Advancing the measurement of cognitive ability: Developing a Cattell-Horn-Carroll computer adaptive screening test

Jake Kraska

BPsychBus (Monash), BA (Murdoch), BTech (Swinburne), PGDipPsych (Monash), MPsych (Monash)

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

October 2020

Faculty of Education

Monash University

Copyright Notice and Declaration

©The author (2020). Except as provided in the Copyright Act 1968, this thesis may not be reproduced in any form without the written permission of the author.

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

This thesis is submitted to Monash University in partial fulfillment of the Doctorate of Philosophy (PhD). The work presented in this thesis is original except as acknowledged in the text. This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Jake Kraska

October 2020

“Failure is an option here. If things are not failing, you are not innovative enough.” - Elon Musk

“Science has not yet mastered prophecy. We predict too much for the next year and yet far too little for the next 10.” – Neil Armstrong

Acknowledgements

Throughout the writing of this thesis, I have received a great deal of support and assistance.

Firstly, several organisations require acknowledgement. This research was supported by the Nectar Research Cloud; a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy. The Krongold Clinic, Monash University, through the Krongold Outreach Program, supported the administration, scoring and interpretation of cognitive ability tests for school aged children by provisional psychologists under the supervision of the myself as a registered Educational and Developmental Psychologist. I also acknowledge the support of the Australian Scholarships Group via the ASG Scholarship for Excellence, and the support of the Australian Government via an Australian Government Research Training Program Scholarship.

The data collection in this PhD relied heavily on the *Concerto Platform* (https://github.com/campsych/concerto-platform) developed by The University of Cambridge Psychometrics Centre. The data analysis relied heavily on the *mirt* package for R (https://github.com/philchalmers/mirt) and *mirtCAT* package for R (https://github.com/philchalmers/mirtCAT) developed by Dr Phil Chalmers. That these tools are open source, freely available and receive ongoing development is a substantial boon to the fields of Item Response Theory (IRT), Computer Adaptive Testing and psychometrics. Without these tools this PhD would not have been possible within the timeframe of a typical candidature. Everyone that researches in this field should support these kinds of developers and initiatives.

I would like to thank Dr Shane Costello (Primary Supervisor) for his supervision of this project. Shane has always been a voice of reason and logic; where others have doubted that something can be done, he has supported me. He has always given me his time freely, both within this project and without. He has also allowed me great freedom and independence in developing ideas and executing them, never forcing me to utilise a method or making a conclusion that I was not satisfied with. Shane demonstrates a fundamental understanding of the intangible nature of psychological constructs and thus, the need for some flexibility in the execution of statistical methods; this PhD benefited substantially from these understandings. Outside of research Shane is a true scientist-practitioner, modelling the importance of understanding the practical implementation and limitations of academic research. Shane has helped develop my professional identity and research interests, and I am glad to call him a colleague and friend. This thesis truly would not be possible without the expert guidance of Shane and I cannot thank him enough. I look forward to our future collaborations.

Thank you also to Dr John Roodenburg (Secondary Supervisor), Associate Professor Wendy McKenzie (Secondary Supervisor) for their flexibility and feedback during key milestones of this project. The insightful feedback of all three supervisors has pushed this thesis and the thinking behind it to a higher level.

The services of Sara Nyhuis, a professional editor, were used for this thesis. Sara's specialisation is not within the field of psychology and editing was focused on matters of language, illustration, completeness, and consistency. It is because of her that readers of this thesis do not have to put up with the word 'whilst' or the overutilization of quotation marks. Thank you, Sara.

Thank you to Elizabeth Kennedy, Adalyn Heng and John Maguire, who contributed significantly to the beginnings of this project through the completion of their Masters projects focusing on the development of items that measure Cattell-Horn-Carroll (CHC) abilities from a theoretical perspective; I enjoyed supervising your projects alongside Shane and I hope that learning about R and IRT fuels your future research or psychological practice.

Thank you to Laura Dye, Dr Kate Jacobs and Karie Stewart for their expertise in teaching, psychological assessment and special education that supported the evaluation of items within this PhD prior to the Item Calibration Study.

Thank you to my PhD peers: Grace Mackie, Simone Gindidis, Andrea Sadusky. Your dedication to your PhDs has awed me and I know that each of your respective fields will not have heard the last from you. Grace, you provided integral consultation on literature reviews and assisted me in deciding how to approach a review of the cognitive ability CAT literature given its sparseness. Your positive attitude, helpful nature, and encouragement were momentously appreciated at a time of great need. Thanks for listening to my PhD lamentations. A project of the magnitude conducted in this thesis would be impossible without the support of others and these contributions truly prove that collaboration enhances the scientific process.

To my family, you have been integral to my academic pursuits. My brother implanted a healthy level of academic competitiveness and passion for scientific rigour, my mother instilled in me a sense of curiosity and inquisitiveness about the world, and my father taught me to be pragmatic and logical. Together these characteristics have prepared me for a competitive academic world. Although I know my career has often limited my time with family, I hope that you know I am doing what I love.

To my wife, you have supported me through 14-years of academic study, from one degree to another. I think it is time I start putting these fancy pieces of paper to use. Your patience and support have been unwavering. It was once a wise young woman that told a (too) serious teenager, “do not take life too seriously, you will never get out of it alive.” You truly are my best friend, and all my achievements would not have been possible without you.

A special thank you must also be expressed to Jack O’Neill, Finn, and John Wick.

Abstract

The current thesis explored the application of a Computer Adaptive Test (CAT) in the measurement of intelligence, the efficiency and validity of a CAT when measuring intelligence across a range of ages and considered correlations of this newly developed CAT with the Wechsler Intelligence Scale for Children – Fifth Edition (WISC-V).

Any measurement tool in psychology must be developed based on a good theoretical framework and sound measurement principles. The background of this thesis (Chapter 2) demonstrated that the Cattell-Horn-Carroll (CHC) theory is the most contemporary and suitable theory of intelligence to form the theoretical basis of a cognitive ability CAT. It is also important that researchers implement appropriate statistical models when developing and implementing a CAT, and thus the background of this thesis introduced many basic concepts related to different item response theory (IRT) models and characteristics of CATs that some readers may be unfamiliar with. Some authors have argued that the measurement of cognitive ability has lacked new innovations despite significant improvements in available technologies in the last two decades. CATs pose an opportunity to improve measurement of cognitive ability through the integration of CHC theory, IRT measurement principles and variation of CAT characteristics.

A review of the literature in this thesis demonstrated that despite CHC, IRT and CAT all being well known concepts within the literature, there has been only limited integration of the three together. Nearly all studies reviewed failed to describe the CAT used with enough detail for their studies to be replicated by researchers or implemented by practicing psychologists. Additionally, none of the CATs investigated demonstrated utility with an Australian sample or truly examined more than one CHC factor at a time.

The four studies in this thesis investigated the use of four sets of items developed from the perspective of CHC theory, designed to measure Lexical Knowledge, Induction, Visualisation and Working Memory. Both the data and statistical analyses for all four studies included in this thesis are accessible at github.com/jakekraska/phd.

The item sets were trialled and evaluated in an Item Tryout Study (ITOS; Chapter 3) and an Item Calibration Study (ICS; Chapter 4). The initial ITOS demonstrated that the Lexical Knowledge items tended to be too easy, Induction items were not always predictable in their ordering, there were possible unidimensionality issues with the Visualisation items, and that there were design issues with the Working Memory item stimuli. The ICS took advantage of further item development and addressed problems with item sets as identified by the ITOS and the analysis resulted in retaining 47 Lexical Knowledge items, 23 Induction items, 30 Visualisation items, 25 Working Memory items. Item parameters were exported for subsequent CAT simulation.

Each item set was included in a CAT simulation (Chapter 5) that made use of 5,000 simulated participant. Simulations were conducted with each item set using item parameters from the ICS and item parameters recalculated using only the school aged sample. Simulations were conducted for varying levels of reliability required utilising a minimum standard error of measurement (SEM) stop rule. A final simulation was conducted for each item set to evaluate where the item sets measured best. Overall, 52 simulations were conducted. It was found that that item administration could be reduced by approximately 50% when aiming for a reliability of .70, and the item sets best measured at very low ability levels through to average abilities. If such a test were implemented in practice, testing time would be approximately 18 minutes depending on the ability of the examinee. Classification of those with deficits in cognitive abilities could be achieved, however differentiation between those with average ability and above average ability may not be possible with the current item sets.

Analysis of the convergent validity of the items was investigated (Chapter 6). Weak to moderate correlations were found between each CHC-CAT subtest and the respective WISC-V subtest, and a moderate correlation was found between a statistically derived *g* factor and the WISC-V Full Scale IQ. Based on the results there is mounting evidence of the psychometric validity of the Lexical Knowledge and Induction item sets from this PhD. Further analysis is recommended to compare the Working Memory item sets into other cognitively complex Working Memory tasks, as well as further development of Visualisation items that provide a better fit to the Rasch model.

A discussion (Chapter 7) of the implications and limitations of this research is also presented. Future research opportunities are identified surrounding multidimensional IRT, use of other IRT models, improvements, and standardisation in item stimuli via funded multimedia design, further item development, and implementation of items that measure other CHC abilities. Overall, it is concluded that this research demonstrates the viability of CAT implementation into the measurement of CHC abilities and hopes to serve as a platform for future innovations.

Other Works During Candidature

Published and Inspired Research

Computational psychology and the intersection between data science and psychological research have inspired several related research projects, including:

Gunadi, M., Kraska, J., & Costello, S. (2017). Using machine learning to predict the relationship between social media use and empathy. Paper presented at the 16th annual Australian Conference on Personality and Individual Differences, Sydney, Australia.

Kraska, J., Bell, K., Robinson, N., Macdonald, P., Chandrasekaran, S., Costello, S. (2020). Computer Adaptive Testing with the DASS-21. Manuscript currently being prepared for submission.

Marinucci, A., Kraska, J., & Costello, S. (2018). Recreating the Relationship between Subjective Wellbeing and Personality Using Machine Learning: An Investigation into Facebook Online Behaviours. Big Data and Cognitive Computing, 2(3), 29.

Tan, A. T., Kraska, J., Bell, K. Costello, S. (2020). CFA of the BASC-3 with an Australian Sample. Manuscript currently under review at the Educational and Developmental Psychologist

Research Supervision During Candidature

Bird, J. (2017). Delay Discounting, Personality, and Machine Learning. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Chandrasekaran, S. (2018). Improving depression measurement with DASS21 Depression scale using Computer Adaptive Testing. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Csepregi, K. (2019). Improving the measurement of Machiavellianism using Computer Adaptive Testing. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Collins, J. (2020). An item response theory analysis of a psycholexical measure of psychological distress. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Duan, M. (2018). Measuring the Narcissism Personality Trait Using Computer Adaptive Testing. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Fleming, S. (2018). Cognitive Assessment and Fluid Reasoning: Moving to the Future with Item Response Theory and Computerized Adaptive Testing. (Master of Psychology), Monash University, Australia.

Gunadi, M. (2017). Using Machine Learning to Predict the Relationship Between Social Media Use and Empathy. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Gunathilake, *T*. (2018). Using Computer Adaptive Testing to Improve the Measurement of Psychopathy. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Heng, A. (2018). Towards a Computerised Adaptive Test of Visual Processing: A Pilot Item Development Study using Rasch Analysis. (Master of Psychology), Monash University, Australia.

Kennedy, E. (2018). The Development of Computer-Administered Items to Improve the Measurement of Working Memory. (Master of Psychology), Monash University, Australia.

Kocic, *D*. (2017). Machine Learning in Social Science: Predicting the Relationship between Life Satisfaction and Empathy. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Lothian, I. (2020). The moderating effect of executive functioning skills on the relationship between intelligence and reading achievement. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Macdonald, P. (2018). Using computerised adaptive testing to improve psychological measurement effectiveness in assessing anxiety. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Maguire, J. (2018). Past and Present Measures of Crystallised Intelligence Inform the Development of an Adaptive Cognitive Ability Scale. (Master of Psychology), Monash University, Australia.

Marrinucci, A. (2017). Recreating the relationship between Subjective Wellbeing and Personality using Machine Learning: An investigation into Facebook online opinions. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Mutchal, M. (2020). Psychologists’ attitudes towards their training, Area of Practice Endorsements, and the Better Access Scheme. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Puhovac, V. (2020). Psychologists perspectives of video gaming as a factor in the development of anxiety disorders. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Robinson, N. (2018). Investigating the suitability of Computer Adaptive Testing (CAT) for the measurement of stress. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Satarawala, *F*. (2020). The moderating effect of executive function skills on the relationship between intelligence and writing achievement. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Tzimas, *G*. (2017). Investigating the Relationship Between Depression and Personality Using Machine Learning. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Yang, *F*. (2020). Exploratory factor analysis of a psycholexical measure of distress. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia

Yu, J. (2020). Australian psychologists' attitudes towards Internet Gaming Disorder. (Postgraduate Diploma of Psychology Minor Thesis), Monash University, Australia.

Online Platforms

This thesis has inspired an interest in utilising open-source tools to enhance the presentation of both psychological data and research data, including:

Psychology Visualisation Tool - This tool is designed to help psychologists visualise psychological test data from cognitive ability tools across a singular assessment time or multiple assessment points, allowing several customisation options. https://jakekraska.shinyapps.io/psychvisualisation/

Australian Psychologists' Training - This dashboard has been set up to accompany an upcoming publication focused on the perceptions of Australian Psychologists' about their training, and the Australian system of "Area of Practice Endorsements" and "Better Access to Mental Health" scheme. https://jakekraska.shinyapps.io/psychtraining/

Table of Contents

List of Tables

List of Figures

List of Code Snippets

List of Equations

List of Abbreviations

CAT	Computer Adaptive Test
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CHC	Cattell-Horn-Carroll
CRC	Categorical Response Curve
CTT	Classical Test Theory
DIF	Differential Item Functioning
DWLS	Diagonally Weighted Least Squares
EAP	Expected a Posteriori
EMA	Ecological Momentary Assessment
FIML	Full Information Maximum Likelihood
g	General intelligence/cognitive ability
G*c*	Comprehension-Knowledge
G*c:*VL	Lexical Knowledge
G*f*	Fluid Reasoning
G*f:*I	Induction
G*v*	Visual Spatial Processing
G*v*	Visualisation
G*wm*	Short Term Working Memory
G*wm:*Wc	Working Memory
G*wm:*Wa	Auditory Working Memory
G*wm:*Wv	Visual Working Memory
ICC	Item Characteristic Curve

ICS	Item Calibration Study
ICS-A	Item Calibration Study Adults
ICS-U	Item Calibration Study School Aged
IRT	Item Response Theory
ITOS	Item Tryout Study
MAP	Maximum a Posteriori
MI	Maximum Information
MI	Multivariate Imputation
MICE	Multivariate Imputation by Chained Equations
MIRT	Multidimensional Item Response Theory
ML	Maximum Likelihood
RMSEA	Root Mean Square Error of Approximation
SB	Stanford-Binet
SD	Standard Deviation
SE	Standard Error
SEM	Structural Equation Modelling
SEM	Standard Error of Measurement
SRMR	Standardised Root Mean Square Residual
TIC	Test Information Curve
WISC	Wechsler
WISC-V	Wechsler Intelligence Scale for Children – Fifth Edition
WJ	Woodcock Johnson
WJ IV	Woodcock Johnson Fourth Edition

Chapter 1: Introduction and Thesis Outline

With the evolution of modern technology has come the ability to automate processes and efficiently perform complex algorithmic calculations. This thesis focuses on the development of a Computer Adaptive Test (CAT) that measures cognitive ability as conceptualised by the Cattell-Horn-Carroll (CHC) theory. CATs have the advantage of offering unsupervised administration, increased efficiency, improved precision of measurement, automated scoring, and a chance to further support construct validity. Therefore, CATs pose an opportunity within the field of psychology to reduce the burden of repetitive, structured and time-consuming psychological assessment processes.

While adaptive testing has existed for over a century, it has not been until recent technological developments that adoption of these techniques has started to become widespread; this is largely a result of computers enabling automated scoring and algorithm calculation. CATs have been shown to be as reliable as traditional tests in the measurement of various psychological constructs and perhaps more utilitarian in a testing scenario (Weiss, 2011). The advantages of CATs over traditional paper-and-pencil tests, when mixed with the power of web technologies, poses an interesting way forward for the improvement of cognitive ability measurement utilising contemporary perspectives of intelligence, such as CHC theory. Given the increasing demands on, and for, Educational Psychologists in the Australian community (Department of Employment, 2019; Lyonette et al., 2019) there are opportunities to evaluate new methods of measuring cognitive ability to increase efficiencies in the industry. The current thesis aims to contribute to the literature surrounding measurement of cognitive ability utilising CATs.

1.1 Progressing Psychological Assessment

CATs have been discussed in the research literature since at least the 1970s. Much early work relied on identifying algorithms that could efficiently calculate an individual's ability while determining the next appropriate item (e.g. Lord, 1980). Given the technology of the time, user interfaces and item presentation were basic, and calculation of the various mathematical principles of CATs was resource intensive.

With increases in personal computing in the 1990s, utilisation of the Internet and more customised CATs were being considered. Barak (1999) proposed several advantages of the Internet as a medium for psychological testing, including heightened accessibility, ease of scoring, decreased administration errors, ease of ability in updating test materials, removal of financial, geographic or time limits on administration, and the possible utilisation of dynamic graphical interfaces. Despite these advantages, many tests that were ultimately utilised on the Internet were simply electronic copies of paper-and-pencil tests.

Notwithstanding the proposed advantages of assessment via CATs or the Internet, the measurement of cognitive ability has barely evolved over the last 100 years. The most utilised cognitive ability tests in Australia, such as the *Wechsler* and *Woodcock-Johnson* batteries (James et al., 2015), either remain in paper-and-pencil form, or where administered via iPads, are simply replications of their fixed form. While CATs that measure cognitive ability exist, they tend to be proprietary or unpublished in the scholarly literature. Measurement of many other psychological traits has benefited from technological innovations, but intelligence testing remains quite traditional. Overcoming the lack of advancement in the measurement of cognitive ability will be addressed in this thesis by demonstrating the utility of CATs**.**

1.2 The Current Project

The current thesis details the background, methods, and results of a pilot Computer Adaptive Test (CAT). The underpinning psychological theory for this project was CHC theory, a contemporary taxonomy of human intellectual abilities, which to date (at least in the peer reviewed literature) has not been operationalised using a CAT. Item Response Theory (IRT) is the framework used for measurement. The goal of this project was to design a CHC-CAT screening tool that was transparent in its design, had a pre-determined methodology, and post-hoc analysis and decision making were avoided. By adhering to these principles throughout development, it was believed this could help address the issue of cognitive ability CATs being restricted to private organisations or being left as some side note in a larger methodological study about IRT or CAT. In the development of any new tool or process it is important to establish validity through a rigorous psychometric evaluation and comparison to existing tools assists with establishing such validity. As an exploratory research project, it was hoped that this PhD could explore the application of CATs in the measurement of intelligence as characterised by CHC theory, the efficiency and validity of CATs in measuring intelligence across a wide range of ages and compare such a tool to mainstream cognitive ability tests. Such a tool can continue to be iteratively developed, improved, and ultimately implemented in a variety of contexts. More specific goals based on the background literature are presented at the end of Chapter 2.

1.3 Perspective of the Author

1.3.1 Practical Focus

As both an early career researcher and a registered Educational and Developmental Psychologist, the focus of this thesis is on psychologists and teachers. It was believed that coming from this perspective would help with understanding the practical challenges of implementing screening tools in educational contexts and addressing sometimes uncertain referral questions from schoolteachers. Much of the CAT literature focuses on the medical field, academic achievement, has been limited to research purposes, or has been implemented in a proprietary format. There is a lack of practitioner or teacher friendly cognitive ability measurement tools that rely on CAT technology. The focus on developing a screening tool as a measure of cognitive abilities provided an opportunity to demonstrate the utility of CATs for teachers and psychological practitioners in Australia.

1.3.2 Open Source

Recent trends in many fields of academia and science have promoted the concept of 'Open Science'. The focus of Open Science is to provide open access to research, data, and transparency of methods, whether providing access to statistical analyses or pre-registering planned methodologies (McKiernan et al., 2016). Problematically, most CATs that measure cognitive ability are proprietary in nature which limits the ability to ensure that they are reliable or valid tools. Open science approaches pose opportunities in this field of research to ensure that CATs utilised by schools are transparent and demonstrate enough reliability and validity to make decisions.

The publication of the R code and data sets utilised as part of this project further add to the ability to replicate the analyses included in this thesis (github.com/jakekraska/phd). R is an open-source statistical programming language and also allows the statistical analyses to be repeated in the exact same way as the original author (Mair, 2018). It is hoped that this will allow any researcher the opportunity to evaluate the outcomes of this research.

1.3.3 Transparency

Concerningly, methods to improve the results and outcomes of studies in a post hoc manner are rife in the psychological sciences (Head et al., 2015). It is my perspective that we have a responsibility to the profession as well as the public to ensure we carry out research, particularly publicly funded research (which this PhD would be considered), in a robust and transparent manner. It will also be evident upon reading this thesis that there is an increase in the sophistication of the analyses as the thesis progresses. This reflects my further learnings, implementation of these learnings, and an explicit attempt to avoid manipulating data or statistical analyses after the fact. While the studies in this thesis were not pre-registered, it is believed the evidence presented demonstrates alignment with the underlying principles of the Open Science philosophy and are sufficiently transparent to establish that post hoc analysis was avoided.

1.4 Thesis Outline and Structure

Improvements in information technology, wider availability of CAT platforms, and the never-ending desire for more efficient and valid psychological tests creates an ample opportunity to develop a cognitive ability CAT based on the CHC theory of cognitive abilities. It is the fundamental principles underpinning these advancements that is the focus of this thesis, along with the outcomes of studies designed to develop and pilot such a test.

Chapter 2 covers the background and literature that form the basis of the four studies conducted as part of this thesis. This includes a review of the historical and contemporary perspectives of cognitive ability and the alignment of contemporary test batteries with CHC theory. The chapter then proceeds to compare Classical Test Theory (CTT) and IRT, different perspectives that assist researchers and test developers in analysing their items and scales. IRT underpin CATs, but they also make use of a variety statistical techniques in order to start and stop a test, as well as engage in item selection. The background and critique of CHC, IRT and CAT independently demonstrate the opportunity to integrate these concepts together. A subsequent review of existing cognitive ability CATs is conducted to demonstrate a gap in the literature for a CHC based CAT.

Chapter 3 introduces the Item Tryout Study (ITOS). This first study utilises items developed in four prior research projects to evaluate the qualitative and quantitative validity of the items. A large sample of adult participants were utilised to trial the first set of items, and pilot the possibility of achieving a set of items for four CHC narrow abilities. This chapter implements conservative approaches to psychometrics in order to demonstrate that the items meet the assumptions of Rasch modelling (a somewhat conservative model in itself) allowing for identification of areas for further item development and improvement.

Chapter 4 builds upon the findings in the Item Calibration Study (ICS). Several problems with items are addressed through the development of more items, and the calibration of existing items. This study also included school aged children to show the applicability of the test items to a school setting. The statistical analysis in this chapter focuses on Rasch modelling and preparing items for use in a CAT.

Chapter 5 simulates a CAT utilising computer generated participants. This is based on the findings of the ITOS and the ICS; the simulation utilises the item parameters from the ICS as well recalculating item parameters using only school aged participants.

The study discusses the efficiency and limitations of using the developed items in a CAT format.

