244	-0.02
25	-0.28	80	0.59	135	0.87	190	-3.11	245	-1.71
26	0.16	81	-2.22	136	0.96	191	2.42	246	1.33
27	0.57	82	-0.08	137	-0.72	192	-0.26	247	-1.87
28	1.15	83	0.75	138	0.47	193	-0.27	248	0.69
29	0.12	84	-0.84	139	-0.91	194	-0.93	249	0.43
30	-0.19	85	0.92	140	1	195	-1.09	250	-0.96
31	-0.42	86	-1.36	141	-2.52	196	1.02	251	1.55
32	-1.05	87	1.82	142	0.39	197	2.01	252	-1.27
33	-0.07	88	-2.95	143	-1.1	198	-0.62	253	-0.29

Item Number	Anchored Measure	Item Number	Anchored Measure	Item Number	Anchored Measure	Item Number	Anchored Measure	Item Number	Anchored Measure
34	-0.33	89	-0.91	144	-0.9	199	1.62	254	-1.05
35	0.39	90	-0.59	145	-0.56	200	-1.15	255	-0.21
36	1.19	91	0.56	146	-0.93	201	0	256	-1.02
37	0.89	92	-0.52	147	-1.31	202	0.94	257	0.58
38	0.01	93	-0.66	148	-0.57	203	-2.16	258	0.34
39	0.32	94	1	149	0.1	204	-0.79	259	2.32
40	1.21	95	-0.78	150	-0.48	205	-0.47	260	-2.57
41	0.06	96	-1.49	151	0.39	206	0.87	261	-0.6
42	2.21	97	-1.41	152	0.54	207	0.09	262	1.74
43	0.8	98	DELETED	153	-3.67	208	1.65	263	-0.49
44	-1.28	99	0.31	154	-1.92	209	-0.66	264	0.78
45	-0.09	100	2.02	155	0.99	210	1.06	265	-5.31
46	2.15	101	1.05	156	0.3	211	-2.08	266	0.83
47	0.68	102	0.61	157	-0.94	212	2.29	267	0.33
48	1.14	103	-0.62	158	-2.08	213	1.98	268	-0.3
49	0.45	104	-0.21	159	-0.6	214	-0.28	269	-0.11
50	-0.08	105	-0.74	160	0.83	215	0.05	270	0.85
51	-2.99	106	-1.46	161	2.18	216		271	0.98
52	1.13	107	0.21	162	0.71	217		272	2.49
53	1.91	108	-0.77	163	1.73	218	1.92		
54	-0.26	109	-1.71	164	-0.01	219	-0.44		
55	1.27	110	-0.89	165	0.78	220	0.37		