Chapter 6 explores the validity of the screening tool under development. School-aged children participants were recruited and administered both the screening tool under consideration and the *Wechsler Intelligence Scale for Children – Fifth Edition* (WISC-V). Correlational analyses are carried out. While there were limitations identified in terms of data collection, efforts were also made to establish a statistically derived measure of *g* based on measures of Lexical Knowledge, Induction and Visualisation to correlate with the WISC-V Full Scale IQ.

Chapter 7 concludes the thesis with a general discussion of conclusions, implications, and recommendations for future research. Focus is placed on the strengths and limitations of the developed CHC-CAT, and how this may be interpreted within the concepts of ethical use of technology and automation of the psychology workforce.

Chapter 2: Background and Literature

The measurement of cognitive ability or intelligence has a long history. Cognitive ability is perhaps “the most researched topic in the history of psychology” (Wasserman, 2012, p. 3). A substantial literature base supports the relationship between cognitive ability and a range of life outcomes, including school performance (Evans et al., 2002; Floyd et al., 2003; Floyd et al., 2008; Walker et al., 1994) and career development (Lang & Kell, 2019; Sternberg, 2003). Such relationships have resulted in efforts to implement assessments and interventions during the school years to facilitate positive academic and career outcomes for students.

Interactions between cognitive ability, academic achievement and other psychological constructs have also been of interest to scholars. Many researchers have investigated whether relations with other psychological constructs such as personality, learning styles or motivation explain the importance of cognitive ability in task performance (Busato et al., 2000; Farsides & Woodfield, 2003; Kraska, 2013; Laidra et al., 2006). However, outcomes of many studies are inconsistent (Wettstein, 2017) likely due to the wide number of contributing factors, and the statistical methodologies used within each study. At the least, a recent meta-analysis found that “the effects of ability and motivation on performance are additive rather than multiplicative” (Chad et al., 2017, p. 270). Thus, cognitive ability measurement has predominately remained a key focus of psychological assessment when addressing academic, behavioural, or day-to-day functioning deficits. Assessment of cognitive ability by psychologists and teachers remains an important aspect of the work that they do, and this chapter focuses on underlying theory and measurement principles that support that testing.

2.1 Chapter Outline

In improving the measurement of cognitive ability, it is important to understand the evolving conceptualisation of intelligence as a construct over a century of research. Any psychological tool should fundamentally be ‘useful’, and this can only be achieved through sound theoretical foundations. The evolving understanding and measurement of cognitive ability is discussed in this chapter throughout Sections 2.2 to 2.5.

Beyond theory, a measurement tool must be based on sound measurement science; researchers, practitioners and the public must be able to ensure that the tool is measuring what it says it is measuring, and that it is doing so in a reliable manner. Not only have our understandings of cognitive ability evolved over time, but so have the sophistication of our statistical and measurement methods. This includes perspectives on Item Response Theory (IRT) and Classical Test Theory (CTT). Conceptualisations of measurement theory and the IRT framework for which CATs are ultimately based on is discussed throughout Sections 2.6 to 2.9.

It is the application of IRT via CAT that poses opportunities for improving measurement of cognitive ability. Section 2.10 to 2.15 introduces basic concepts of CATs and discusses the advantages of and purpose of different characteristics that can be customised when developing or implementing a CAT. The broad advantages and disadvantages of CATs as well as the availability of CAT platforms is also discussed.

Considering the potential opportunities that the integration of CHC, IRT and CAT offers to the measurement of cognitive ability, it is important to understand the strengths and limitations of research in this field. Section 2.16 includes a review of the literature focused on cognitive ability CATs. This section summarises the gaps in the wider empirical works and subsequently Section 2.17 outlines the goals of the four studies conducted as part of this thesis.

Overall, this chapter focuses on drawing together the background and literature of cognitive ability (defining what we want to measure), IRT (how do we measure it), and computer adaptive testing (application of combined frameworks) to demonstrate the opportunity for advancement of the measurement of cognitive ability. This chapter aims to identify opportunities for future implementation of technology and improved statistical methodologies in test administration and development within the field of cognitive ability.

2.2 A Brief History of Cognitive Ability Tests

The measurement of cognitive ability has undergone several changes over the last century but remains relatively uninfluenced by technology, adaptive testing methodologies or contemporary statistical methods. A full history of intelligence testing is beyond the scope of this thesis, but a brief review demonstrates that while there has been advancement of intelligence theory, there has only been slow advancement of cognitive ability measurement tools.

Wasserman (2012) considers many of the efforts to measure intelligence in the 19th century to be pseudoscientific antecedents to the contemporary approaches to intelligence testing. Galton’s Anthropometry eventually led to efforts to create scientifically based intelligence measurement and Cattell’s reaction time experiments. Nearing the end of the 19th century, and challenging of much of Galton and Cattell’s work, Alfred Binet, described by Wasserman as an “innovative outsider”, was conducting experiments with his own children and publishing prolifically before seeing an opportunity in the early 20th century to develop a norm referenced intelligence test as a solution to France’s desire to apply the mandatory public education laws to “abnormal” children. That is, the now well-known *Binet-Simon Intelligence Scale*. This led to refinements in the form of the *Stanford-Binet* (SB), the “most frequently used psychological test in the United States for decades” (p. 20). The SB is now in its fifth edition (Roid & Pomplun, 2012).

Initially thought to be hereditary, it was largely Piagets’ theories that began considering the impact of a child’s environment on intellectual ability. This was subsequently supported by observations of the poor performance of children from lower socio-economic status families (Raiford & Coalson, 2014). The recognition of the interplay between environment and intellectual ability led to significant policy changes in the spheres of disability and education, the production of a vast range of assessment tools, and the importance of reliable and valid tools of intellectual ability assessment in these policies (Raiford & Coalson, 2014). These early pioneers of intelligence tests paved the way for the Army Mental Tests and David Wechsler’s now well-known intelligence scales. The evolution of these tests has largely been dependent on the changing understanding and operationalisation of ‘cognitive ability’ as conceptualised by key researchers throughout the 20th century. This was only the beginning of a long journey of iterative theory development in the field.

2.3 Contemporary Theories of Intelligence

Modern theories of intellectual ability are conceptually linked to the progress made throughout the 20th Century. David Wechsler (1944) defined intelligence as:

> *The aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment. It is global because it characterizes the individual’s*

> *behaviour as a whole; it is an aggregate because it is composed of elements or abilities which, though not entirely independent, are qualitatively differentiable.*

Within this definition is the clear consideration of a global intelligence that is composed of individual but inter-related abilities. Since this seminal definition, intelligence has been operationalised in many varying ways (McGrew, 2009b; Wasserman, 2012).

Acknowledging that any theory is ultimately going to be deficient in exactly describing human intelligence, it is generally accepted that “human intelligence is a melange of many abilities that are interrelated in many ways” (Horn & Blankson, 2012, p. 73). Thus, contemporary theoretical perspectives on intelligence generally reflect this: Multiple Intelligences Theory (Chen & Gardner, 2018), the Triarchic Theory of Successful Intelligence (Sternberg, 2018), and the Planning, Attention, Simultaneous, and Successive cognitive-based processing theory (Naglieri & Otero, 2018). Most contemporary theories of ability are either multidimensional or process driven.

The most contemporary and well-validated understanding of human cognitive abilities is the Cattell-Horn-Carroll (CHC) model (Kamphaus et al., 2018). CHC theory has been described as "a systematic synthesis of hundreds of studies spanning more than a century of empirical investigations of cognitive abilities" (Schneider & McGrew, 2012, p. 100). This theory is a culmination of the Carroll (1993) Three-Stratum theory with Cattell (1971, 1982) and Horn's (1986, 1988) *Gc-Gf* theory. Although originally developed via factor analytic studies it has been validated through developmental, neurological, and biological research (Horn & Blankson, 2012).

CHC theory is a three-tier hierarchical model of abilities, with broad abilities subsuming narrow abilities, and broad abilities being subsumed by general intelligence (*g*) (Flanagan et al., 2013). This theoretical model has error associated with each broad ability, narrow ability and measuring subtest. Such a model is particularly conducive to further research and refinement (McGrew, 2009a). In practice, broad abilities are theoretical constructs, and each narrow ability is measured by a specific subtest. Recent examples include modifications to the model (v2.4) to conceptualise long-term retrieval (G*lr*) into two separate abilities known as learning efficiency (G*l*) and retrieval fluency (G*r*) (McGrew, 2016) and increasing the numbers of broad abilities (Schneider & McGrew, 2018). Broad abilities continue to be explored and argued in support of, such as 'emotional intelligence' (Evans et al., 2019). The latest broad ability model (v2.5) is represented in Figure 2-1, with examples of narrow abilities for eight of the core broad abilities exampled in Figure 2-2.

2.4 Current Challenges to CHC Theory

CHC theory is not without its critics. Predominately criticism relates to the interpretation of *g* versus "group ability factors", such as the broad cognitive abilities included in CHC theory (Beaujean & Benson, 2019). Other concerns include the perception of the proliferation of abilities within the CHC taxonomy as evidence of its own shortcomings (Wasserman, 2019) as well as failures to replicate the CHC structure in a range of tests (Canive*z* & Youngstrom, 2019). Wasserman (2019) raises "twenty challenges to CHC" ranging from concerns surrounding CHC theories basic adherence to scientific principles of parsimony, falsifiability, and replicability, to criticisms of CHC theories conceptualisation of *g,* Gf-Gc, memory, speed and quantitative reasoning.

Figure 2-1. CHC Model (v2.5) with 18 Broad Abilities

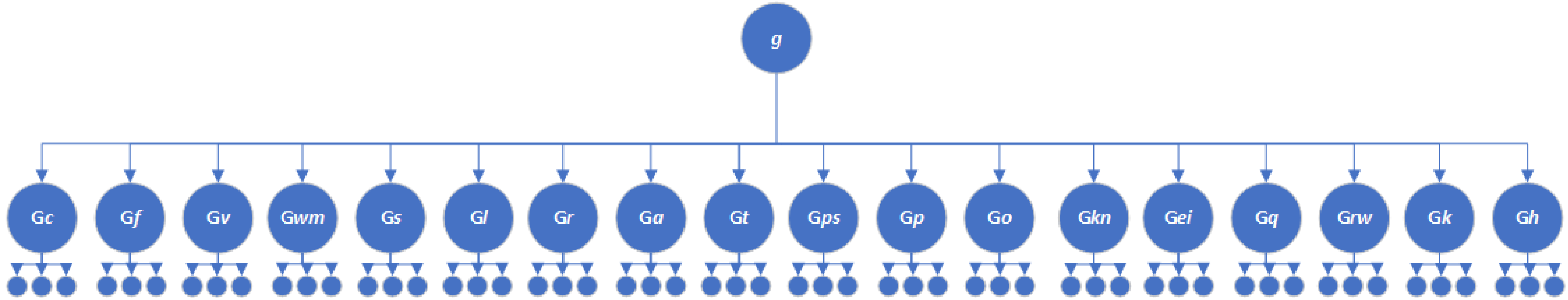
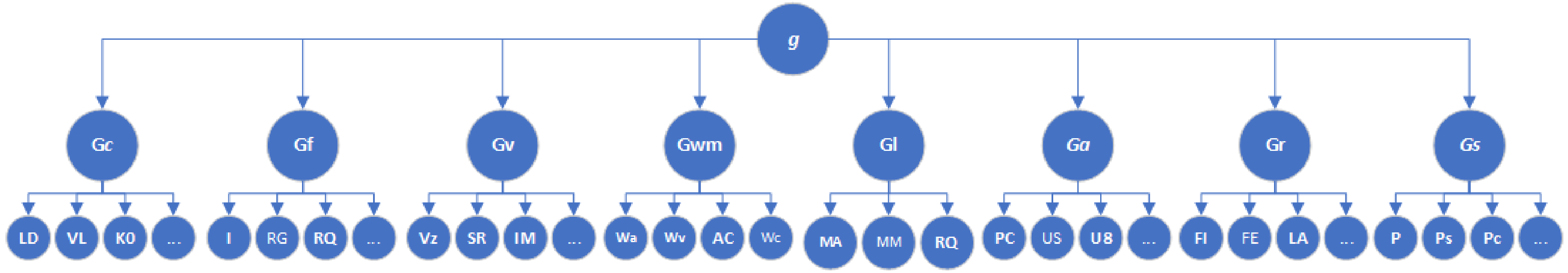


Figure 2-2. CHC Model (v2.5) with Narrow Abilities for 8 Core Broad Abilities



Note. Gc = Comprehension Knowledge, Gf = Fluid Reasoning, Gv = Visual-Spatial Processing, Gwm = Working Memory, Gs = Processing Speed, Gl = Learning Efficiency, Gr = Retrieval Fluency, Ga = Auditory Processing, Gt = Reaction and Decision Speed, Gps = Psychomotor Speed, Gp = Psychomotor Abilities, Go = Olfactory Abilities, Gkn = Domain-specific knowledge, Gei = Emotional intelligence, Gq = Quantitative Knowledge, Grw = Reading and Writing, Gk = Kinesthetic abilities, Gh = Tactile (haptic) abilities. While Version 2.5 of the CHC model includes 18 abilities, focus is largely placed on the 8 core abilities where links have been demonstrated with academic achievement.

These criticisms often imply that the *entire* CHC theory is invalid and should be discarded, which completely discounts decades of iterative improvement in our understanding of the structure of cognitive abilities. Even Wasserman (2019) acknowledges that both Spearman (1927, 1938) and Thurstone (1948), who both started at polar ends of the debate, ultimately acknowledged the existence of both a general factor and "group factors". Some broad abilities within CHC are arguably more "defensible" and this seems implicitly understood by key CHC authors as they have established criteria for updating the theory, as well as labelling certain abilities as "tenuous" (Schneider & McGrew, 2018). Outside of these yet to be fully supported abilities, over a century of research suggests that key abilities measured by tools such as the *Wechsler Intelligence Scale for Children – Fifth Edition* (WISC-V), such as Verbal Comprehension (VC), Fluid Reasoning (FR), Working Memory (WM), Processing Speed (PS) and Visual Spatial (VS) do in fact account for variance outside of *g* (Keith & Reynolds, 2012).

Findings of studies on the factor structure of existing tests often depend on the statistical methodology utilised. For example, initial confirmatory factor analyses by the test publisher of the WISC-V utilised a higher order model (with a general factor indirectly influencing subtests) which resulted in a five-factor model (VC, VS, FR, WM and PS) (Wechsler, 2016). Criticism of this included the use of weighted least squares estimation, preferencing an overly complex model, identification of an FR factor that had a standardized path coefficient of 1.00 with the *g* factor (making FR redundant), and a failure to test "rival" bifactor models (Canivez et al., 2016). Canivez, Watkins and Dombrowski's resulting exploratory factor analysis found support for a four-factor model, with VS and FR combined into a Perceptual Reasoning (PR) factor (Canivez et al., 2016). Subsequent confirmatory factor analysis using nine different theoretical models, including higher order and bifactor models, found that a bifactor model with four group factors (VC, PR, WM and PS) was the best fitting, whereby *g* accounted for most subtest variance (Canivez et al., 2017) and was generally supported across most age groups (6-8, 9-11, 12-14, 15-16) (Dombrowski, Canivez, et al., 2018) and cultures (Canivez et al., 2018; Fenollar-Cortés & Watkins, 2018; Lecerf & Canivez, 2018). Concerned about the impact of initial starting values for the factor communalities or method of bi-factor rotation on results, Dombrowski, Beaujean, et al. (2019) conducted a "sensitivity study" which found the ideal structure was that of five factors (one general, four group; VC, VS, WM, PS). They also found that a single subtest (i.e., Cancellation) can have a substantial impact on the factors extracted. Taking a balanced view there appears to be at least consistent evidence of a WM, VC, and PS factor, and ongoing debate about whether VS and FR are separate, as in a five-factor model, *g* often accounts for variance of FR.

These analyses can only go as far as the subtests that are included in the WISC-*V:* "Intelligence is what the intelligence test measures" (Maas et al., 2014). Therefore, additional subtests with unique task demands may introduce unique variance. The technical manual for the Woodcock Johnson IV Tests of Cognitive Ability (WJ IV COG) (McGrew et al., 2014) suggests that with additional *types* of tasks, seven group factors can be derived beyond the variance accounted for by *g*. This has been challenged in both the WJ III (Dombrowski & Watkins, 2013) and WJ IV (Dombrowski, McGill, et al., 2018), finding a similar four factor alignment as found in their studies of the WISC-V (G*wm*, Perceptual Reasoning, G*s*, G*c*). However, it may be also that the overlap between G*v* and G*f* is due to the use of non-verbal stimuli, suggesting some degree of method error, conflating the relationship and factor structure. The status of a visual processing factor and fluid reasoning factor is a continuing debate.

Some argue that this points to a “replication crisis” (Dombrowski, McGill, et al., 2019). To address this Dombrowski and colleagues utilised Monte Carlo simulations to conduct 1,000 replications of various CFA models of the WISC-V, WJ IV COG, Kaufman Assessment Battery for Children – Second Edition (KABC-II), and Differential Abilities Scales – Second Edition (DAS-II). In all simulations the publisher theory was not the best fitting, with previously discussed bifactor models remaining supported for the WISC-V and WJ IV COG, and higher order models being supported for the KABC-II and DAS-II. For the latter, the models were “essentially consistent with, yet offered superior modification statistics relative to, the test publishers’ theoretically proposed” (p. 8) models. Regardless, the authors acknowledge that nearly all models tested, including both the publisher models and more parsimonious independent models were able to be replicated almost 1,000 times. Whether a bifactor or higher order model is a better fit depends on the test instrument being analysed. And this is only complicated by concerns about the suitability of bifactor models when trying “to represent the general and group factor structure of an entire domain of psychological functioning” (Bonifay et al., 2017, p. 185). It seems unlikely that these outcomes are going to settle the disagreement within the field, as various statistical methodologies continue to permeate. Larger samples and increased representation of various levels of the latent traits under measurement may assist, which seems only likely through the implementation of technological solutions to cognitive ability testing.

Beyond the structure of cognitive ability, the application of these theoretical models is contentious. There is also ongoing debate in the literature about the predictive validity of CHC when relying on broad abilities (McGill, 2017), particularly when attempting to diagnose Specific Learning Disorders (Kranzler et al., 2016; Miciak et al., 2016). In response, Flanagan and Schneider (2016), point out that these authors have confused various terms, such as Cross-Battery Assessment (XBA), Pattern of Strengths and Weaknesses (PSW), and the Dual-Discrepancy/Consistency Model (DD/C Model), which are all separate concepts.

Such criticisms also conflate problems regarding the structure and theory of intelligence with the diagnosis of a heterogeneous disorder. One does not need to cross-batteries to obtain a CHC profile (Flanagan & Schneider, 2016), and in fact XBA guidelines state that this should only be done on occasions where a norm derived score is not available or valid (Flanagan et al., 2013). CHC as a theory has applicability beyond Specific Learning Disorders, in terms of both developing deeper understandings of existing measurement tools as well as further investigation of ability deficits in other neurodevelopmental disorders[1] (e.g. Abu-Hamour & Al Hmouz, 2018; Bench et al., 2019; Jacobs & Costello, 2013; Lemann et al., 2019; Warne, 2015). It was hoped that CHC theory "provides a common framework and nomenclature for intelligence researchers to communicate their findings without getting bogged down in endless debates about whose version of this or that construct is better" (Schneider & McGrew, 2018, p. 73). Despite this, many criticisms continue to advocate for significant modification or cessation of use of the CHC model, based on literature that tends to focus on Specific Learning Disorders.

[1] Reference is made here to "neurodevelopmental" disorders within the context of the DSM-5, as it is acknowledged that profile analysis of intellectual test batteries to derive understanding of psychopathology (e.g. mood disorders, personality disorders etc.) has been thoroughly debunked.

Notwithstanding wide adoption of the CHC taxonomy in academia and research, for CHC advocates there remains a theory-practice gap whereby many cognitive ability assessment tools used in practice are perceived to be based on outdated understandings of intelligence (Jacobs et al., 2013). Insufficiency in test tools to measure a broad range of abilities has been argued to hamper psychologists' abilities to identify *specific* weaknesses for school aged people (Jacobs, 2015). As argued by Flanagan and Schneider (2016), cognitive deficits increase the risk of academic deficits (i.e. probabilistic causation), rather than cause academic deficits (i.e. deterministic causation). There continues to be a need for a comprehensive theory that allows practitioners to measure a wide range of cognitive abilities while having a common understanding of what they are measuring.

Overall, the literature reveals a polarisation of perspective on the *g* versus "group ability factors" debate. As with any necessary scientific progress, regardless of theoretical orientation, only further development or improvement of tools and measures that align with CHC theory will enable researchers to either replicate or improve the CHC taxonomy of abilities. Taking into consideration an amalgamation of giants in the field (such as Wechsler, Spearman, Thurstone, Carroll, Cattell, Horn, McGrew, Flanagan, Schneider) over the last century, who provide substantial evidence for the presence and interpretation of both a general factor *and* broad abilities, it is difficult to conclude that a complete "annulment" (Canive*z* & Youngstrom, 2019, p. 243) of the CHC theory of cognitive abilities is the appropriate course of action. It is important for the development of new tools (such as in this thesis) and the refinement of existing tools that there is a common understanding of constructs amongst researchers, even if there is some disagreement about the finer points of a theory. Let's not throw the baby out with the bath water.

Kamphaus et al. (2012), in their considerations of what might constitute a "fifth wave of intelligence test interpretation" believed that test interpretation would ultimately be impacted by improvements in test technology. It has been almost a decade since their prediction, yet other than moving existing models of paper-and pencil-testing onto PCs, the Internet and tablets, there has been little change in the linear pattern of test administration.

2.5 Current Implementations of Cognitive Ability Measurement

As of 2015, in Australia, the *Wechsler* scales (WAIS-IV, WPPSI-III, WISC-IV) and the *Woodcock-Johnson batteries* (WJ III COG) were the most utilised cognitive assessment batteries used by psychologists that participated in a study on CHC adoption (James et al., 2015). Since then there has been the release of new versions of the WISC (Wechsler, 2016), WPPSI (Wechsler, 2014), and WJ COG (Howe & Dailey, 2015). At current, the Wechsler scales are available to be administered in both paper based and iPad versions (known as Q-interactive), while the *Woodcock Johnson* (WJ) batteries are only available in paper format. While not intended to be CHC tools, both the WISC-V (Flanagan & Alfonso, 2017) and WPPSI-IV (Raiford & Coalson, 2014) align well with, and can be interpreted via, the CHC taxonomy of abilities.

Though there has been some advancement in online administration of cognitive ability assessments, they are largely just replications of paper-and-pencil tests on electronic screens. For example, many tests published by Pearson are now available via their "Q-interactive" system (Pearson, 2013). These tools still predominantly utilise the same norms, basal rules, ceiling rules, test stimuli and score structures as their paper-

and-pencil counterparts and are generally only partly adaptive; specific item sets are often required to be administered, with raw scores still being calculated based on basal and ceiling rules. As far as can be found on others publisher's websites, or in the extant literature, administration of other popular cognitive ability assessments such as the SB and WJ have not, despite their wide use, become available in computer administered format nor adapted to technological advancements beyond online scoring and reporting.

Psychological assessment and research have not yet fully taken advantage of the significant improvements in technology, and continues to be dominated by pen-and-paper tests (Gibbons, 2017). Despite advancements in almost all other industries, the psychological test industry remains stagnant (Barrett, 2018). Cognitive ability tests are highly time consuming and despite being extremely structured (i.e. administered in the exact same way every time), they can only be administered by a psychologist, and most often only in a one-to-one setting. While some other international jurisdictions make use of test examiners, this is generally not the case in Australia outside of the military. Beyond this, there is a utility for a cognitive ability CAT that has high reliability and validity. For example, many teachers are increasingly expected to engage in 'on the fly' differentiation, and a brief cognitive ability CAT can assist with differentiated instruction (Scalise, 2009). There are clearly advantages to a robust CHC measure that can be administered either individually or in a group setting, and that is adaptive in nature to reduce the time commitment of psychologists. It is the analysis and collection of data and the practice of psychological counselling and interventions that has predominantly benefited from technological innovations. However, the *measurement* of latent psychological traits can also be improved via technological solutions.

2.6 Psychological Constructs

When aiming to improve the way in which we measure a latent psychological construct, it is important that researchers ensure that they are measuring what they intend to measure. One of the long-standing issues in psychological research is the measurement of psychological constructs. This challenge is demonstrated by the above discussion of CHC theory. Having a strong definition of the construct under investigation assists with the development and rationale of items.

Lilienfeld et al. (2015) state that psychological phenomena are open concepts and are “characterised by fuzzy boundaries, an indefinitely extendable indicator list, and an unclear inner essence” (p. 1). In their analysis of inaccurate, misleading, misused, ambiguous and logically confused words and phrases in psychology and psychiatry, the authors challenge many of the common terms that psychology researchers often mistake as represented by or related to “good measurement”. For example, they argue that “operational definitions are unrealistic in virtually all domains of psychology, because constructs are not equivalent to their measurement operations” (p. 6), recommending the use of the term “operationalisation” instead. They also contend that the way in which many researchers discuss reliability and validity implies that a test is either valid, reliable or not, but “at best, these measures are ‘empirically supported’” (p. 6). No measurement tool in psychology is without some level of error or inaccuracy, including CATs.

Psychological assessment can be conducted through various means, including tests, interviews, portfolios, case history, behavioural observation, and role play (Cohen & Swerdlik, 2005). With the uptake of computers and the Internet, the ability to integrate technology into assessments has long been known. Zenisky and Sireci (2002) showed that technology allowed innovations to be made in terms of how participants responded to questions, the types of items presented, and the inclusion of media (i.e. graphics, video, and audio) in assessment tools. The ability to measure psychological constructs via technology has long been investigated by pre-employment settings, military organisations and educational institutions (Drasgow & Olson-Buchanan, 1999). Despite this, there has been very little adoption by those that regularly conduct psychological assessment in practice.

Notwithstanding early suggestions of innovation, integration of technology into psychological assessment tools has largely been confined to academia, whereby researchers have utilised online surveys and other electronic tools. In practice, online formats have been limited to the same structure and stimuli as their paper-based origins (Bartram, 2000). This lack of innovation appears to be somewhat related to a continuing predominance of CTT compared to IRT; these underlying measurement theories have implications for how a test is designed, developed, calibrated, administered, and scored. More significantly, these theories, when integrated with technology, have consequences for the "operationalisation" – and therefore validity – of psychological constructs (Wools et al., 2019).

2.7 Classical Test Theory

To advance the measurement of cognitive ability, an understanding of the predominant measurement theories is required. Much of statistical analyses in the 20th century focused on 'classical test theory' (CTT), also known as 'true-score theory' (Embretson, 1995). At its essence is the concept that observed scores on a scale or psychological tool are composed of a person's 'true score' on the trait or variable of interest in combination with error (Iramaneerat et al., 2008):

Equation 2-1. True Score

$$X = T + E$$

where X is the participants observed score, T is the particpants true score, and E is the error in measurement which comes from a large variety of sources.

At the core of CTT is an assumption that all people have a quantifiable amount of each psychological construct. That is, a score on a psychological test or tool is a function of this quantity plus measurement error (DeVellis, 2006). Because psychological phenomenon is subjective, we rely on psychological tools as a proxy for measuring a person's level of a construct or trait. Unlike being able to measure the number of white blood cells in a sample of blood, it is not possible to observe a person's level of depression directly and objectively. Thus, CTT allows us to measure a person's internal state by using self-report or performance measures. If a person responds positively to a question, this suggests that they possess the characteristic of interest in higher quantities than someone who responds negatively (Cappelleri et al., 2014); there is a linear relationship between an observed score and the true score (Rusch et al., 2017). The person has a 'true' level of depression (T), but the tool and other extraneous variables introduce error (E), producing an observed score (X). Because items are imprecise measures of the true score, there is error inherent in its measurement.

The error that is measured in CTT is assumed to be random. The error is "as likely to increase as to lower the observed score for any item" and "errors for items are assumed to be independent of one another" (DeVellis, 2006, p. 51). Because of this assumption, when combined across all items, the random error cancels each other out. Because the amount of measurement error is only known for the full item set, this means that any person who completes the test has the same standard error of measurement (SEM) (Embretson, 1996). This conceptualisation of error has implications for how we statistically derive the reliability of a psychological test's items.

If item error can vary, but the overall test error has a mean of zero (DeVellis, 2006), it is assumed that the SEM is equal for all levels of measurement but can vary for different populations (Embretson, 1996). Test and scale developers can improve the reliability of their tools by increasing the number of items. A good item in CTT is one that correlates well with the person's true score (DeVellis, 2006). Therefore, not only do more items produce better reliability coefficients, but more items at the mean trait level (as they correlate well with the hypothetical true score) are desirable, rather than items at the extreme ends of the distribution. This means that the more similar the items are with respect to the mean trait level, the 'better' the test is according to CTT.

All the items in a set must be measuring the same underlying trait. This is known as 'unidimensionality' (DeVellis, 2006). If an observed score is meant to represent a construct, such as depression, then the items should only reflect that content. While scales or tests can be multidimensional in totality, the items themselves should represent independent constructs that align with a theoretical or statistical model. Such models can usually be demonstrated via Structural Equation Modelling (SEM) – or Confirmatory Factory Analysis (a type of SEM) – and Exploratory Factor Analysis.

While these methods allow scale scores to be calculated that reflect multidimensional constructs, there is difficulty in utilising these raw scores in a meaningful way. In CTT this is often achieved via norm referencing scores (Embretson, 1996). Therefore, when a person obtains a raw score of 23 out of 50 on a depression scale, rather than saying that this person has a depression score of 23, we can interpret this in the context of the normative sample that was gathered during test development.

It may be that only 1% of the normative sample obtained a score of 23 or above, and thus a raw score of 23 implies a high level of depression. This interpretation is based on the sample gathered during normalisation of the scale rather than the raw total score.

To compare an individual's performance on one test to other people's performance on the same test in a reliable way we need to ensure that the normative sample is representative of the population (Embretson, 1996). Using the example above, if the normative sample only consisted of people that never endorsed items measuring 'depression' (i.e., low depression), then it is likely that even very low scores out of 50 would be indicative of high levels of depression even if the test taker had actually only responded affirmatively to a couple of items. In contrast, if the normative sample only consisted of people that always endorsed items measuring depression (i.e. high depression) then it is likely that very high scores out of 50 could result in low standard scores or percentile ranks, erroneously suggesting a lack of symptoms. Thus, each scale developed using CTT must rely on a representative sample of the population. Scores can only be utilised to compare an individual to the population under which the sample was derived. This also means that two tests of the same trait may result in different scores, making comparisons difficult.

In CTT, comparing test forms requires a significant amount of work. Most importantly, there must be equality of means, variances and covariances across test forms (Embretson, 1996). Newer methods of equating test forms have been demonstrated by regressing scores from one another, however this means that equating of test forms is reliant on tests with high reliabilities and similar score distributions (Embretson, 1996). Again, interpretation and scoring are reliant on the sample utilised.

CTT has several shortcomings (Iramaneerat et al., 2008). Firstly, because scales are developed using a specific sample of the population, interpretation of scores are relative to that sample. That is, the difficulty of the items cannot be generalised to other samples that may have different distributions. The second problem with CTT is that missing data can be quite problematic; because the true score is represented by an observed score that is summative of items, then a missing item causes problem for calculation of an observed score. Thirdly, that the SEM is assumed to be the same across all scores in a population is a strong assumption given the reliance in CTT on having more items at the mean level of difficulty/trait than at the extremes; having more items in the middle of the distribution should mean that there is less error measurement at the peak than at the extremes where there is more variability in performance, less people in the normative sample, and fewer items to measure these extreme trait levels. The fourth problem is that all items are assumed to be of equal importance. In a measure of depression, dichotomous yes or no items such as “I think about killing myself” and “I am sad” increase the raw score of the scale by one. In this hypothetical two item scale, a total raw score of one is quite clearly different based on which item the person endorsed, but from a CTT perspective the person is assumed to have the same ‘level’ of depression because they both received a raw score of one. A fifth shortcoming is that because there is an assumed deterministic linear relationship between observed scores and a true score, there are few ways to validate a final score. That is, consistent with the problem of equating all items as equal, if a person with high depression does not endorse an item that represents low levels of depression, even though they may have endorsed many items indicative of high depression, it can be difficult to identify this anomaly. In an overall sense, CTT tends to produce scales and tests that have many items that may be redundant, and the scores from these tools can be difficult to compare to other tests or scales, or to other samples. Many of these problems can be addressed by IRT which is fundamental to progressing the measurement of cognitive ability.

2.8 Item Response Theory

Item Response Theory (IRT), a set of models that focuses on the relationship between latent traits and observed outcomes, has a long history. While both CTT and IRT aim to measure a construct based on some sort of indicator (i.e., item performance), CTT assumes a linear relationship between item performance and a true score, and IRT tends to rely on probabilistic perspectives of whether a person will endorse an item or not (Rusch et al., 2017). IRT allows the calculation of probability of an endorsement or correct item (i.e., probability X = 1 given theta; Figure 2-3 black line), or vice versa, the probability of not endorsing or not obtaining a correct item (probability X = 0 given theta; Figure 2-3 red line). To do this, just as with CTT, IRT has basic assumptions.

When using IRT there are specific measurement properties of items and of respondents that are estimated for scales. Firstly, within IRT the latent trait is represented by the Greek symbol theta (θ) (Yang & Kao, 2014). As in CTT, in IRT the latent trait is assumed to be unidimensional, although there are multidimensional IRT models (Bonifay, 2020). A second assumption of unidimensional IRT is local independence, which means that once the latent trait is statistically accounted for, there are no remaining relationships between the items (i.e., residual dependencies) (Yang & Kao, 2014). Thirdly, IRT assumes that the items demonstrate monotonicity, which means that as a person's level of the latent trait increases, the probability of endorsing or correctly answering an item increases (Bonifay, 2020). Monotonicity is represented in IRT via 'item characteristic curve' (ICC; Figure 2-4) for a dichotomous item and ‘categorical response curves’ (CRC; Figure 2-5) for polytomous items (Nguyen et al., 2014). This is generally represented by an ‘S’ shape curve that shows the increasing probability of a respondent answering a question as their trait level increases.

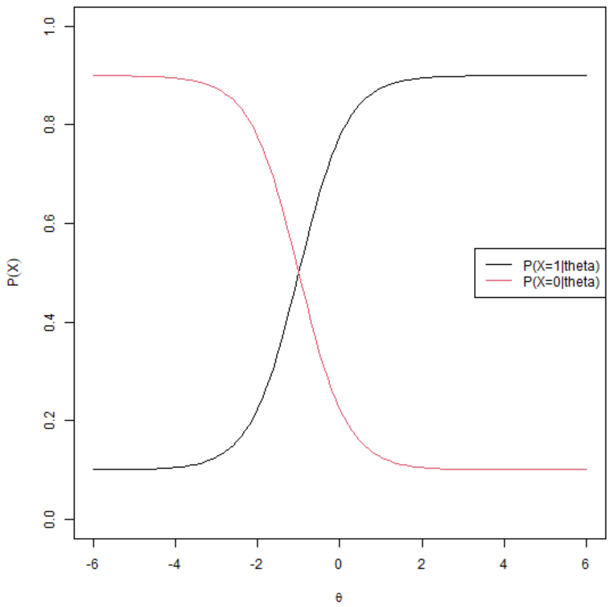
Another concept introduced with IRT is that of information. This is a concept that somewhat replaces ‘reliability’ in a CTT sense. In IRT, more information implies more reliability. However, because IRT demonstrates that reliability varies depending on the level of the latent trait, this means that information varies too (Bonifay, 2020). This results in an item information function (Equation 2-2) that is bell shaped by summing the information from each item into a test information function. This allows direct comparison of tests and how much information they provide, as well as to calibrate items and tests to ensure that tests are measuring effectively at different levels of θ.

Equation 2-2. Item Information Function for the Four Parameter Model

$$I(\theta) = \frac{a_i^2(P(\theta) - g_i)^2(u_i - P(\theta))^2}{(u_i - g_i)^2P(\theta)(1 - P(\theta))}$$

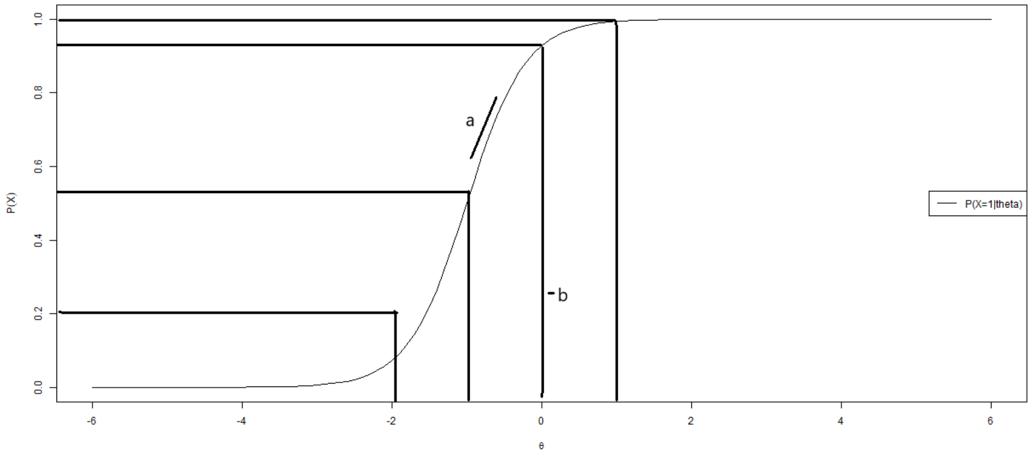
While IRT has existed for almost a century it did not come into prominence until the 1970s (de Ayala, 2009). Initially known as ‘latent trait theory’, there was a divergence in work between Rasch (1960) and Birnbaum (1968). The former focused on invariance and making sense of a theoretical framework, and the latter focused on mathematical models that explain the most variance in the data (Iramaneerat et al., 2008). That is, does the data fit the model versus does the model fit the data? Over time, IRT models have expanded from a focus on dichotomous item formats into ordinal rating scales and multidimensional models (Embretson, 1996). Now, there are a wide variety of models with applications for different contexts and that utilise a variety of different parameters.

Figure 2-3. IRT Probability of Obtaining a Correct Item



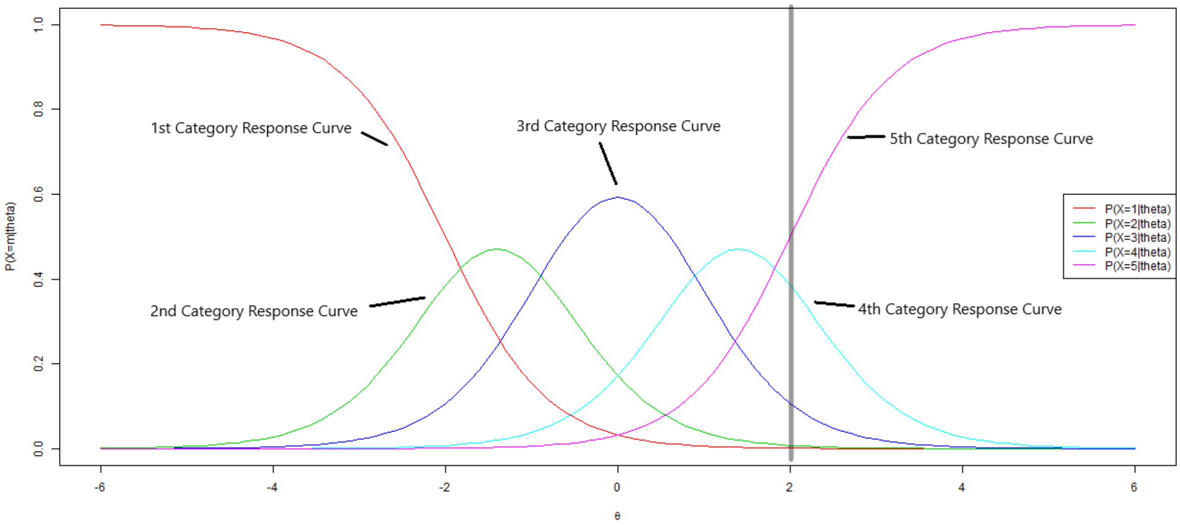
Note. The red curve represents the probability that a person will obtain an incorrect score while the black curve represents the probability that a person will obtain a correct score.

Figure 2-4. Item Characteristic Curve



Note. At -2 theta there is a 20% probability of correct answer, at -1 theta there is an approximate 55% chance of a correct answer, at 0 theta there is an approximate 95% chance of a correct answer and then at theta of 1 and above there is an almost 100% chance of a correct answer.

Figure 2-5. Category Response Curve for Polytomous Items



Note. This Category Response Curve represents an item with five response categories. As a person increases in their level of theta (their internal amount of the theoretical construct), the probability that they will endorse a higher category on the item.

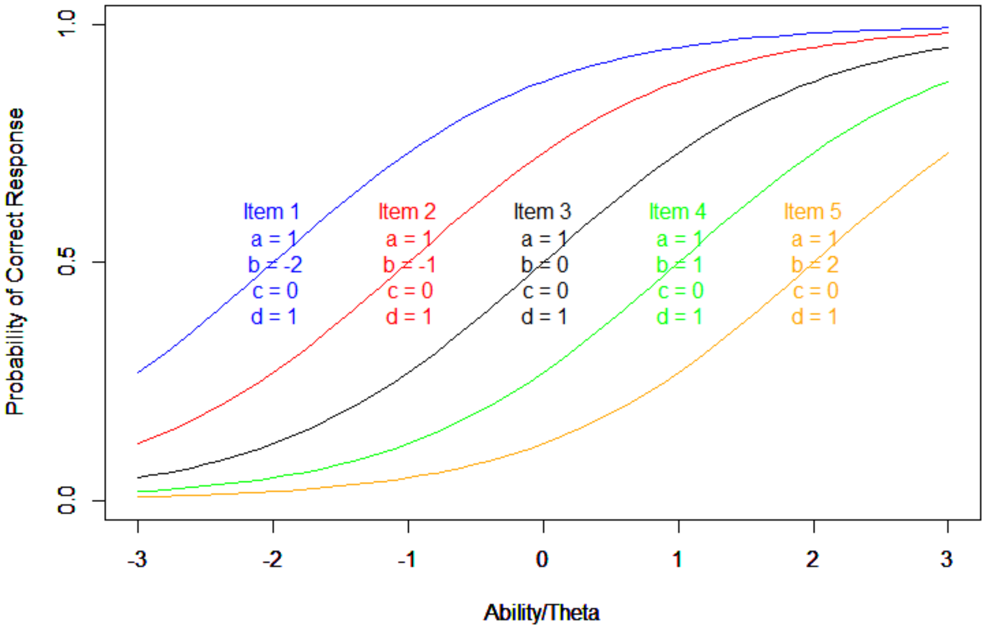
The Rasch model, essentially a One Parameter Logistic (1PL) model (Baker & Kim, 2017), is perhaps the most well-known IRT model. This model focuses on the 'difficulty' parameter. Using the Rasch model it is easy to compare items regardless of the population under analysis, as well as allowing people to be compared regardless of the items utilised. The Rasch model is sample independent, which the other IRT models are not. There is some disagreement about whether the Rasch model and the 1PL model are different in their approach or not (Andrich, 2004). The perspective of the former would be that if the model does not fit, then alterations to the data are needed, whereas the latter would suggest adding additional parameters (Bond & Fox, 2015). Rasch supporters would argue that if items do not fit the Rasch model, they "should be dismissed out of hand for failing to meet the minimal standards required for measurement" (p. 308). In Rasch, the discrimination parameter is fixed. Ultimately, while there are theoretical and conceptual differences between the two approaches, for most practical purposes the models are the same (Institute of Objective Measurement, n.d.).

The probability of a correct answer under the Rasch Model can be calculated using Equation 2-3 where e is the base of the natural logarithm that is a constant 2.718, b is the item difficulty parameter, and θ is an ability level. In this model all other parameters are fixed. While the theoretical range of the difficulty parameter is $-\infty \leq b \leq +\infty$, it is usually set to $-3 \leq b \leq 3$ (Figure 2-6).

Equation 2-3. Rasch Model

$$P(\theta) = \frac{1}{1 + e^{-1(\theta - b)}}$$

Figure 2-6. ICC calculations under the Rasch Model

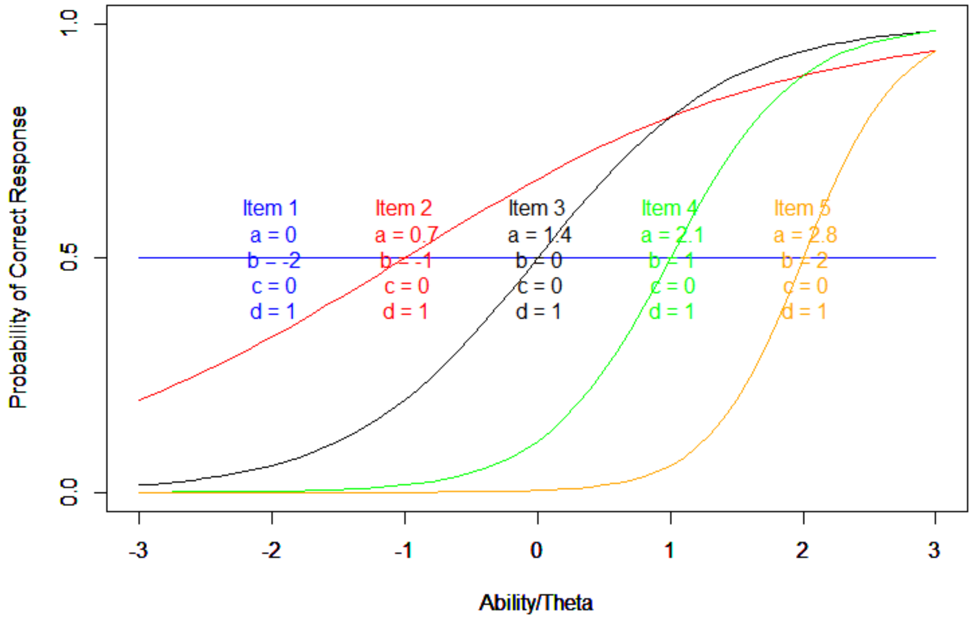


The 2PL model introduces the discrimination parameter which can also be described as the slope of the ICC (Baker & Kim, 2017). This can be thought of as how rapidly the probability of a correct answer increases with ability. The probability of a correct answer under the 2PL model can be calculated using Equation 2-4 where e is the base of the natural logarithm that is a constant 2.718, b is the item difficulty parameter, a is the item discrimination parameter, and θ is an ability level. The range of the difficulty parameter for the 2PL model is the same as in the Rasch model and 1PL model. The range of the discrimination parameter is theoretically $-\infty \leq c \leq +\infty$, it is usually set to $-2.8 \leq c \leq 2.8$ (Baker & Kim, 2017). As the discrimination value increases, the slope of the curve increases (Figure 2-7). Item 1 has a discrimination parameter of 0, meaning there is a 50% chance that any person, regardless of their ability level, may or may not endorse this item. Such an item would not be good for discriminating people of different abilities. In contrast, Item 5 has an item discrimination parameter of 2.8, causing a steep slope in the ICC and thus a strong ability to discriminate between those that are above and below a θ level of 2. While high discrimination has its positives, this item has an extremely low likelihood of endorsement by almost anyone below a θ of 1 and thus depending on the purpose of the test, may not be a desirable item either.

Equation 2-4. 2PL Model

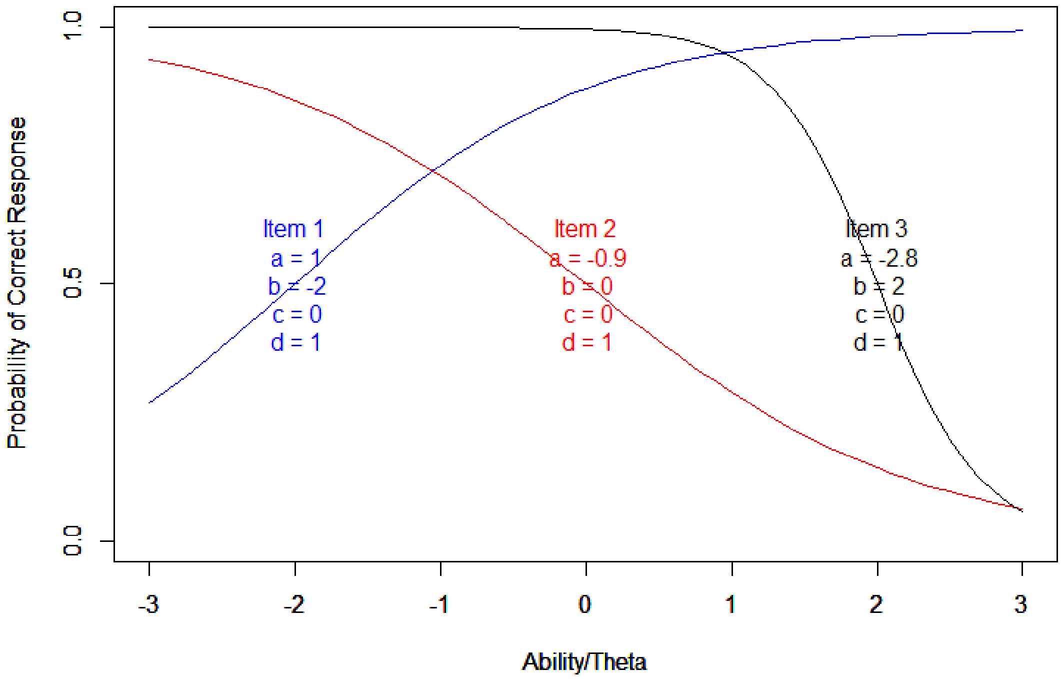
$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}}$$

Figure 2-7. ICC calculations under the 2PL Model



It is possible to have an item with negative discrimination. This may be an indicator of a poorly designed item, because as a person's trait level increases, they have a lower probability of endorsing the item or getting a correct answer. In Figure 2-8, Item a discrimination of 1; in contrast, Item 2 and Item 3 both have negative discrimination values and thus as the person's θ level increases, the person's probability of endorsing these items decreases. This is contrary to the goal of any scale attempting to measure a particular trait or ability. Generally, this can be addressed by reverse scoring an item.

Figure 2-8. ICC calculations comparing negative and positive discrimination in 2PL



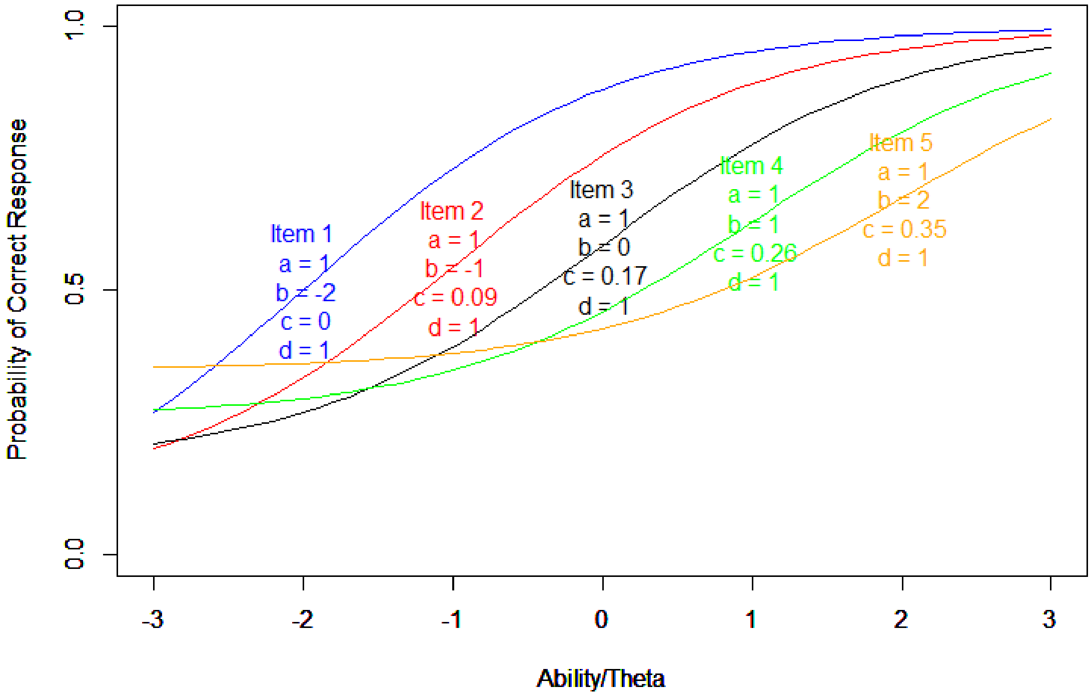
The 3PL model introduces the 'guessing' parameter, representing the contribution of guessing to the probability of a correct response (Baker & Kim, 2017). The probability of a correct answer under the 3PL model can be calculated using Equation 2-5, where e is the base of the natural logarithm that is a constant 2.718, b is the item difficulty parameter, a is the item discrimination parameter, c is the guessing parameter, and θ is an ability level. The c parameter has a theoretical range of $0 \leq c \leq$

1.0, but anything above 0.35 is not considered acceptable (Baker & Kim, 2017). Another way to conceive of the guessing parameter is as a lower asymptote (Primi et al., 2018). As the guessing parameter increases, the lowest point of the ICC rises up the y axis, suggesting a higher likelihood a person will obtain a correct answer (Figure 2-9).

Equation 2-5. 3PL Model

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}}$$

Figure 2-9. ICC calculations under the 3PL Model

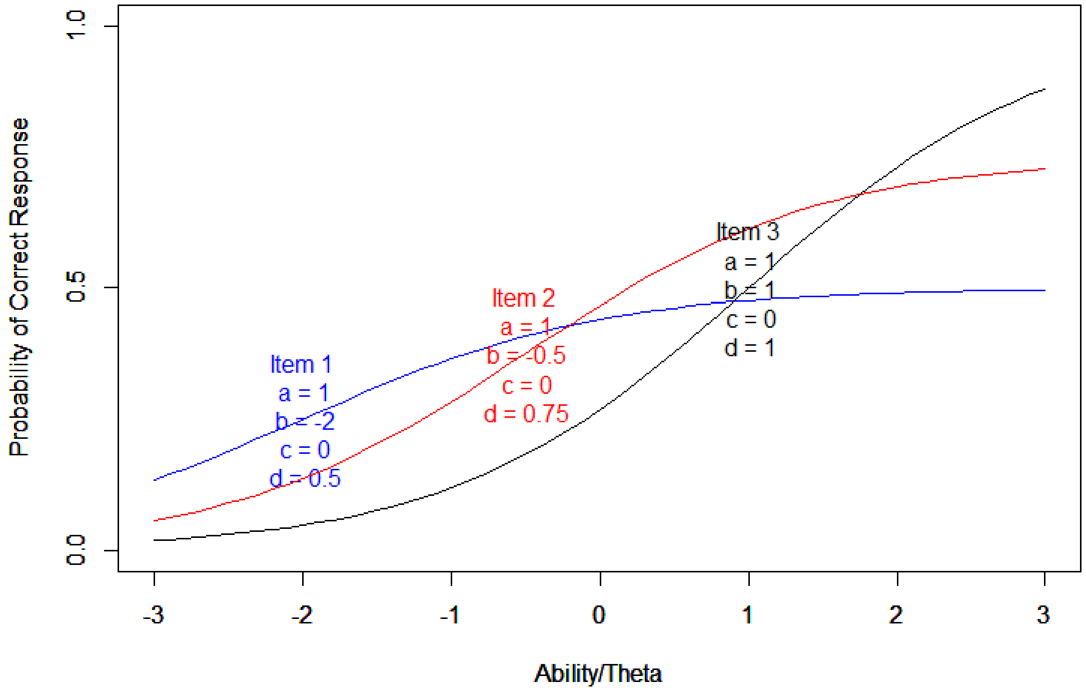


The 4PL model extends the 3PL model by introducing an item specific upper asymptote (Baker & Kim, 2017). This is considered an upper limit on the probability of a correct response. The probability of a correct answer under the 4PL model can be calculated using Equation 2-6 Where e is the base of the natural logarithm that is a constant 2.718, b is the item difficulty parameter, a is the item discrimination parameter, c is the guessing parameter, d is the upper asymptote parameter, and θ is an ability level. In contrast to the guessing parameter, the upper asymptote parameter lowers the top of the ICC (Figure 2-10).

Equation 2-6. 4PL Model

$$P(\theta) = c + (d - c)\frac{1}{1 + e^{-a(\theta - b)}}$$

Figure 2-10. ICC calculations under the 4PL Mode



2.8.1 Other IRT Models

The models discussed thus far are the main ones present in the literature for use with dichotomous items. IRT models for polytomous items do exist. These include, but are not limited to, the Rating Scale Model (Andrich, 1978), Graded Response Model (Samejima, 1997), Rasch Partial Credit Model (Masters, 1982), Generalised Partial Credit Model (Muraki, 1992). IRT models for polytomous items are particularly useful for scales designed to measure mental health constructs (e.g., depression, anxiety, stress) or personality constructs (e.g. ways of thinking, extraversion), but are less useful for tests of academic achievement or cognitive ability where there is generally one correct answer and a range of incorrect answers. It is also possible to convert polytomous response options into dichotomous response options. This is useful if there are small samples and a parsimonious model such as the Rasch model is preferred (Nguyen et al., 2014).

2.9 Comparing CTT and IRT

Ultimately both CTT and IRT are attempting to model latent traits. The advantage of CTT is that it tends to be mathematically simpler and well understood. Most practicing psychologists are likely to have been taught the basic assumptions of CTT and thus can easily understand tools that have been developed in such a manner. Despite this, IRT offers advantages in being able to scale people and items on the same metric, and generally allows for stronger claims to be made about the items and people's performance on those items.

One common criticism of IRT is the requirements for large sample sizes. Such samples are often difficult to obtain in "low-incidence or difficult to sample populations" (Finch & French, 2019, p. 78) and thus methods to address such limitations are required. Finch and French compared the Maximum Likelihood (ML) Estimation, Markov Chain Monte Carlo and Pairwise methods of estimating item difficulty, discrimination, and pseudo-chance parameters and found that the latter two were often more robust in situations of small sample sizes, even as low as 100 participants. While this is relatively recent research, there is an increasing body of research in the psychological sciences about the practical utility of IRT methods in differing circumstances. Such findings allow researchers to apply IRT techniques to cases where participants are difficult to recruit.

Rather than suggesting IRT is more utilitarian than CTT, it seems more logical to select statistical methodology based on the goals of the research and intended audience rather than biased personal preferences. Petrillo et al. (2015) demonstrated a working example of using CTT and IRT measurement theory and found that each approach was relatively consistent in the analysis of a patient reported outcome measure. Each methodology produced additional information about items, including poorly fitting items, poor targeting and disordering of response categories. Ultimately the authors recommended that CTT analyses in isolation are only applicable in situations of an "instrument being developed for descriptive purposes and on a restricted budget" (p. 33). Whereas IRT and Rasch modelling is more appropriate in 'high-stakes' situations but should be still supplemented with CTT approaches. While the current project may be on a restricted budget, cognitive ability tests are within the boundaries of high stakes and thus this thesis will make use of differing balances of CTT and IRT throughout.

IRT is particularly useful when developing tests that need to measure a wide band of a latent trait for decision making. Cognitive ability is one such construct; researched for over 100 years, cognitive ability has proven to be a multidimensional construct with extreme differences in ability between young children through to adults. CTT has consistently been utilised to refine theory about cognitive ability, demonstrating the reliability and validity of different cognitive ability batteries. The integration of IRT with tests of cognitive ability poses interesting opportunities to further improve scoring accuracy and thus make better decisions about individuals. The use of IRT in measures of cognitive ability is not novel (particularly in development) but is hardly mainstream (particularly in scoring).

2.10 Advancing Measurement Using Computer Adaptive Testing

Over the past few decades, no industry has gone unaffected by technology, and this includes psychology and the mental health industry. Technology has been implemented to great effect in automating processes, creating efficiencies, improving the fidelity of treatment, and easing accessibility. This includes making improvements in data collection (Hamilton & Bowers, 2006), the analysis of data (Press, 2013), changing the nature of psychological intervention and conjunctive therapies through online therapy or use of Apps (Hawn, 2009; Hides, 2014; Kavanagh, 2014; Kyrios & Thomas, 2014), disseminating mental health information and challenging stigma (Christensen, 2014), and automating business processes (*Cliniko*, 2020; *HealthKit*, 2018). While technology has been implemented into cognitive ability testing, mainstream tools in Australia have been largely stagnate. It is evident that there is still much to discover about the utilisation of technology in psychological practice and research. It is this context that provides an important basis for which to justify further advancements in measurement methodologies, particularly in the measurement of cognitive abilities.

In the literature, a variety of recent technologies are proposed as solutions to the issue of psychological measurement innovation. The most prominent amongst them are ecological momentary assessment (EMA), computer adaptive testing (CAT), gamification, and machine learning. At times there is a significant overlap between these assessment methodologies.

Gamification and machine learning are particularly contemporary concepts in assessment, with little literature evidence prior to the turn of the century. There is some initial research surrounding the use of gamification (Nikolaou et al., 2019; Tong & Chignell, 2014), stealth assessment (the use of games and other activities to measure constructs without the examinees' knowledge) (Reichenberg, 2018) and machine learning to measure cognitive ability (Dawadi et al., 2013; De Marco et al., 2017). However, they tend to focus on how performance on one measure relates to other measures (i.e. predictive validity), rather than demonstrate construct validity in other ways. In addition, these studies tend to suffer from heterogenous and small samples (Lumsden et al., 2016). If the research does use a more traditional psychological construct in its study or large samples, it often considers intelligence at a reductionist level. For example, Kosinski et al. (2013) utilised the Ravens Matrices as a measure of *g*, which has little applicability outside of identifying general cognitive difficulties. Ultimately gamification and machine learning measures of cognitive ability are currently less useful for utilisation by teachers or psychologists who are required to identify skill deficits or strengths in children and adults and adapt curriculums or provide recommendations accordingly.

EMA relates to the use of measuring a construct in real time (Stone & Shhiffman, 1994). While a useful concept for teachers and psychologists, such tools are inadequate without psychometric validation (Gibbons, 2017). Additionally, EMA is of most use for constructs that are expected to respond to intervention (e.g., pain, depression), but are less useful for relatively stable traits such as personality and cognitive ability. It is these latter traits that tend to be of interest in schools and educational settings.

These attempts at technology integration into psychological measurement are interesting and innovative but tend to be atheoretical, led by the technology rather than strong psychological theory. The use of IRT and CAT allows the integration of psychological theory (such as CHC theory), permits consideration of important concepts such as reliability and validity, and while not implemented widely in Australia, has an array of literature that supports its usage.

Computer testing via the Internet has been considered an option since at least the 1990s. The desire for early CAT researchers was heightened accessibility, ease of scoring, decreases in errors, ease of updating test materials, reduced geographical impact, higher cost efficiency, and the implementation of dynamic graphical interfaces (Barak, 1999). Compared to machine learning, gamification and EMA as solutions to the stagnation of psychological assessment innovations, CAT allows conceptualisation of a construct that can be measured using performance-based items such as those required in a cognitive ability test. Despite early considerations and predictions about the Internet and CATs, outside of private organisations (Gibby et al. as cited in Kantrowitz et al., 2011), government/military agencies (e.g. Segall & Moreno, 1999) and project/research specific considerations, CATs have been underutilised in psychological practice. With significant improvements in Internet and computer technologies in the past decade, there is now further opportunity to create customised CATs that measure cognitive ability.

2.11 Characteristics of CATs

Adaptive tests are not necessarily new. Alfred Binet's original intelligence test was considered 'adaptive' (Weiss, 2011). According to Weiss, five characteristics differentiate adaptive tests from conventional tests. Firstly, an item bank with known psychometrics is required. While both conventional and adaptive tests require an item bank, conventional tests tend to rely on items that focus on a specific level of trait; in contrast, an adaptive test can measure a wide range of a latent trait because not all items are required in order to be administered. Second, prior information can be used to inform an examinee’s starting point. Conventional tests tend to rely on the examinee completing every item from first through to last. Third, adaptive tests require the scoring of items as they are administered, and the ultimate test score can differ based on the subset of items given to each examinee. Contrastingly, conventional tests rely on total test scores derived from the same item set for every examinee. Fourth, each adaptive test has a rule that determines the next item to be administered based on the examinee’s previous item responses, whereas a conventional test administers items in a linear fashion. Lastly, an adaptive test ends when a certain ‘termination criteria’ is reached. This means that the last item administered can differ for every single examinee, and this increases in variability based on decisions the test developer has made regarding the first four adaptive test characteristics.

At first look some of these characteristics may suggest that modern cognitive ability and academic achievement tests such as the Wechsler and WJ scales are adaptive in nature. While these tests may utilise ‘basals’ and ‘ceilings’ in order to skip items based on an individual’s age, grade, pattern of incorrect responses, or some other known characteristic, the calculation of a total score still relies on the same summative raw total of items regardless of the items that the examinee was exposed to. That is, for example, even if the examinee was only exposed to 23 of 42 items in an item set due to having met a basal and a ceiling, there is an assumption that the examinee obtained a correct answer for every item prior to the basal and incorrect answer to every item after the ceiling. From a statistical perspective this implies every item within the scale is on a Guttman scale rather than a Mokken scale. The former implies that there is a deterministic relationship (very strong assumption) between performance on every item, while the latter implies a probabilistic relationship (more realistic assumption).

Additionally, the SEM is assumed to be the same for all individuals of a particular normative group regardless of the number of items they have completed; whether you have completed 23 items or 37 items of a 42-item set, you are assumed to have the same SEM because your score is calculated out of 42. In contrast, a truly adaptive test can calculate the SEM based on the actual items you have completed (Gershon & Cook, 2011) and not an assumption of performance on other items. Ultimately, while modern cognitive ability and academic achievement tests may appear partially adaptive, their basic disadvantage is the continued reliance on concepts of CTT.

While there are significant advantages to even a basic adaptive test, there are limitations to a paper-based adaptive test. They require individual administration by a psychologist, require all items at a certain level to be administered before adaptation occurs and are as such only partially adaptive, and bear no method for controlling score precision (Weiss, 2011). The introduction of computers, however, addresses these issues, turning adaptive tests into CATs.

CATs possess several characteristics that enhance the efficiency of psychological measurement while attempting to avoid compromising test validity or reliability. CATs use adaptive algorithms to select items to ensure precise measurement (Scalise & Allen, 2015). CATs allow us to “simultaneously model the level of underlying construct that a person has, and the level of the underlying trait that the item or questionnaire assesses” (Gibbons, 2017, p. 2). That is, a CAT places the items and the person on the same difficulty/trait scale, represented by theta (θ), in the same way that IRT does; however, this occurs after *each* item administered. The integration of computers with adaptive tests provide several customisation options that allow CAT developers an opportunity to create innovative test tools for the measurement of a variety of psychological constructs, including cognitive ability.

These tools are intimately linked with IRT. According to Gibbons (2017), a CAT's accuracy can be mathematically expressed in terms of standard error, item information or reliability. Weiss (2011) explains that "information in IRT replaces the concept of reliability in classical test theory" (p. 2). Because tests that are based on CTT depend on a high number of items with similar difficulties, this reduces the variance of each item but increases the variance of the total score. This results in an increased variance to number of items ratio, and subsequent increases in reliability. Additionally, removal of items that have low correlations with the total score also increases this ratio, and thus the reliability of the test. This produces a single reliability value that represents a set of items. Due to this procedure of item selection tests based on CTT tend to measure well at the mean of the latent trait but measure poorly for individuals that deviate from that point. CATs utilising IRT pose a significant opportunity to practitioners wishing to measure a construct via fewer items and a high level of reliability across the *spectrum* of that construct.

The characteristics of a CAT are somewhat like that of a paper-and-pencil adaptive test, with some improvements. The main characteristics of a CAT are that it has an item bank, there is methodology for selecting the first item presented to the participant, there is a statistical methodology of calculating θ for the examinee after each item administration, a methodology of selecting the next item to be administered, and some sort of stop rule (Figure 2-13). The latter four characteristics are generally algorithms that collaborate to allow delivery of the CAT (Thompson & Weiss, 2011).

Some CATs have additional constraints placed on them, but these are generally optional. Each characteristic is discussed individually below.

2.11.1 Item Bank

Items within a CAT based IRT can include a variety of item difficulties for which each item can have a different amount of information (as discussed above and displayed in Figure 2-11 and Figure 2-12). In the first example, the curves represent a low difficulty item (blue) and a high difficulty item (red) that both have high information. The second example shows a more realistic example of how a CAT might be visually represented whereby Item 1 (red) produces the most amount of information and Item 7 (aqua) produces the least. The information of an item may differ depending on which IRT model is used in calculating item psychometrics. The advantage of a CAT over paper-and-pencil is that the items are in an electronic format which can be efficiently accessed by computer algorithms.

Figure 2-11. Item Information Curve

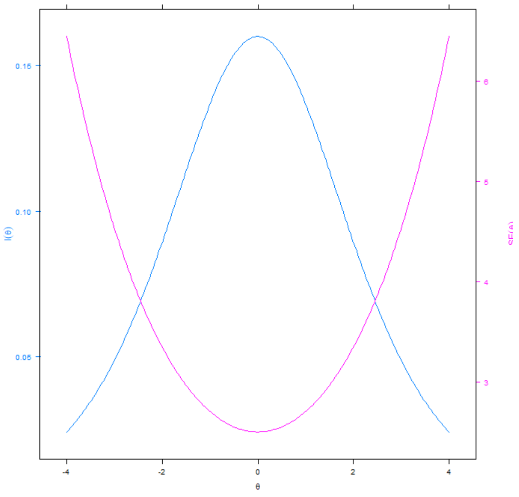


Figure 2-12. ICCs for Theoretical Assessment

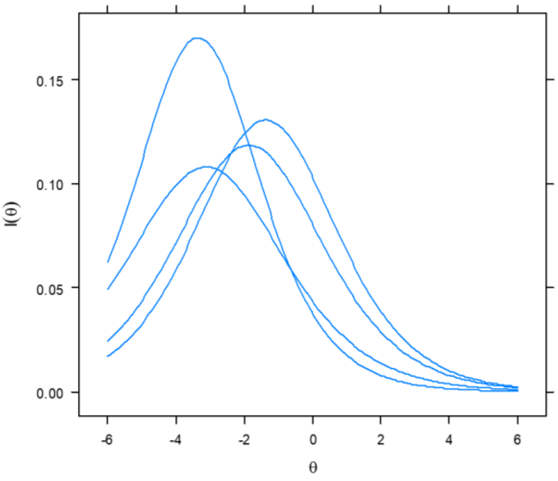
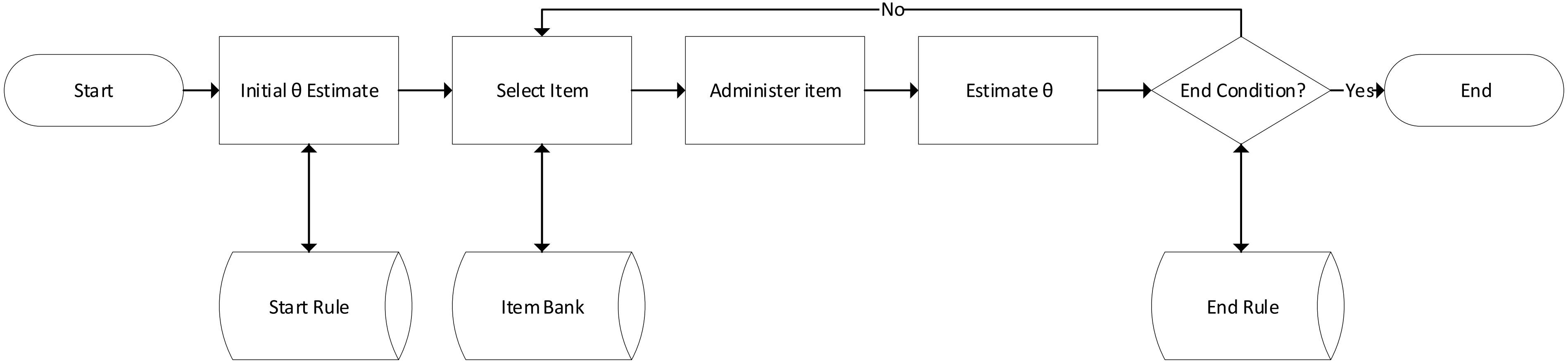


Figure 2-13. Characteristics of a Computer Adaptive Test



Note. Computer Adaptive Tests possess these common characteristics. Many steps involve the use of various algorithms that are constantly being refined within the research literature to make CATs more effective and reliable.

2.11.2 First Item Selection

For CATs, the first item administered can be based on the item psychometrics, an algorithm, or selected either randomly or using prior information. Choosing the first item via item psychometrics may include selection based on specific difficulty parameters, or some other psychometric characteristic. For example, perhaps all participants are administered the item that is closest to mean of θ. Algorithms that can be used for first item selection correspond to those algorithms that are utilised for selecting subsequent items (as discussed below in 2.11.4) (Chalmers, 2016). Randomly selecting the first item means each examinee has a higher probability of being exposed to different initial items, which is useful for test security; according to Weiss (2011) “CATs can recover quickly from incorrect starting points” (p. 12). Lastly, using prior information is likely to improve the efficiency of a CAT greatly; this can be done by using an estimation of an examinee’s trait level on the construct of interest, age, gender, or some other known characteristic for which there is a useful starting point in the CAT item bank.

2.11.3 Estimating ϑ

IRT allows the estimation of an examinees θ after the administration of any set of items. For example, an examinee would have different θ calculations for any pattern of answers for four different items (i.e., 1-1-0-0 versus 1-0-1-0 versus 1-0-0-1) (Weiss, 2011). Implementation of IRT models into CATs therefore allows the calculation of θ after administration of each individual item. After each correct answer, the θ estimate increases, and after each incorrect answer the θ estimate decreases; the additional advantage being that as more items are administered, the SEM of θ decreases. The SEM is calculated based on the items administered, and not the full item set.

The most utilised algorithm for CATs is ML estimation. Introduced in early IRT research (e.g. Lord, 1986), ML is a method of estimating a probability distributions parameter. For CATs, this estimation method allows the use of "all information in an examinee's responses in conjunction with the information available for each test item" (Weiss, 2011, p. 12). The other popular estimation method is Expected a Posteriori (EAP) Estimation, which is closely related to the ML estimation method (Chen et al., 1997). Being able to include all the respondent's previous responses into a θ calculation after the administration of each item allows for increased precision of the estimate and ensures the item selection algorithm can correctly select the next appropriate item.

2.11.4 Item Selection Methodology

Item selection methodologies of CATs are predicated on the idea that if an examinee provides a correct answer then the next question should be more difficult, and if the examinee provides an incorrect answer, then the next question should be easier (Chang, 2015). However, the actual method of achieving this process is varied. Early methods used pre-structured item pools, which were limited in their adaptability and did not use all of the examinee's response pattern to select the next item (Kingsbury & Zara, 1989). With the increase in performance of computers, more procedurally complex methodologies began to emerge. This began with Fred Lord (1980) who applied the 'Robbins and Monro process' (1951). This allowed a test developer to calculate item information separately from test information and update the latter at each stage of the test. It also identified the next appropriate item based on the performance of the examinee on every single item administered previously. Another mainstream implementation in CAT item selection was that of Bayesian item selection, which relied on a normal prior distribution and, after each item administration, a recalculation of a posterior distribution of trait estimates – or, “the item that reduces the posterior variance to the smallest value is chosen [next]” (Kingsbury & Zara, 1989, p. 363). Since this early research, interrelated but arguably conceptually similar methodologies have dominated in item selection procedures for CATs.

Significant improvements in computer technologies bring forth discussions of different algorithms for item selection, largely focusing on improving CAT efficiency and accuracy. Item selection methods for unidimensional CATs include, but are not limited to, Maximum Information (MI), Minimum Expected Posterior Variance, Maximum Likelihood Weighted Information, Maximum Posterior Weighted Information and Maximum Expected Information (Chalmers, 2016). There is also a variety of item selection algorithms specifically developed for multidimensional CATs (Chalmers, 2016). These include D-optimality, Kullback-Leibler Information Index, Mutual Information, Continuous Entropy Method, Modified Method of Posterior Expected KL Information, and the Modified Method of Continued Entropy (Tu et al., 2018). Suffice to say many of these methods have a wide literature base explaining their mathematical principles when they are useful and the strengths and weaknesses of each. Generally, these methods are somewhat related and ultimately aim to reduce the number of items that require administration by choosing the item that provides the maximum amount of information possible.

Outside of specific algorithms it is also possible to simply administer a random item that is within a predefined distance from the previous item. For example, Lun*z* et al. (1994) used a method whereby subsets of items were selected at random within a certain target difficulty, *then* during item presentation the CAT selected two items in the background, one for an incorrect answer and one for a correct answer. At the time, such a method was necessary to ensure there was no lag time between item presentation and item choice. Since then, computers have advanced exponentially in their ability to calculate complex algorithms in a fraction of a second once an examinee has entered an answer. Item selection continues until the CAT meets a stop or termination rule.

2.11.5 Stop Rule

The stop rule for a CAT is utilised to finalise the CAT after a predetermined condition. The most basic stop rule is to have a maximum number of items administered; each examinee is administered the same number of items, but a potentially different set. A more commonly adopted stop rule is to set a minimum SEM; this relies on the CAT stopping when the test has reached a pre-determined level of reliability (i.e. their response pattern is stable enough that the estimate of their ability is reliable enough). This is particularly useful for when a test is designed to measure an individual's level of some trait (Weiss, 2011). A third stop rule is by determining a cutoff θ score; the CAT will continue until the examinee's θ (with included confidence band) is entirely above or below the predetermined cutoff. This enables test administrators to classify examinees into different groups with less focus on precision of scores. The next stop rule is only applicable for CATs that are readministered to measure the change in level of trait between the two test administrations. To achieve this the CAT can continue item administration until the error bands of θ at the two different times no longer overlap (indicating a significant change in the investigated trait), or until a sufficient number of items have been administered where it is clear the error bands will continue to overlap (indicating no significant change) (Nydick & Weiss, 2010).

The most appropriate stop rule for a particular CAT often depends on the test characteristics. Babcock and Weiss (2013) evaluated several stop rules and found that stopping a CAT based on a minimum SEM was useful when there were many items. They also suggested the possibility of combining different stop rules, and that 10-15 items should be administered as a minimum for dichotomously scored items. A minimum SEM rule combined with a minimum item information level would be appropriate for peaked tests. They also found the change in θ stop rule was generally as efficient as other stop rules for repeated administrations. Ultimately, stop rules should be implemented based on the purpose of the tool and the psychometrics of the items.

2.11.6 Constraints

Constraints are optional when designing a CAT. An unconstrained CAT relies on the characteristics, or psychometric details, described above (Weiss, 2011), whereas a constrained CAT includes deliberate design decisions that may run counter to the algorithms and core CAT components. The purpose of such constraints is often to ensure item security, manipulate the length of the test, or ensure the test is measuring all desired aspects.

Georgiadou et al. (2007) identifies over 20 exposure control methods, grouped into five types of strategies: randomization, conditional selection, stratified, combined, and multiple stage adaptive test designs. These methods are useful for ensuring test item security while still attempting to maintain the efficiency advantages of CATs.

Content balancing is another potential constraint. This involves ensuring a certain percentage of items from different content areas within an item set are administered (Kingsbury & Zara, 1989). For example, within a psychological distress tool, it may be that the examiner wants to ensure a certain number of items related to physiological, cognitive, and affective depression are administered. Although the items alone may not be required to obtain an accurate θ score for overall distress, they may be useful from a treatment or diagnostic perspective.

As a practical example of the implementation of constraints, *mirtCAT* (Chalmers, 2016), a free open-source CAT platform, provides test designers with the option to select a range of constraints. The platform includes the ability to include non-scored items (for use in testing experimental items), exclude certain items (e.g. for retesting of certain examinees and to avoid previously administered items), select independent items (e.g. to avoid certain pairs or groups of items appearing in the same testing session), and a choice between ordered and unordered (e.g. for administration of an ordered or randomised group of items after administration of a specified individual item). A combination of these constraints can be applied at once.

2.12 Applications of CATs

While there has been a limited utilisation of CATs (relative to tests developed using CTT), where CATs have been implemented, they have been found to be effective in their goal of measuring certain constructs with more efficiency.

Recently, there has been an increase in use of CATs in patient reported outcome and health related quality of life measures in the medical and rehabilitation fields (Gibbons et al., 2016). This has included the use of CATs for joint awareness based on the Forgotten Joint Score test (Giesinger et al., 2013), emotional functioning using items from the EORTC Quality of Life Questionnaire (Petersen et al., 2016), self-report shoulder functioning (Hart, Cook, et al., 2006; Wang et al., 2010), lack of appetite (Thamsborg et al., 2015), lumbar spine impairments (Hart, Mioduski, et al., 2006), impact of asthma on quality of life (Stucky et al., 2014), health related quality of life across different stages of HIV disease (Revicki & Cella, 1997), and general physical functioning (Haley et al., 2006). Gershon and Cook (2011) state that “patients can typically complete a [patient-reported outcomes] type CAT in an average of five questions with reliability similar to a typical 15-25 survey measure” (p. 1450).

Gershon and Cook argue that the principles underpinning CATs are like that of machine learning, whereby the use of a minimal set of variables could also be used to predict patient-reported outcomes. However, medical research into machine learning techniques has the benefit of incorporating a range of measurements that psychological measurement may not have access to. For example, Tighe et al. (2011) utilised variables such as prior medication use and a basic pain rating scale. In contrast, when attempting to predict psychological constructs, due to their latent nature there may not be prior observed behaviours or prior observations of the construct.

Research into the military’s use of CATs has been longstanding. As early as the 1980s the United States Department of Defence began investigating the use of CATs for aptitude testing (Kathleen et al., 1984; Weiss, 1985) and non-cognitive tests (Stark et al., 2014; Stark et al., 2012). Again, many of these studies suggest the use of CAT is not only feasible but investigates a variety of non-standard item types (such as pairwise preference items). Despite this, these studies also suggest that ongoing evidence needs to be established regarding validity, and further investigation related to “test design, (content, dimensionality, item composition), proctoring, warnings, and cognitive load” (Stark et al., 2012, p. 482). Despite some publications by military agencies across the world, unsurprisingly there is limited insight into the workings of their CATs. This further demonstrates the need for a transparently designed CAT in the Australian context.

Multidimensional CATs (MCATs) have also become popular. Makransky and Glas (2013) demonstrated the applicability of MCATs with the NEO PI-R whereby a reduction of 120 items was possible while still retaining the tool’s accuracy. Bass et al. (2015) utilised MCATs with Patient Reported Outcomes Measurement Information System datasets and found increased reductions compared to unidimensional tools.

Mental health is also an area of investigation for CATs. Fliege et al. (2005) developed a CAT out of 144 items across 11 questionnaires designed to measure depressive symptoms; they found that latent trait could be measured with approximately six items. Devine et al. (2016) utilised CATs in conjunction with the GAD-7, PHQ-9 and PSQ to assess depression, anxiety and stress, finding that similar levels of precision were found with approximately 5-7 items. Some research also considers the adaptation of existing mental health questionnaires, such as the Center for Epidemiologic Studies Depression Scale (Loe et al., 2017; Smits et al., 2017) and Beck Depression Inventory (Gardner et al., 2004), into CATs. In these cases, CATs have demonstrated increased efficiencies while maintaining precision of measurement.

CATs are widely used in achievement and educational testing. In Australia this discussion is largely dominated by the adaptation of NAPLAN into a computer adaptive test (Martin & Lazendic, 2018; Thompson, 2017). CATs that measure cognitive ability have been developed, but they tend to focus on assessment of geriatric patients (e.g. Konsztowicz et al., 2011; Lebedeva et al., 2015; Wouters, Zwinderman, et al., 2009) and automatic item generation in specific domains of cognitive functioning (Arendasy et al., 2011; Hines, 2018). There is significant opportunity for CAT utilisation in the educational and intellectual assessment domains.

2.13 CAT Platforms

While researchers and organisations wanting to use CATs can design and implement them using the basic mathematical principles of IRT in combination with programming knowledge, this can be a time consuming and complex task due to the technological requirements and software development expertise needed. To address this, there is a range of commercial or open source platforms available.

Given the wide range of design decisions available to test developers, Oppl et al. (2017) developed a list of generic requirements which a CAT platform should support. These are:

1. Flexibility in testing strategy and item pool design: The platform should be able to be adaptable to a range of IRT models, item psychometrics, item set sizes, order of items and other test design factors
2. Flexibility in item selection algorithm: The platform can draw from a variety of item selection methodologies (as discussed above), but there is also an ability to use different methods for the first item to be selected versus any following items.
3. Flexibility in specifying the stop rule: Variety in the use of different stop rules must be allowed, as well as allowing for a combination of stop rules.
4. Possibility of technical integration with learning platforms on different layers: This involves the platform being able to integrate with other technologies including functionality in managing data, external platform integrations, and user interface adaptability.
5. Ability to display and evaluate items stemming from arbitrary domains: The authors describe this as ensuring the “presentation and evaluation of items that require domain specific display and data representation”; the platform allows for a variety of response styles that are consistent

with the content being tested, including how the examinees interact with the items.

Oppl et al. (2017) evaluated six platforms: Concerto, IRT-CAT, CAT-MID, SIETTE, MISTRAL, a platform developed by Duda and Walter (2012), and a platform developed by Huang et al. (2009). They found that none met all the identified requirements. Concerto was the best performer (so was utilised in this thesis), meeting the first three requirements, with the fifth requirement reliant on how creatively HTML and other web coding is used to provide response options for examinees. While they proposed potential solutions to this gap, there has not yet been wide adoption of any single CAT platform.

There are several commercial options available to test developers, many of whom will also provide psychometric consultation and CTT to IRT conversion. ASC (2020) provides test developers a platform to host and manage CAT administration of their pre-existing item sets. Pearson VUE (2020a), Prometric (2020), and McCann Associates (2019) provide content design, access to CAT hosting and usage by examinees.

The problem for many of the platforms discussed thus far is that they are difficult to access due to significant technical knowledge requirements, are limited in their functionality, or have substantial cost barriers. Another issue is that platforms are not updated. OSCATS (Open Source Computer Adaptive Testing System) hasn’t been updated since 2011 (Culbertson, 2011). IRT-CAT has not been updated since 2014 (huhs & yoonani, 2014). Since these platforms have stopped being developed, we have seen the increasing implementation of HTML5 (W3C, 2014), wider development and adoption of APIs (Boyd, 2017), and the replacement of Adobe Flash Player with HTML5 standards (Adobe Corporate Communications, 2017), to name just a few web technologies. While six years may only be a short while in academia, this is a product ending timeline in technology.

There are two CAT platforms that integrate well with web technologies, have a continuing development regime and are open source. These are *Concerto* (Scalise & Allen, 2015) and *mirtCAT* (Chalmers, 2016). Both have ongoing development and demonstrate increasing adherence to the requirements listed above. Both platforms still require some technical knowledge for implementation, however there are extensive knowledge bases and how-to guides from the platform developers and the public. Both platforms are based on the R programming language, a free software environment for statistical computing (R Core Team, 2020). While the *Concerto* platform relies on PHP and *mirtCAT* relies on *Shiny* (an R package), both allow the utilisation of HTML5, CSS, and JavaScript to customise the user interface. These platforms are quite flexible to different test developer needs and present opportunities for open source, public domain CATs. Use of these tools in the development of a cognitive ability CAT would allow for easy implementation by practitioners (with the support of IT professionals) and researchers because the underpinning technologies are not locked behind a paywall.

2.14 CAT Advantages

2.14.1 Unsupervised Administration

Since CATs administer an almost unique set of items determined by the individual's own ability, one key advantage is that the security of the test is enhanced. That is, because individuals are not exposed to all the items, and even those that they are exposed to are in a non-predetermined order, it is more difficult to remember item answers. Subsequently, this allows for increased use of tests reliant on CAT technology in unsupervised settings. This is particularly prevalent in the space of organisational and pre-employment testing and enables hundreds of thousands of participants to be tested on their ability, personality and knowledge (Kantrowitz et al., 2011). This is often referred to as unproctored testing. While unproctored testing brings about other challenges with a plethora of associated literature (Aguado et al., 2018; Cavanaugh, 2018; Nesnidol & Highhouse, 2018), implementation of large item sets and CATs ensures examinees are unlikely to receive the exact same item sets, or even the same items in the same order.

2.14.2 Efficiency

A CAT provides the ability to administer fewer and more relevant items. There has been a substantial increase in evidence that CATs can greatly reduce test length and improve reliability. For example, Fliege et al. (2005) was able utilise items from 11 mental health questionnaires to develop a CAT to measure depressive symptoms. They argue that "there is almost no appreciable difference between the total 640 item test score and the [computer adaptive test]-score, which is based on an average of approximately six items" (p. 2289). Gibbons et al. (2016) used simulated CATs relying on IRT to reduce the 100 item World Health Organisation Quality of Life questionnaire by between 45% and 75% depending on the pattern of responding. In these cases, there is clearly an advantage to CATs. On the other hand, Delgado-Gomez et al. (2016) found that in reducing the Personality and Life Event scale to predict suicidality using CATs or decision trees, decision trees outperformed the CAT in reducing the number of items required. Unfortunately, there is insufficient research into both decision trees and CATs to take a confident stance on their reliability in clinical settings.

2.14.3 Precision

As has been discussed in several sections of this thesis, CATs provide opportunities to measure individuals precisely at various levels of a trait or difficulty spectrum without having to administer copious numbers of items. Well-designed CATs have large item sets that assist with targeting the test at the examinee's individual trait level. In contrast, fixed tests tend to provide the best precision at the mean ability level and are less precise at the extreme ends of the trait (Weiss, 2011). Given the uncertainty of a person's cognitive ability prior to testing, an intelligence test using CAT technology allows for a quick adjustment of the difficulty of items and a well targeted measurement of their ability.

2.14.4 Construct Validity

CATs do not rely on what some argue are outdated perspectives regarding the nature of psychological constructs and their measurement (Barrett, 2018). Due to their reliance on IRT, many researchers are supportive of CATs simply because of their opposition to what they perceive as deficits in CTT or a lack of progress in psychological measurement. CATs specifically allow examiners to avoid reliance on the assumption that all items are of equal measurement value. Through this avoidance, construct validity can be further established; completing a harder item implies you have a higher trait level, and thus further demonstrates that the items possess construct validity. Because items are administered via computers, there is an increased variety of item types available (Zenisky & Sireci, 2002)

In contrast to the contention that CTT-developed tests are outdated, some suggest that CTT-based paper and pencil tests, and IRT-based CATs, are useful for different purposes; Cappelleri et al. (2014) suggest descriptive assessments using CTT are suitable when attempting to gather information about the content validity of an instrument, whereas IRT-based tests are suitable when the sample size is larger, and the construct being measured is well understood. Intelligence is a well understood construct.

2.14.5 Automated Scoring

While not unique to CATs specifically, the fact that they are administered on computers and related devices means items are immediately scored. Utilising algorithms built into the CAT, a θ score for each examinee is available immediately after testing. This may include classification of performance (e.g., pass/fail or some other grouping) or comparison of previous scores. Such functionality enables unique opportunities to provide quick feedback to examinees with related interpretations and recommendations.

2.15 CAT Limitations

While the literature on CATs is generally positive and suggests they can address many concerns relating to CTT, there are several limitations.

2.15.1 Limited Research

There are still substantial gaps in the literature when it comes to CATs. Validity of psychological constructs remains a contentious area and proposing to administer items in a non-linear fashion and in differing quantities poses further challenges to this debate. Further, there are many choices to be made when both developing and administering a CAT. Kantrowitz et al. (2011) suggest further research is required relating to the validity of CAT in selection settings, the effectiveness of different item exposure strategies, cross-cultural applicability of the tests, candidate perceptions of fairness, and various implementation challenges. While significant research continues to occur in the interim, there are still many gaps in the literature regarding the use of CATs with various psychological constructs, populations and in differing settings.

2.15.2 Perceptions of CATs

There are several concerns that CATs may fail to address test anxiety (Colwell, 2013) and may cause a decrease in test-taker motivation and self-confidence (Frey et al., 2009). In contrast to concerns about motivation, Martin and Lazendic (2018) used a large Australian sample (*n* = 12,736) of Year 3, 5, 7 and 9 school students and found there were test-relevant motivation and engagement improvements in CATs relative to fixed order tests, particularly in Year 9 students. These positive effects were also achieved while maintaining measurement precision. As a relatively recent and large study, this counters earlier concern about the use of CATs and their perceived fairness, but further research is still required.

Other concerns relate to the difficulty of items; CATs tend to focus on presenting items the examinee has a 50% chance of getting correct, which some argue is too difficult (Kimura, 2017). This can be addressed by incorporating constraints or custom functions designed to keep difficulty levels lower. An example of this is provided by Phil Chalmers, developer of *mirtCAT (2017)*, whereby constraints can be placed on the CAT so that only items that the examinee may have an 80% chance of obtaining a correct answer are presented. Such methods may improve examinee confidence and thus perception of CATs in general.

2.15.3 Complexity

The development of a CAT is complex and requires the establishment of a range of characteristics. While both CATs and CTT depend on items that can cost substantial sums (Downing, 2006a), because CATs are automated, the items must be valid and reliable while also scoreable in real time. Depending on the IRT model, sample sizes required for CAT development may be substantial; a three-parameter IRT model may require a sample of between 500 to 1,000 participants (Yoes, 1995). Additionally, some technical know-how is required to design a CAT that is accessible by examinees. Even with the use of pre-packaged CAT systems, unless developers want to pay for a commercial product (e.g. ASC, 2020), some knowledge of programming languages, relational databases, user interface design and other software related concepts is required (discussed in 2.13).

2.15.4 Calibration of the Item Pool

Outside the cost of developing items, the item pool requires calibration. This generally requires statistical analysis of both IRT and classical statistics. Depending on whether the items are for a new item bank or linked to an existing scale, there are various methods available to ensure the items are calibrated on a common scale (Thompson & Weiss, 2011). While statistical analysis is required in any test development, educators and subject matter experts are often concerned about the loss of domain coverage or confused by the statistical concepts underpinning IRT (Kimura, 2017).

The purpose of the test also impacts the item pool calibration. If the test requires measurement of a wide range of traits, then a range of item difficulties is required to ensure enough level of precision can be achieved; demonstrating construct validity in such cases requires careful statistical analysis. On the other hand, if the test is designed to classify examinees then the test needs to measure very well at a specific point of θ and other items can be sacrificed (Thompson & Weiss, 2011).

2.15.5 Item Repetition

Because an item needs to be scored immediately after administration to select the next, the option of reviewing items presents challenges for CATs. Depending on which CAT platform is used, examinees may not be able to repeat questions. Some have even argued that if item reversal is permitted, examinees might identify whether they have obtained a previously incorrect or correct answer based on the difficulty of the item they are presented with afterwards, potentially encouraging them to go back and change their response (Rudner, 1998). This also presents an issue when re-administering a test. Despite these potential concerns, models have been developed to detect compromised items (Liu et al., 2019)

2.15.6 Implementation of Time limits

By their very nature, fully adaptive CATs administer a different number of items for each examinee, meaning it is impossible to always predict the amount of time required for the examinee to sit the CAT. If a maximum time limit on the CAT is set too low, examinees with trait levels or abilities that are not consistent with the item set within the CAT are likely to be administered more items than the average examinee. In these situations, examinees may be inaccurately classified or scored, rendering the time limit unsuitable.

2.16 Cognitive Ability CATs

While acknowledging that some niche CATs exist, the previous discussions show that despite the proposed advantages of CATs the measurement of cognitive ability has barely evolved beyond turning traditional pen-and-paper tests into electronic formats. Furthermore, the progress of science in relation to intellectual abilities has been hampered by proprietary measures, which also makes research that relies on commercial tests an expensive process (Condon, 2015). As far as can be found in the extant literature, non-proprietary CATs have not been used to measure cognitive ability as per the CHC taxonomy. The following sections detail a review of the literature in this space.

2.16.1 Search Strategy

Electronic searches were undertaken in June 2020 in MEDLINE (via Ovid), ERIC (via Proquest), Scopus, and PsycInfo (via Ovid) to identify relevant articles. Terms related to cognitive ability, computer testing, and adaptive testing were combined using the 'AND' operator; within each set of terms the 'OR' operator was used. Search results were restricted by document type: Original articles, trade journals, scholarly journals, conference proceedings, books, editorials, magazines, working papers, theses, reports, and systematic reviews were included; newspapers, wire feeds, blogs, podcasts, conference reviews, short surveys, letters, notes, historical newspapers, and other sources were excluded. Only articles written in English were included.

Code Snippet 2-1. Search Term Methodology

```
("intelligence" OR "cognitive ability" OR "iq" OR "CHC" OR "cattell horn
carroll" OR "cattell-horn-carroll" OR "lexical knowledge" OR "verbal
comprehension" OR "comprehension knowledge" OR "crystalli*ed
intelligence" OR "fluid reasoning" OR "inductive reasoning" OR "working
memory" OR "visual spatial" OR "visuali*ation" OR "visual processing" OR
"processing speed" OR "perceptual speed") AND ("computer* adaptive" OR
"unproctored" OR "cat")
```

2.16.2 Exclusion and Inclusion Criteria

Many references were removed due to being unrelated to CATs; for example, many references related to the 'California Achievement Test', 'Cognitive Adaptation Training', 'Crying cat syndrome', 'Cognitive Abilities Test', 'Category Test', 'Cat swarm optimisation', 'Clinical Assessment of Attention Test', 'Digit Cancellation Test' (D-CAT), ‘Communication Assessment Tool’, or ‘CAT scans.’ Many sources discussed various computer based or unproctored measures of cognitive ability, or somewhat related psychological constructs such as executive functioning or academic achievement, but these were not adaptive in nature and thus were excluded.

Articles evaluating CAT development in areas other than cognitive testing (e.g., academic achievement, patient-report outcomes, or personality testing) were not included. Many search results surrounding CATs related to measurement of academic achievement which includes word reading, word decoding, reading comprehension, spelling, sentence writing, written expression, numerical operations, applied mathematics, science knowledge, specific knowledge domains (e.g., geology, history), or academic fluency. For example, Cokely et al. (2012) developed a CAT for numeracy and risk literacy, and Martin and Lazendic (2018) found educational testing that makes use of CATs can lead to better measurement precision and increase test-relevant motivation, but neither evaluated psychometric *g*, theoretically driven measures of CHC abilities, nor their own defined intelligence constructs. In fact, some literature argued for the development and implementation of CATs given the many gaps that currently exist in their application (e.g. Kantrowitz et al., 2011).

Some articles focused on adaptive tutoring systems. Where possible, articles were included if they focused on a well-defined construct within intelligence research or CHC theory but excluded if they made ambiguous statements about the nature of the psychological construct under investigation. For example, Pelánek and Jarušek (2015) utilised concepts of IRT to develop an adaptive tutoring system for ‘problem solving’ (graphs, programming, logic puzzles) that would not be considered a measure of intelligence from the perspective of CHC.

Articles that focused on development of cognitive ability measurement tools were only included if they focused on *development* or *use* of a CAT to measure cognitive ability. Using IRT methods for reduction of lengths of tools or measuring improvement in response to training were not included. Methodological studies on certain aspects of CAT (e.g., item selection, ability estimation, stop rules) were excluded. Additionally, any book chapters or articles that simply described the function and utility of CATs were also excluded.

Studies that focused on cognitive training were only included if the focus was on measurement of an ability rather than improvement. For example, Roberts et al. (2016), Flak et al. (2019) and Brehmer et al. (2009) all evaluated the outcomes of Cogmed, a working memory intervention that adaptively increases difficulty of training; to do this, it would be logical to assume the program is tracking the users progress on the latent trait, working memory, through their performance on individual items. However, these studies do not provide any information regarding the adaptive functionality and this could not be found via Google searches, literature searches, or evaluation of the Cogmed product page (Neural Assembly, 2019). Therefore, these studies (and others like it) were not included.

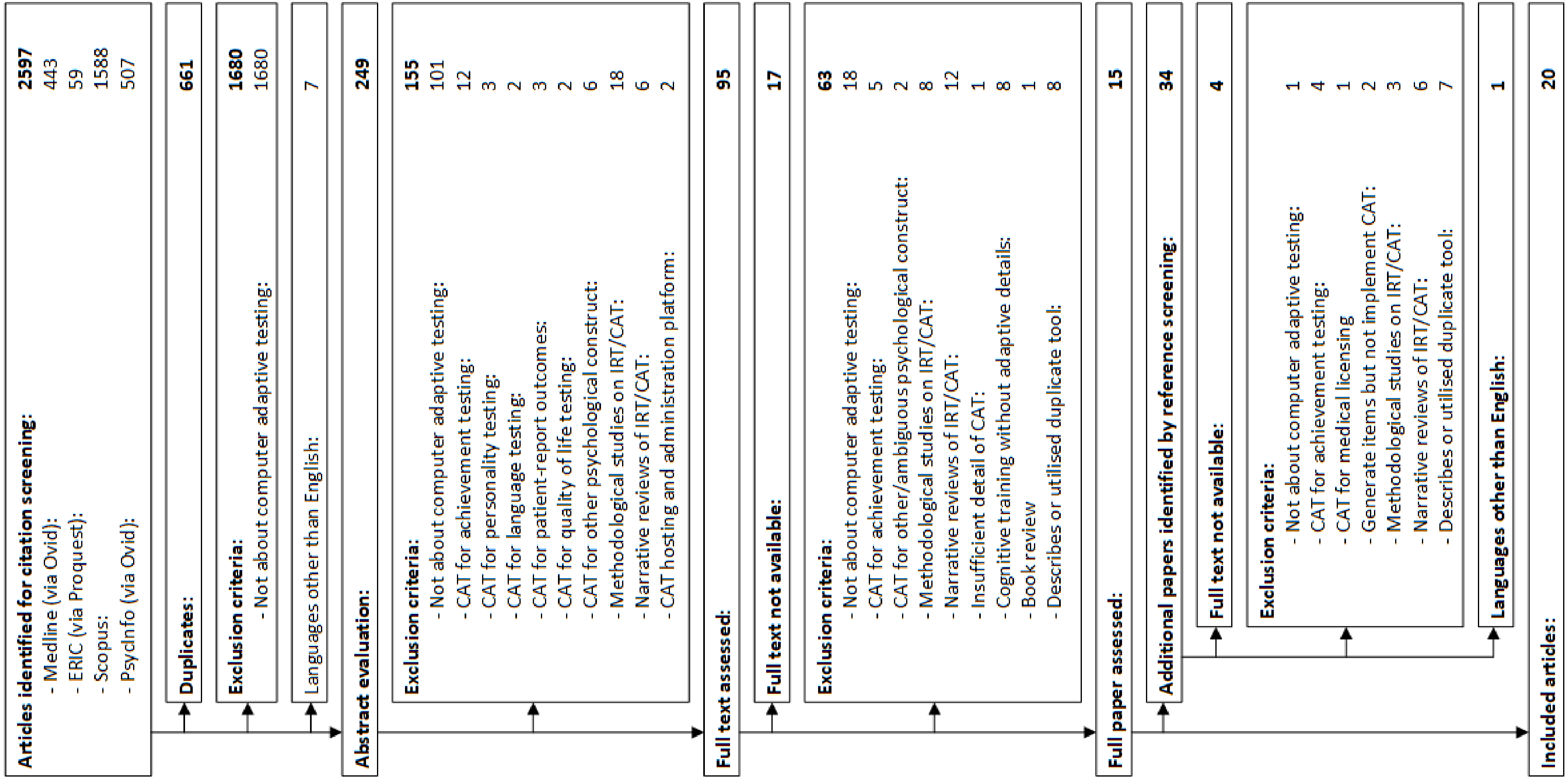
Some references were retained for full text evaluation due to it being initially unclear from the abstract whether they were computer adaptive measures of cognitive ability. Reference lists in each retained full text were also evaluated for possible additional papers (Greenhalgh & Peacock, 2005). Where possible, studies which utilised a CAT but did not provide details about its structure were discarded in favour of the source that detailed the development of said CAT; in some cases, these articles were already obtained during the initial database search, and in other cases reference lists had to be consulted. This was intended to ensure accurate reporting of each identified CAT and reduce duplicate discussions of the same tool. An example of this is the CAT-ASVAB which was utilised in several studies, but few details were provided regarding its development or underlying structure in these studies, usually being used as a measure of *g*; in this case, a single source focused on its underlying features was retained.

Several full text articles could not be located via Google Scholar or university library database searches. These were predominantly international sources, with a majority being from mainland China related to ‘cognitive diagnosis’ models. Other sources appeared to be from Spain, Romania, Germany, and Italy. Evaluation of the abstracts suggested that while they may have focused on cognitive ability, there was unlikely to be a comprehensive intellectual ability CAT underpinned by CHC theory identified in these sources – many of them *appeared* to be methodological in nature, focusing on information systems and artificial intelligence theories related to CAT.

2.16.3 Results

A comprehensive search of the literature for CATs that measure cognitive ability resulted in retaining 20 sources for evaluation (Figure 2-14 and Table 2-1). This comprised book chapters (*n* = 1), conference proceedings (*n* = 1), dissertation (*n* = 1) and original journal articles (*n* = 17). Sources ranged in date from 1997 through 2020, with a range of scopes such as psychometrics, intelligence, psychological assessment, business intersection with psychology, health psychology, learning, and individual differences. While all sources were in English, a variety of jurisdictions were involved including the US, Germany, the Netherlands, South Africa, the UK, Turkey, Hong Kong, Taiwan, and the Chinese Mainland. Evaluation of these 20 sources suggests that there are several gaps for future research into the use of CATs in measurement of cognitive abilities.

Figure 2-14. Literature Review Outcomes



Articles identified for citation screening: **2597**
- Medline (via Ovid): 443
- ERIC (via Proquest): 59
- Scopus: 1588
- PsycInfo (via Ovid): 507
Duplicates: **661**
Exclusion criteria: **1680**
- Not about computer adaptive testing: 1680
Languages other than English: 7
Abstract evaluation: **249**
Exclusion criteria: **155**
- Not about computer adaptive testing: 101
- CAT for achievement testing: 12
- CAT for personality testing: 3
- CAT for language testing: 2
- CAT for patient-report outcomes: 3
- CAT for quality of life testing: 2
- CAT for other psychological construct: 6
- Methodological studies on IRT/CAT: 18
- Narrative reviews of IRT/CAT: 6
- CAT hosting and administration platform: 2
Full text assessed: **95**
Full text not available: **17**
Exclusion criteria: **63**
- Not about computer adaptive testing: 18
- CAT for achievement testing: 5
- CAT for other/ambiguous psychological construct: 2
- Methodological studies on IRT/CAT: 8
- Narrative reviews of IRT/CAT: 12
- Insufficient detail of CAT: 1
- Cognitive training without adaptive details: 8
- Book review 1
- Describes or utilised duplicate tool: 8
Full paper assessed: **15**
Additional papers identified by reference screening: **34**
Full text not available: **4**
Exclusion criteria:
- Not about computer adaptive testing: 1
- CAT for achievement testing: 4
- CAT for medical licensing 1
- Generate items but not implement CAT: 2
- Methodological studies on IRT/CAT: 3
- Narrative reviews of IRT/CAT: 6
- Describes or utilised duplicate tool: 7
Languages other than English: **1**
Included articles: **20**

Table 2-1. Literature Search

Author	Instrument Name	Study Objectives	Setting and Sample	Latent Trait	IRT	CAT Software	N Items	Simulated/ Real World	Starting Item	Stop rule	Item Selection	Theta Estimation
Arendasy and Sommer (2012)	INSBAT	Evaluate factor structure of INSBAT items in combination with automatic item generated items	Unknown setting 12 years to 77 years old *n* = 481	Quantitative Knowledge, Fluid Reasoning, Comprehension Knowledge	1PL	?	166	Real	?	?	?	?
Arendasy and Sommer (2017)	Figural Inductive Reasoning Mental Rotation Verbal Fluency Arithmetic Flexibility	Vary item administration format and retest form to reduce effect size of retest effect	Unknown setting 16 to 64 years old *n* = 960	Inductive Reasoning Visualisation Verbal Fluency Arithmetic	1PL	TestWeb 2.0	FID: 120 ELT: 120 V*F:* 100 A*F:* 100	Real	Mean item difficulty	Max items FID: 14 items ELT: 14 items V*F:* 17 items A*F:* 13 items	MI	?
Balas-Timar and Balas (2009)	MAB-II	Integrate Fuzzy Logic into CAT estimation	Engineers *n* = 200	Gc, Gv, Gs, Gwm (Jacobs & Costello, 2013)	1PL? 3PL?	?	?	Real	?	Minimum SE?	MI?	MAP?
de Beer (2005)	Learning Potential Computerised Adaptive Test	Develop culture free test measuring non-verbal reasoning	Multicultural Grade 9 and 11 students from South Africa *n* = 2,454	Non-verbal reasoning	3PL	MicroCAT	188	Simulation	Mean item difficulty	Minimum SE Maximum number of items	?	?

Author	Instrument Name	Study Objectives	Setting and Sample	Latent Trait	IRT	CAT Software	N Items	Simulated/ Real World	Starting Item	Stop rule	Item Selection	Theta Estimation
Hausler (2006)	Adaptive Matrices Test	Evaluate adaptive success control algorithm	Unknown setting German Individuals aged 18-81 years *n* = 392	Induction	1PL	?		Real	?	Minimum SE SE < .60	MI based on variations of base success probability	?
Hausler and Sommer (2008)	Lexical Knowledge Test	Compare and simulate item selection methods	Simulated examinees *n* = 2000	Crystallised Intelligence	1PL	?	126	Simulation	Mean item difficulty	Max items 20 items	MI based on variations of base success probability	ML
Hines (2018)	N/A	Develop an experimental non-verbal measure of cognitive ability through automatic item generation	Amazon Mechanical Turk *n* = 333	Fluid Reasoning	1PL	?	?	Real	?	?	MI	?
Kantrowitz and Dainis (2014)	?	Examine inconsistent test scores for possible cheating	US candidates for 11 jobs *n* = 4,026	Deductive reasoning	?	?	Propri etary >300	Real	?	Unspecified Minimum SE	?	?
Konsztowicz et al. (2011)	Geriatric Rapid Adaptive Cognitive Estimate	Adaptive administration of the MMSE and MoCA	Patients referred for geriatric cognitive assessment *n* = 137	Cognitive impairment	1PL	?	?	Real	Clock Test	?	?	?

Author	Instrument Name	Study Objectives	Setting and Sample	Latent Trait	IRT	CAT Software	N Items	Simulated/ Real World	Starting Item	Stop rule	Item Selection	Theta Estimation
Legree et al. (1998)	Telephone Test	Correlate the Telephone Test with the Armed Forces Qualification Test	US military enlistees *n* = 144	Crystallised Verbal Aptitude	?	?	?	Real	?	?	?	?
Liao and Ho (2011)	Computerized Figural Testing	Integrate CAT with virtual item banks and automatic item generation to improve test security	Six graders from Taiwan *n* = 310	IQ	1PL	?	?	Real	?	?	?	?
Makransky and Glas (2013)	Adjustable Competence Evaluation	Investigate the use of MCAT in personnel selection	Test-takers as part of personnel selection, recruitment, and individual development *n* = 1350	Numeric, spatial, and verbal ability	2PL	FORTRAN 6.0	201	Simulation	Random	Unspecified max items	Bayesian	EAP
Reise et al. (2011)	SCoRS CGI-CogS	Run a CAT simulation of existing rating scales of cognitive impairment	Patients with schizophrenia and schizoaffective disorder *n* = 176	Working memory, attention, verbal learning, spatial learning, reasoning, processing speed	GRM	Firestar	41	Simulation	$\theta = 0$	SE < .25 SE < .30 SE < .40 SE < .50	MI	EAP

Author	Instrument Name	Study Objectives	Setting and Sample	Latent Trait	IRT	CAT Software	N Items	Simulated/ Real World	Starting Item	Stop rule	Item Selection	Theta Estimation
Sands et al. (1997)	CAT-ASVAB	Descriptive chapter on the technical aspects of the ASVAB	Military personnel	Arithmetic Reasoning Word Knowledge Paragraph Comprehension Mathematics Knowledge General Science Mechanical Comprehension Electronics Information Auto and Shop Coding Speed Numerical Operations	3PL	NA	~200	Real	Select a random item from the five best items	Unspecified minimum SE Unspecified max items Unspecified Time limit	MI and random item selection	Owen Bayesian
Segall (2001)	MIRT-AVSAB	Utilise MIRT to measure hierarchical factor model of intelligence	US military applicants *n* = 12,000	General, Verbal and Math Ability	MIRT	IFACT	420	Simulation	?	Max items 60 items	Self-authored algorithm	MAP
Sommer et al. (2018)	Adaptive Tachistoscopic Traffic Perception Test	Compare administration methods to determine causes of practice effects	Driving tests *n* = 891	Perceptual Speed	1PL	?	14	Real	Mean item difficulty	?	MI	?

Author	Instrument Name	Study Objectives	Setting and Sample	Latent Trait	IRT	CAT Software	N Items	Simulated/ Real World	Starting Item	Stop rule	Item Selection	Theta Estimation
Thomas et al. (2020)	N-Back Task	Proof of concept study of use of latent variable modelling with small samples	Undergraduate research participants and clinical outpatients *n* = 92	Working Memory	GLVM	CogIRT	300	Simulation	?	Reliability of estimated *d*' intercept > .7 Max items 9 items	MI	MH-RM
Witruk (2019)	?	Compare working memory performance of German and Chinese dyslexic children	German and Cantonese speaking dyslexic children *n* = 192	Working Memory	?	?	?	Real	?	?	?	?
Wouters, de koning, et al. (2009)	CAMCOG MMSE	Investigate use of CAT to maintain precision while shortening testing	Patients with cerebrovascular disease or vascular dementia *n* = 797	Orientation, memory, language, attention, praxis, calculation, perception	1PL	?	67	Simulated	Item 138, 139, 146, 171, 147, 178	Minimum SE < .15 Unspecified Max items	MI	?
Žitný et al. (2012)	Test of Intellectual Potential Vienna Matrices Test	Investigate construct validity of CAT compared to paper-and pencil	Slovak secondary school students *n* = 803	General intellect via Fluid Intelligence	3PL	CATO	?	Simulation	Random selection	Minimum SE < .50	MI	EAP

Note. SE = Standard Error, MI = Maximum Information, EAP = Expected a Posteriori, MLE = Maximum Likelihood Expected, MAP = Maximum a Posteriori

There was some variation in whether the source made use of a simulated CAT (*n* = 8) or whether the source reported results of implementation of CATs with 'real' participants (*n* = 12). Those that used real samples often did not demonstrate where they sourced their sample from (Arendasy & Sommer, 2012, 2017; Hausler, 2006). When they did they used engineers (Balas-Timar & Balas, 2009), participants from employment pre-selection (Kantrowitz & Dainis, 2014; Makransky & Glas, 2013), Amazon Mechanical Turk participants (Hines, 2018), military enlistees (Legree et al., 1998; Sands et al., 1997; Segall, 2001), those referred for geriatric cognitive assessment (Konsztowic*z* et al., 2011; Wouters, de koning, et al., 2009), those undertaking a driving ability test (Sommer et al., 2018), or people undergoing psychiatric review (Reise et al., 2011; Thomas et al., 2020). While it is positive see CAT measurement of cognitive ability across a range of settings, there appears to be substantial lack of robust investigation into the use of CAT in real settings across a range of jurisdictions for a variety of referral questions. Increased usage of CATs and subsequent evaluation is required to demonstrate their applicability.

Across both the simulated and real CAT studies, only four studies made use of school aged children. de Beer (2005) simulated a CAT after splitting items into two groups and administering them to 2,454 Grade 9 and 11 students. While the terminology described in this study differs from contemporary CHC theory, it focuses on the learning potential of participants by measuring non-verbal figural reasoning ability as a form of cross-cultural assessment in South Africa. In a similar vein, (Liao & Ho, 2011) utilised automatic item generation combined with CAT to improve figural test item exposure rates; to do this the authors utilised their CAT with 310 Grade 6 students across 10 classes in Taiwan. Their test demonstrated a Pearson correlation .683 with the Advanced Progressive Matrices test (a measure of Inductive Reasoning). Witruk (2019) administered a self-designed adaptive measure of auditory and visual working memory to 86 dyslexic and non-dyslexic children. Finally, Žitný et al. (2012) compared the performance of 803 secondary school students from Slovakia on paper-and-pencil and computer-based versions of the Test of Intellect Potential and the Vienna Matrices Test (both measures of Fluid Intelligence). The authors then simulated a CAT version of the test and showed that it maintained validity while reducing the number of items administered.

All four studies above demonstrate the viability of CAT for school-based screening, but gaps in the literature remain. Firstly, many of the characteristics of the CAT utilised in each study are unclear; the software utilised is only identified in two studies (both of which were simulation), the IRT model was not identified for one study, and only the Žitný et al. (2012) study makes it clear which item selection procedures and θ estimation method was implemented. Finally, the starting and stop rules were only made clear in Žitný et al. (2012) and de Beer (2005).

These issues make it difficult to compare the effectiveness of CATs with school aged participants, as well as make it challenging to engage in replication efforts. Secondly, none of these studies made use of an English as a first language sample, limiting the generalisability of findings. Thirdly, the psychological construct of interest generally focused on Fluid Reasoning or Working Memory, although was rarely operationalised in a way consistent with CHC theory; instead, they appear to rely on test driven definitions rather than theoretical perspectives of intellectual ability.

Across all studies, the software enlisted to host the CATs was generally not mentioned. In some cases where simulations were carried out the software was identified but insufficient additional information was provided by authors for potential examiners to make use of the CAT, or for researchers to replicate the research. In fact, excluding simulation studies, none of the sources identified a CAT platform. For a school, or even another researcher or psychologist, it would be practically impossible to implement these CATs in the real world. These studies demonstrate viability of CATs within different settings and populations for measurement of different constructs but offer little practical opportunity for application.

Very few studies focused on developing items from the perspective of CHC theory, the most contemporary theory of intellectual abilities. Several studies engaged with existing tests of cognitive ability (e.g. Balas-Timar & Balas, 2009) that can be *classified* under the CHC taxonomy (Jacobs & Costello, 2013), while others generated items using automatic item generation (e.g. Arendasy & Sommer, 2017). Similarly, both Arendasy and Sommer (2012) and Arendasy and Sommer (2017) identify that some of their items measure inductive reasoning but tend to focus on item characteristics as indicators of what they are measuring, rather than theoretical definitions. Interestingly Arendasy and Sommer (2012) visualise how their findings fit with a CHC model that represents *g*, G*f*, G*c*, and G*q* but make no link back to the wider literature about CHC abilities, their conceptualisation, or structure. Hausler and Sommer (2008) utilised the Lexical Knowledge Test as a measure of the latent trait Crystallized Intelligence, again demonstrating a lack of modern conceptualisation of cognitive ability. Out of the 20 retained sources, only Arendasy and Sommer (2012) and Hines (2018) refer to CHC theory, with the latter relying heavily on theory to create a CAT reliant on automatic item generation to measure inductive reasoning. Overall, these sources appear to use a mixture of items that likely possess construct irrelevant variances, are predicated on poorly defined traits, or measure knowledge and achievement rather than cognitive abilities.

Further evidence of inconsistently defining intelligence is demonstrated by studies utilising ratings of cognitive abilities rather than making use of performance measures. Reise et al. (2011) conducted a CAT simulation of the combined 41-items from the CGI-CogS (21-items) and SCoRS (20-items), two interviewer report scales of cognitive impairment. Using Firestar (Choi, 2009) the authors simulated CATs with varying minimum standard error stop rules. They found the number of items could be reduced substantially, however in most cases the CGI-CogS items provided more information than the SCoRS items, and across all 41 items it would usually be the same 10-items administered. There is a significant literature base that continues to question the applicability of questionnaires to measure certain constructs, suggesting the relationship between performance-based measures and questionnaire ratings is poor (Coutinho et al., 2017). It may be the CGI-CogS and the SCoRS are useful in the measurement of cognitive impairment or quality of life, but it is unlikely they are good measures, broadly speaking, of intelligence as conceptualised by CHC theory.

Several studies throughout the review discussed the use of the Armed Services Vocational Aptitude Battery (AVSAB), or simulated different CATs using data from the AVSAB and thus only one source was retained. While the AVSAB has been found to possess a general factor (Ree et al., 1994), the test does include components that are often considered irrelevant to the measurement of general intelligence since they are taught skills. For example, the CAT-AVSAB includes 'knowledge tests' and 'ability tests' (Sands et al., 1997); subtests include coding speed, assembling objects, arithmetic reasoning, word knowledge, general science, paragraph comprehension, numerical operations, automobile and shop information, mathematical knowledge, mechanical comprehension, and electronics information (Hartmann, 2006). Outside of the original CAT conceptualisation of the AVSAB, a multidimensional hierarchical CAT version of the AVSAB also only made use of four subtests measuring verbal and math abilities (Segall, 2001). Many of the subtests across the AVSAB and its various versions are argued to be within the Reading-Writing (G*rw*), Quantitative Knowledge (G*q*), and Domain Specific Knowledge (G*kn*) domains. In contrast some researchers have argued that the AVSAB measures no more than Comprehension Knowledge (Gc) (Roberts et al., 2000) and if true, it matters not whether it is a CAT or not. Such achievement or knowledge-based tasks are rarely included as sole measures in cognitive ability tests; one only need look at the ongoing debates about construct irrelevant variance in the Wechsler scales' arithmetic task (Flanagan & Alfonso, 2017) to become concerned about the AVSAB subtests. The AVSAB may have predictive validity for military service but is not applicable otherwise in the measurement of intelligence.

The IRT model used to fit items and generate parameters for use in the CAT differed across sources. The Rasch Model (or 1PL) was clearly identified in most studies (*n* = 9) with the 3PL model the second most common (*n* = 3). GRM (*n* = 1), an unidentified MIRT model (*n* = 1), and GVLM (*n* = 1) were also used. In all other cases either no specific model was identified (*n* = 3) or the study referred to multiple models but did not indicate which model they ultimately chose (*n* = 1). It is not surprising that the 1PL model was the most utilised as it is the most known among researchers, and due to its specific focus on the difficulty parameter it is easy to compare items. It is also not surprising that the 3PL is a runner up. This model is more lenient and allows researchers to include a guessing parameter; this is useful due to the high likelihood in cognitive ability and academic achievement tests to become increasingly difficult and results in participants guessing. A review of models used in the journal Psychometrika and psychometric textbooks showed a similar pattern, with a Rasch/1PL frequency of 32.36% and a 3PL frequency of 22.55% (Kim et al., 2020). Despite the model choice in articles retained for this review, no sources identified why a certain model was applicable to their items or suitable for their CAT. Further research is required to demonstrate the suitability of different IRT models for cognitive ability CATs specifically, but it is likely this should be driven by theoretical factors rather than researchers' familiarity with specific models or using whichever model 'fits best'.

In relation to starting item selection, sources were again varied. Most sources did not identify the method of first item selection (*n* = 10) but when they did the most utilised was the mean item difficulty (*n* = 4) or item closest to a θ of 0 (*n* = 1). This essentially means the CATs chose an item of moderate difficulty that was likely to have a 50% probability of success for the most amount of people the test was designed for. Other start item selection methods included random (*n* = 2), random from a specific set of ideal items (*n* = 1) or the start item was already specified (*n* = 2). Wouters and colleagues (2009) identified a set of items to administer initially because subsequent items rely on recall of other items. Konsztowicz et al. (2011) used a specific item to inform the selection of subsequent items. It is not unexpected that mean item difficulty and random item selection were utilised in a range of studies as CATs are known to be able to quickly recover from inappropriate starting items.

Similarly, the stop rules for the CATs in each source varied. The stop rule was not identified for many sources retained (*n* = 7), but a maximum number of items to be administered (*n* = 5) or a minimum standard error (*n* = 5) were the most frequent. There was also a mixture of stop rules (*n* = 3). de Beer (2005) and Wouters and colleagues (2009) utilised both a minimum SE or maximum number of items, while Sands et al. (1997) utilised a mixture of time limit, minimum SE, and a maximum number of items. Unfortunately, the only retained study that considered the difference in efficiency across different levels of a certain stop rule was Reise et al. (2011) who employed a range of different minimum standard error levels. Stop rules should be driven by the goal of the CAT, such as whether it aims to classify participants, accurately measure a participant's position on a latent trait, or to track progress over time (Weiss, 2011). Despite this, there is opportunity supported by this literature review to evaluate the impacts of different stop rules on real computer adaptive tests, rather than in simulations.

It was somewhat difficult to identify the item selection method for many of the studies retained. In some cases, this appeared to be because customised algorithms were used. However, MI (*n* = 8) and customised versions of MI (*n* = 3) algorithms appeared to be the most common. In other cases, the retained studies either did not refer to an identifiable item selection method (*n* = 7), a vague reference to Bayesian item selection (*n* = 1) or reported a 'self-authored' algorithm (*n* = 1). Again, despite a wide range of methodological studies (not included in this review) on item selection methods, there appears to be little variation (or consideration) of item selection methods in the utilisation of real CATs.

The method used for θ estimation was the least identified characteristic in the reviewed studies (*n* = 12). Where they were identified, the Maximum a Posteriori (MAP; *n* = 2) and EAP (*n* = 3) were the most frequently used. Other mentioned estimation methods included ML estimation (*n* = 1), Owen Bayesian estimation (*n* = 1) and

Metropolis-Hastings Robbins-Monro (*n* = 1). Some studies suggest that there is little difference in outcomes between MAP, MLE and EAP and the choice between the three differs based on unique characteristics of the test and items (Chen et al., 1998; Chen et al., 1997).

Some articles failed to include enough details to fully understand the appropriateness or performance of the CAT, and the way in which it was designed. For example Balas-Timar and Balas (2009) employed a CAT version of the Multidimensional Aptitude Battery – II (MAB-II) with a focus on 'Fuzzy Logic' to improve ability estimation; the authors provided a descriptive discussion of the decisions a CAT designer should make relating to item pools, starting points, item selection, score estimation and CAT termination, however they did not indicate what choices they made for any of these for their CAT version of the MAB-II. Searches of reference lists, Google and databases failed to produce any results about an accessible MAB-II CAT, making it unlikely that schools or other organisations could implement such a platform in their practices. For the MAB-II, a relatively common tool, making these characteristics known could enhance the future of CAT research, the use of a MAB-II CAT, or enable replication of this research. If studies continue to omit details of their CATs, an organisation will have to develop their own CAT to implement or rely on a proprietary solution that is unlikely to have independent peer-reviewed evaluation. Existing mainstream paper-based cognitive ability tests used by psychologists in Australia benefit from a wide range of research and evaluation, and without the same level of investigation CATs will fail to become mainstream tools for psychologists and educators despite their many advantages.

Despite frequent claims in the literature that IRT and CAT are widely used in cognitive ability testing, the current review suggests otherwise. When focusing on measures of cognitive ability, some CAT tools do exist in the literature, but they generally focus on the use of intellectual ability as part of a larger study rather than being the focus. Further, the gaps in the literature suggests the use of CATs for cognitive ability measurement is particularly limited in Australia. This is unsurprising given that "only 4% of IRT abstracts mentioned terms related to computerized adaptive testing" (Thomas, 2019, p. 1447). It appears CATs have largely been adopted by organisations in a proprietary manner for personnel selection in military and business environments. Ultimately there appears to be no comprehensive CAT exists that can be utilised in schools or by psychologists in their day-to-day practice that relies on CHC theory.

2.17 Conceptualising a CAT Screening Tool

In Australia, cognitive ability is measured by organisations, psychologists, and teachers to assist with understanding job performance, academic concerns, response to trauma, and people's development. Australia has begun to see shifts in academic measurement via CATs (ACARA, 2016). However, excluding organisational or research specific tools, there does not appear to be a tool that measures cognitive ability from the perspective of CHC theory while also using CAT technology. Such a gap poses interesting areas for research that could have significant implications for psychologists and teachers in Australia.

The following studies detail the development and evaluation of a CAT using items developed in recent research. Without the ability to rely on decades of copyrighted item development, or significant organisational or publisher capital, the development of CAT aims to be a screening tool for identification of children or adults with deficits in overall intellect or specific cognitive abilities. Such a tool is more useful for assistance in referring students and adults with potential traumatic brain injuries, domain specific deficits, or generalised learning difficulties for further comprehensive evaluation. As an entirely new set of items is being developed, it is believed that further research following this thesis would be required before the CAT could be relied upon for high stakes decision making or be used in measuring specific ability strengths and weaknesses.

Four stages of research were carried out. For the first stage an Item Tryout Study (ITOS) was conducted. This investigated the operationalisation of CHC abilities known to be important to measurement of overall cognitive ability as well as related to academic outcomes, with associated item development for initial administration to participants. As a completely new tool there was no reliance on previous psychometrics of older items and as such very conservative psychometric analyses were carried out. In order to access many participants, large samples of adults were recruited via social media; this had disadvantages in terms of participant drop out and possible motivation, guessing, and distraction issues, however it allowed a large number of participants to be quickly exposed to significant numbers of items. As IRT is largely probability based, it was believed that this would be suitable for an ITOS. The second stage, Item Calibration Study (ICS), involved taking the outcomes of the initial ITOS to develop new items, adjust previous items, and attempt administration of items to both adults (via social media) and school aged participants (via supervised visits to their school) in a planned manner to achieve more robust psychometric outcomes, with a focus on IRT analysis. The third stage, CAT Simulation, involved using the Rasch item parameters from the ICS and simulating the CAT with 5,000 simulated participants with a wide range of abilities; a focus was placed on performance of the CAT when allowing for different levels of SEM. The fourth stage, Validity, took the school aged participants from the ICS, who had also been administered the WISC-V, and analysed the convergent validity between tools.

With these stages taken together it was believed that this thesis would form an innovative pilot study into the feasibility of a CHC-CAT screening tool using CHC abilities known to be important to learning.

The goal of this project was to design a CHC-CAT screening tool that was transparent in its design, that the methodology of design was pre-determined, and post-hoc analysis and decision making was avoided where possible. As an exploratory research project, the following research aims were formulated:

1. Evaluate and calibrate a set of items using CTT and IRT methods to identify four sets of items that measure cognitive ability in line with the CHC theory of abilities.
2. Simulate a CAT using known Rasch item parameters to explore the efficiency of such a tool.
3. Compare and contrast the outcomes of the WISC-V with scores from item sets derived from the current project.

By addressing these research aims it was hoped that we could establish a solid foundation for future research into an open-source CAT that measures cognitive ability based on CHC theory; such a tool can continue to be iteratively developed, improved and ultimately implemented in a variety of contexts.

Chapter 3: Item Tryout Study

3.1 Introduction

The first stage of the research to develop a ‘CHC-CAT’ screening tool was to build a set of items and evaluate these in an Item Tryout Study (ITOS). While such a procedure is common in the development of psychological inventories and educational tests, there appears to be little in the way of a formalised procedure other than general guidelines, or what previous studies have conducted. Additionally, commercial publishers tend to not make their item design phase publicly available for scrutiny, making it difficult to design items in a manner that follows procedures of mainstream cognitive ability tools.

The overall goal of the ITOS was to build item pools for new tests from which an item calibration study could be conducted; this involved trialling items and identifying potential problems and strengths with item design and psychometrics based on participant performance. Because we did not have an original set of items like other tools such as the *Woodcock-Johnson* (WJ) or *Wechsler Intelligence Scale for Children* (WISC) batteries (due to their multiple editions), we were unable to develop a set of “carefully selected linking items” (McGrew et al., 2014, p. 48). Therefore, the ITOS in this thesis evaluated the psychometric characteristics of a set of items recently developed as a pre-cursor to this project; items were designed to measure Lexical Knowledge (G*c*:VL), Induction (G*f*:I), Visualisation (G*v*:Vz) and Working Memory (G*wm*:Wc) as these are core abilities known to be important to learning (Flanagan et al., 2012). Because the goal of this thesis was to focus on technical application of CAT rather than attempting inform CHC theory, there is a focus on the theoretical underpinnings of the items developed early in this chapter, but not necessarily a continued focus on how the psychometrics support CHC theory. This is partially accounted for when looking at convergent validity with the WISC in a later chapter; if the current tool correlates with the WISC, a tool that aligns with CHC theory, then it can be assumed that the current items align similarly well with CHC theory. It is anticipated that after the ITOS, items will be calibrated and additional items developed, and a more comprehensive study of the items will be conducted.

3.2 Chapter Aims

This chapter consists of a secondary analysis of existing data gathered during four Master of Psychology projects. These projects were supervised by me and Dr Shane Costello. As part of these projects each Master of Psychology candidate (John Maguire, Elizabeth Kennedy, Adalyn Heng, and Sarah Fleming) wrote a literature review about their construct of interest, and a subsequent research report about the outcomes of the data analysis. After the four Master of Psychology candidates finalised their theses, additional data collection was conducted by me, and thus analyses detailed below will differ from their original outcomes.

Each project individually focused on one of four CHC narrow abilities; Lexical Knowledge, Induction, Visualisation, and Working Memory. These abilities were chosen as they are an integral aspect of many mainstream cognitive ability tests (i.e. WISC-V, WAIS-IV, WJ IV COG) and significant bodies of research suggest they are either significantly correlated with academic achievement, or are considered intermediary abilities within the most recently published CHC taxonomy (Schneider & McGrew, 2018). Items were therefore developed in line with these tools and prior theory.

Items for Lexical Knowledge were conceptualised and sourced by John Maguire; audio files for these items were sourced by the current PhD thesis author via a third-party contractor. Items for Visualisation were conceptualised and developed by Adalyn Heng and a third-party contractor completed the graphic design for these items. Items for Induction were conceptualised and developed by Sarah Fleming. Items for Working Memory were conceptualised and developed by Elizabeth Kennedy, with the graphic design completed by me. A significant focus of the Master of Psychology theses was the theory underpinning item development and thus only brief details of the item development are within this PhD thesis; the current thesis focuses on the application of CHC theory, Item Response Theory (IRT) and CAT, rather than a discourse on the nature of each CHC narrow ability.

As well as contributing significantly to the original item development, I also present new ideas, new ideas, analysis, and methodology in the current chapter. I wrote the R code that completed the analysis (github.com/jakekraska/phd), the supplemental HTML, CSS and JavaScript code required for the testing platform to work with our items (Concerto, discussed below), authored the ethics application and assisted with item development for all four constructs. For the current chapter the data analysis was standardised across all four constructs to ensure consistency of approach; this was an effort to align the analysis of existing data in this chapter with the philosophy of Open Science discussed earlier in this thesis (Section 1.3). That is, a standard methodology for analysis of all item sets was developed and statistical cutoffs were consistently applied to avoid "hypothesising after the results are known" (Kerr, 1998, p. 197).

3.3 Method

3.3.1 Participants

The entire sample collected consisted of 2,776 participants. These participants were recruited through paid Facebook advertisements, distribution of a link (chctest.com.au) on social media websites, a Facebook page (facebook.com/chctest) and via snowball recruitment. Participation was voluntary and no inducement was provided. Ethics approval was gained from the Monash University Human Research Ethics Committee (Project 13912).

Data was cleaned, organised, and recoded using the *dplyr* (Wickham et al., 2019), *tidyr* (Wickham & Henry, 2019), *stringr* (Wickham, 2019), *tibble* (Müller & Wickham, 2019) and *knitr* (Xie, 2015) packages, via *R v.3.6.0* (R Core Team, 2019) within the *R Studio Integrated Development Environment v.1.1.456* (R Studio Team, 2015).

Data with missing demographics was removed from the data set. Only 47 (1.69%) participants identified as a gender other than male or female and were removed to ensure that differential item functioning analysis could be conducted with confidence. As this stage of the research consisted only of adults, removal of those under the age of 18 and over the age of 90, reduced the participant count by 38. After this data cleaning, 2,691 participants remained. There were 1,428 females (53.1%) and 1,263 males (46.9%).

Items that the participant did not complete within the item time limit were scored as 0. Items that the participant did not complete were also marked as incorrect which is in line with non-adaptive paper-based tests when individuals reach the ceiling. Given that the difficulty of each item is uncertain, the desired linear order of the items is unknown and as such it is acknowledged this was a strong assumption to make.

Each of the 2,691 participants were able to select which item sets they would like to complete, resulting in varying numbers for each item set. Because the data was collected largely through social media and administered online without supervision (i.e. unproctored), three steps were taken with each item set to improve the robustness of the data. Firstly, participants aged below the minimum (Q1 – 1.5 * IQR) and above the maximum (Q3 + 1.5 * IQR) were removed. The mean age was 47.44 (*SD* = 16.47); participants were split across 7 age groups (Figure 3-1). Using a boxplot (Figure 3-2) no outliers were identified. Secondly, participants that were below the 5th percentile or above the 95th percentile in terms of time taken for the item set were removed. Thirdly, participants that obtained all items correct or all items incorrect were removed to assist with Mokken scaling. While these steps are quite conservative, and produced a data set with less variability, it enabled full-case analysis. A summary of the steps, and a summary of demographics, is in Table 3-1.

3.3.2 Materials

3.3.2.1 Test Platform. The platform used to host the ITOS was Concerto v.5.0.beta.7.2 (Lis, 2018). Concerto is an open-source online adaptive testing platform first described in the literature by Scalise and Allen (2015). This was used because it has CAT capabilities. By using this platform, it was hoped that there will be a reduction of aesthetic changes and functionality across future research. Additionally, this platform allowed for significant customisation of item format and web page presentation utilising HTML, CSS, and JavaScript.

The Concerto Platform allows a test designer to use nodes to create a test flow of differing components (Figure 3-3). For the current study, the test flow was linear in nature; participants were required to review the explanatory statement, enter basic demographic information, select the subtest(s) they wished to participate in, then items were presented in a sequential order. Participants were given feedback about their performance on each of the constructs at the end of each subtest. Using a mouse to select response options reduced the impact of writing or verbalisation difficulties.

Figure 3-1. Participant Age by Group

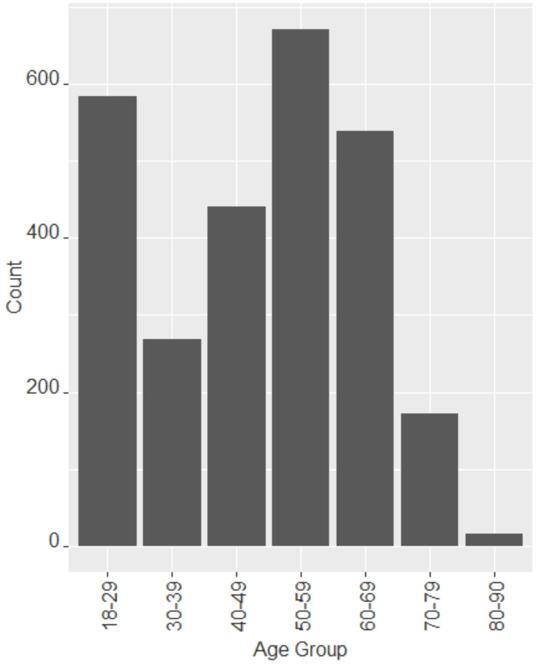


Figure 3-2. Age Outliers

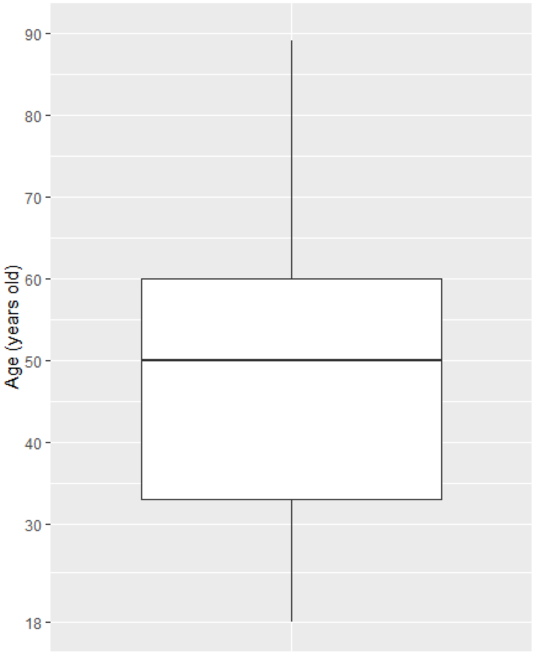
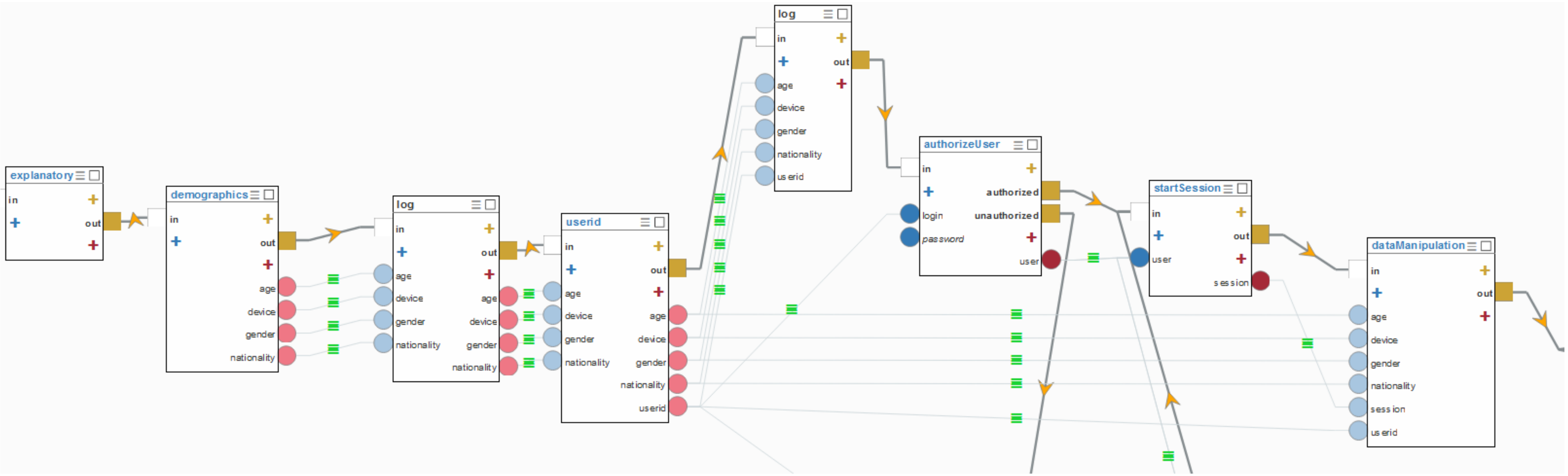


Table 3-1. Data Cleaning Stages and Demographics

	Gc:VL	***Gf*:I**	**Gv:Vz**	**Gwm:Wc**
Initial *n*	1241	673	889	529
Data Cleaning				
Gender/MD[1]	21	12	11	8
Age Outliers	0	0	0	0
Time Outliers	122	66	88	52
Score Outliers	20	1	4	9
Final *n*	1078	594	786	460
Gender[2]				
Male	497 (46.1%)	250 (42.1%)	353 (44.9%)	181 (39.3%)
Female	581 (53.9%)	344 (57.9%)	433 (55.1%)	279 (60.7%)
Mean Age[2] (SD)	47.27 (15.84)	46.69 (15.30)	45.78 (16.01)	46.21 (15.41)

Note. [1]MD = Missing Data; [2]After removal of outliers

Figure 3-3. Example of Concerto Nodes



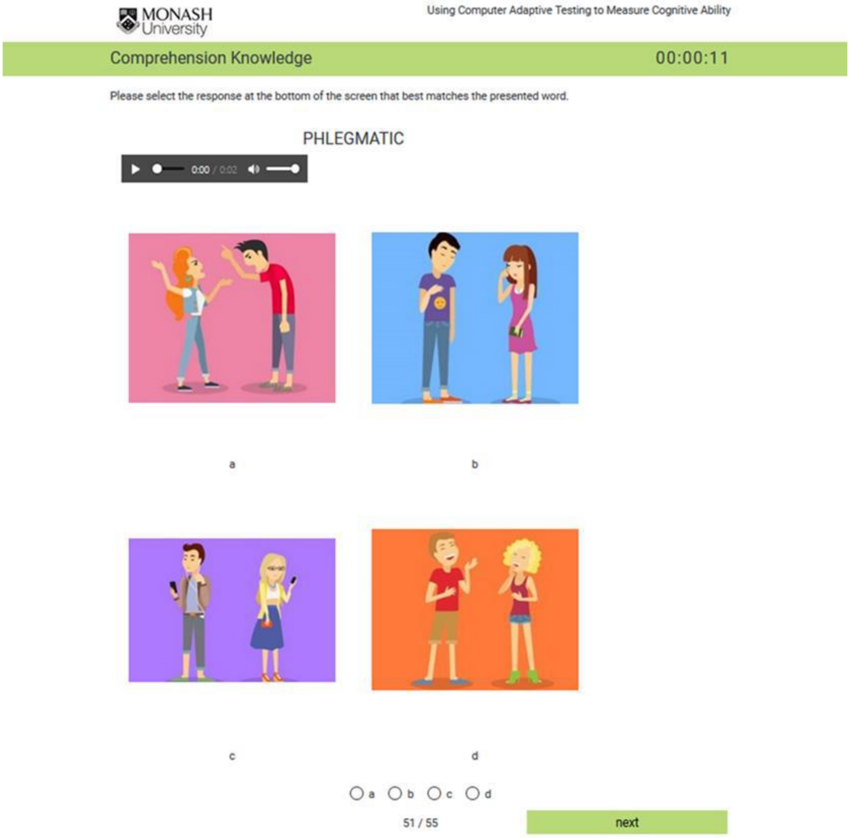
Note. This demonstrates the separate nodes of Concerto. Each node represents a step that a participant flows through, sometimes associated with test materials, and at other times associated with backend calculations or storage of data in the Concerto database.

3.3.2.2 Item Development and Validity. Overall, the goal of test validation is to ensure the items are valid and measure what they are intended to measure. Downing and Haladyna (1997) argue there has been too much emphasis on the statistical evidence at a test level, and that an “ideal process for item validation” begins with a “careful and systematic approach to the task of creating test items” (p. 63). Downing and Haladyna highlight the need to focus on both qualitative and quantitative evidence. Much of the qualitative evidence that items within this study are valid are within the early research that established items prior to this PhD. The quantitative evidence will be established in the results and discussion of this chapter and following chapters.

The goal of the initial item development was to design items that could categorise participants into those that possess below average ability and those that possess average ability. Little focus was placed on the design of items that were significantly difficult. As discussed in the prior chapter, IRT assists in the development of invariant items that have quantifiable difference in their level of difficulty. It is possible to design items across the spectrum of difficulty. While ethical challenges (discussed further below) resulted in initial items being administered only to adults, initial items were designed with the developmental trajectory of children in mind.

3.3.2.3 Lexical Knowledge. The initial set of Lexical Knowledge (G*c:*VL) items were conceptualised in Maguire (2018). To develop these items mainstream tests of Lexical Knowledge were analysed, including the *Wechsler* scales and WJ batteries. Items were developed to reflect increasingly complex vocabulary words. Each item had a stimulus word and corresponding audio presentation of a female voice saying that word. Each item consisted of four response options for the participant to choose from. Qualitative evaluation of these items by three co-researchers and two research supervisors indicated these items aligned well with the CHC theory of intelligence; the items each tested participants' "knowledge of the meaning of words and their underlying concepts acquired through reading and listening" (p. 13). All items had a 30 second time limit to reduce opportunities to cheat as well as to reduce the overall test time. CSS and HTML coding was utilised to restrict the maximum width of images to 300 pixels while maintain the aspect ratio of the image (Figure 3-4).

Figure 3-4. Example of Gc:VL item from Maguire (2018)



MONASH University

Using Computer Adaptive Testing to Measure Cognitive Ability

Comprehension Knowledge 00:00:11

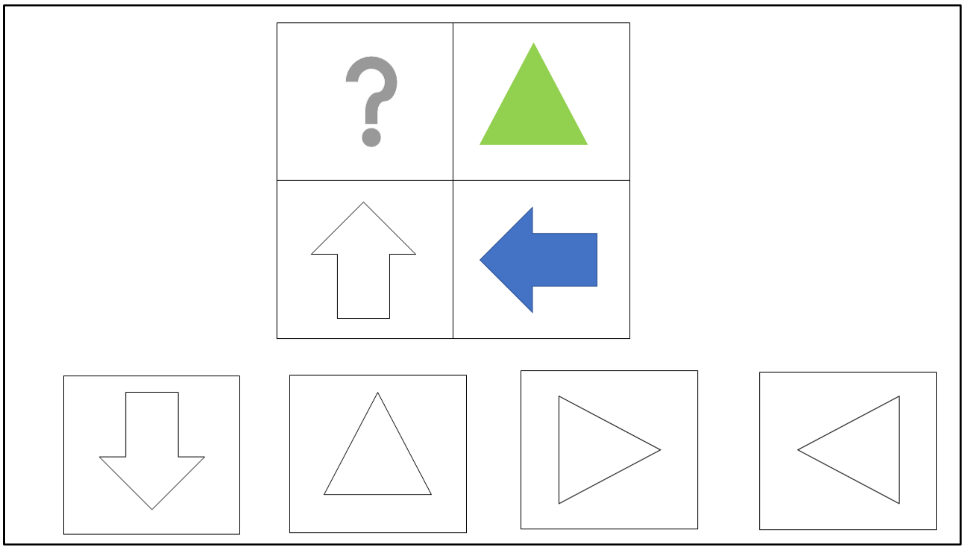
Please select the response at the bottom of the screen that best matches the presented word.

PHLEGMATIC

Maguire (2018) analysed modern test batteries (e.g., WISC-V, WJ IV, PPVT) to develop items that reflected difficulty and content of other subtests that measure Lexical Knowledge. Due to the desire to develop an automated test, it was important to develop items that had minimal language demands. Stimuli for visual representation of words was sourced from 'creative commons' or 'available for commercial reuse' images on the Internet, and the same female voice was recorded for each word. The combination of the auditory and visual stimuli avoids the impact of reading/language difficulties and differentiates the task from subtests such as the WISC-V Vocabulary, WISC-V Similarities and WJ IV Oral Vocabulary which require a face-to-face examiner.

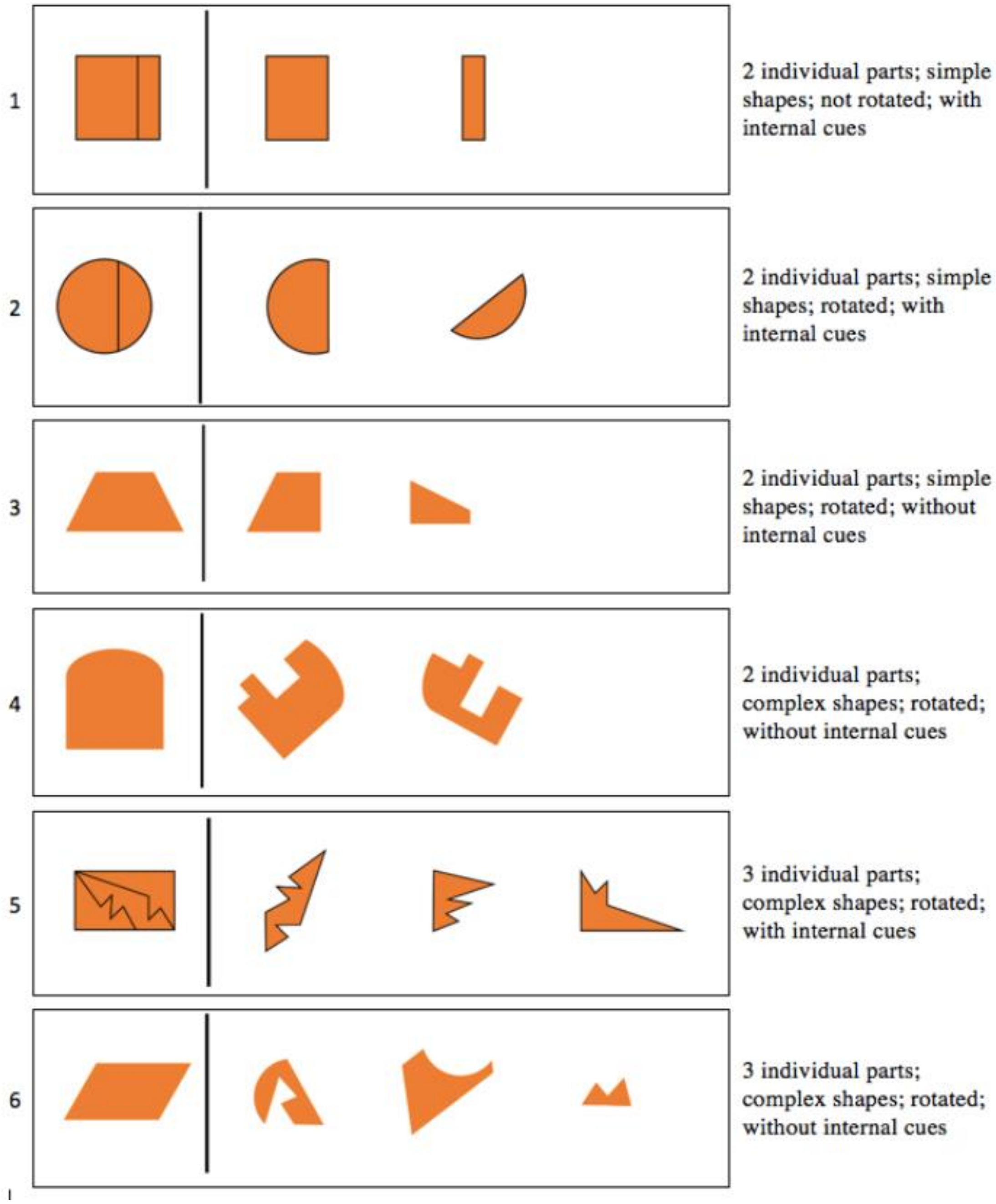
3.3.2.4 Induction. The initial set of Induction (G*f*:I) items were conceptualised in Fleming (2018). Like other types of matrices tests the items were built on the underlying principle of Induction in cognitive ability tests (i.e., the participant detecting the rule for a problem and then applying that rule to determine the missing component). The implementation of this resulted in 33 items that had varying matrix sizes (5 x 1, 2 x 2, 4 x 1, 3 x 3 and 4 x 2). Each matrix had four to six response options (e.g., Figure 3-5). Items 1-15 had a 30 second time limit and items 16-33 had a 60 second time limit.

Figure 3-5. Example of Gf:I item from Fleming (2018)



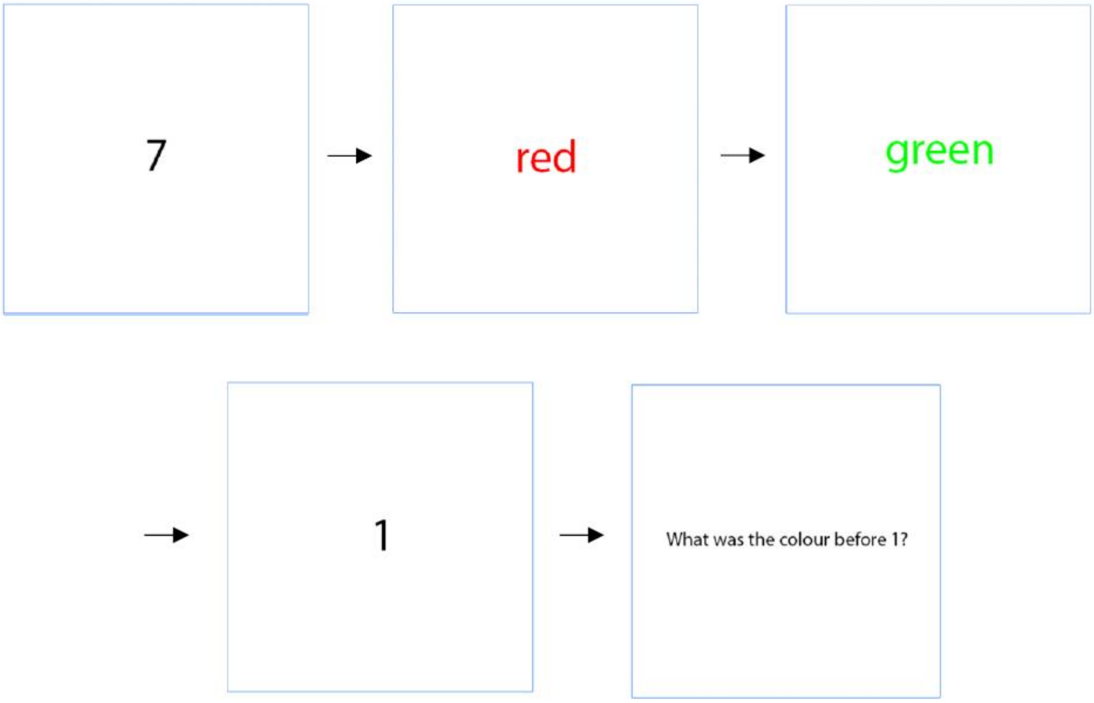
3.3.2.5 Visualisation. The initial set of Visualisation (G*v*:Vz) items were conceptualised in Heng (2018). In conceptualising items, a wide variety of previously developed measures were consulted, including the *Revised Minnesota Paper Form Board Test*, *Mental Rotation Test*, *Paper Folding test*, *Wechsler* scales, WJ batteries, and the *Differential Aptitude Test: Space Relations*. The resulting items consisted of a completed shape and puzzle pieces for which the participant needed to determine if the completed shape could be constructed from the pieces; they then selected a ‘same’ button and if they thought that it could not then they selected a ‘different’ button. To increase difficulty four variables were manipulated: (1) number of pieces, (2) rotation of puzzle pieces, (3) complexity of shapes and (4) internal cues (Figure 3-6). Heng developed 52 items of varying difficulty with a time limit of 30 seconds per item.

Figure 3-6. Difficulty variation of Gv:Vz items in Heng (2018)


1
2 individual parts; simple shapes; not rotated; with internal cues
2
2 individual parts; simple shapes; rotated; with internal cues
3
2 individual parts; simple shapes; rotated; without internal cues
4
2 individual parts; complex shapes; rotated; without internal cues
5
3 individual parts; complex shapes; rotated; with internal cues
6
3 individual parts; complex shapes; rotated; without internal cues

3.3.2.6 Working Memory. The initial set of Working Memory (G*wm*:Wc) items were conceptualised in (Kennedy, 2018). Kennedy stated that the items were "heavily influenced by Engle and Kane's (2004) dual-component model of working memory, which emphasises that working memory involves the ability to conduct a cue dependent search for recently activated information" (p. 16). While Kennedy's item difficulty was based on the *Verbal Attention* subtest in the Woodcock-Johnson IV Tests of Cognitive Abilities which uses a former model of CHC (i.e. Schneider & McGrew, 2012) that placed less emphasis on attentional control (as discussed in Section 2.5), this item format aligns well with the newer conceptualisations of the narrow abilities within Working Memory (i.e. Schneider & McGrew, 2018). That is, these items require both Attentional Control (G*wm:*AC) and Visual-Spatial Short Term Storage (G*wm:*Wv) in order to measure Working Memory (G*wm:*Wc).

Figure 3-7. Example of Gwm:Wc item from Kennedy (2018)



Thirty-eight items were developed utilising a sequence of random numbers and colours in a GIF (Graphic Interchange Format; a type of lossless image with multiple frames). Each GIF was made up of frames that included each number or colour. An example is in Figure 3-7. Each individual frame was presented for 1 second, with the question frame showing until the time limit of the item was reached (40 seconds). Due to limitations in the version of Concerto, answer responses were available as text and visible during the presentation of the GIF.

3.3.3 Procedure

Participants were able to access the Concerto platform via a link on any electronic device of their choosing, at a time of their convenience. The Concerto platform was hosted via an Amazon Web Service Elastic Compute Cloud t2.micro server; the server utilised Ubuntu and Concerto was installed via the recommended instructions provided by the Concerto developers. All recruitment was complete online as per the Participants section.

3.3.4 Data Analysis

The data analysis for this study can be found on GitHub (github.com/jakekraska/phd). Approximately 1,300 lines of R code were written.

3.3.4.1 Software. The analysis in this study was conducted using *R v.3.6.0* (R Core Team, 2019) within the *R Studio Integrated Development Environment v.1.1.456* (R Studio Team, 2015). Packages used for the analyses are specified below.

3.3.4.2 Reliability. The reliability of each CHC item set was measured using the Cronbach alpha statistic (Cronbach, 1951) for each set of items. Cronbach's alpha is a measure of internal consistency, which is argued to represent how well all the items are measuring the same underlying construct. As discussed in earlier sections of this thesis this statistic is largely used in Classical Test Theory (CTT); while it is unlikely subsequent chapters in this thesis will make a strong attempt to improve the Cronbach's alpha (as items at the extreme ends of difficulty are likely to reduce Cronbach's alpha but are required for effective CATs), it is a useful diagnostic tool for this stage of the research. There is a significant literature base on the interpretation of Cronbach's alpha, and this study classified .7 and above as good (Pallant, 2011; Tavakol & Dennick, 2011). Given the goal is to develop each item set to measure a wide range of ability, a very high Cronbach's alpha would suggest the items are too closely correlated and thus too similar. In contrast, a very low Cronbach's alpha may suggest problems with unidimensionality.

Reliability analysis was completed using the *psych package, v.1.8.12* (Ravelle, 2018) for R.

3.3.4.3 Confirmatory Factor Analysis. Confirmatory factor analysis (CFA) was completed to ensure that each item set was unidimensional in preparation for IRT modelling. Five fit indices were utilised: the chi-square statistic (Bollen, 1989), the Comparative Fit Index (CFI; Bentler, 1990), the Tucker Lewis Index (TLI;Tucker & Lewis, 1973), the Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1992) and the Standardised Root Mean Square Residual (SRMSR; Hu & Bentler, 1999). This is a balanced selection of fit indices (Beaujean, 2014) that rely on comparison with the null model (incremental indexes), fit indexes that take into consideration model complexity (parsimony indexes) and models that measure the absolute fit of the model (absolute indexes). As all fit indices have a range of strengths and weaknesses, no singular index was relied upon in isolation.

The chi-square statistic has been found to be biased for large samples, so several cutoff criteria is relied upon for other fit indices. The CFI and the TLI measured whether the designated model fit the data better than a restricted model, with greater than .9 considered an acceptable fit (Hu & Bentler, 1999). Some authors argue the RMSEA is more appropriate than the CFI for confirmatory analyses (Rigdon, 1996). The RMSEA statistic determines how closely the model replicates covariances, with values below .01 considered excellent, .05 considered good and .08 considered mediocre (MacCallum et al., 1996). The SRMR represents the average discrepancy between an implied correlation matrix and the observed correlation matrix; a value less than .08 is considered good (Hooper et al., 2008).

The extant literature recommends a sample of anywhere between 250 to 500 participants (Lewis, 2017) to provide adequate power to a CFA. Each item set in the current study meets this requirement, with working memory displaying the lowest participant count at 529.

CFA was carried out using the lavaan v0.6-3 package (Rosseel, 2012) for R, relying on the diagonally weighted least squares (DWLS) estimator.

3.3.4.4 Mokken Analysis. Mokken analysis was carried out to determine if the items fit a Mokken scale, which assumes unidimensionality and an increasing level of the underlying trait (known as monotonicity). Mokken scale analysis was originally designed for dichotomous items (Mokken, 1971) and was extended into polytomous items by Molenaar (1991). Unlike a Guttman scale (Stouffer et al., 1950), which assumes that if a respondent answer correctly/positively towards one item, that they *will* have answered a lower difficulty/level item positively, a Mokken scale assumes that a respondent is *more likely* to answer in this pattern. The analysis produces a Loevinger's

H value (Loevinger, 1948). Items that receive a Loevinger *H* value of below 0.3 lack scalability, between 0.3 and 0.4 indicates weak scalability, between 0.4 and 0.5 to have moderate scalability, and values over 0.5 suggest good scalability (Ligtvoet et al., 2010; Sijtsma & Meijer, 1992). When item shave good scalability this means that items can be predictability ordered.

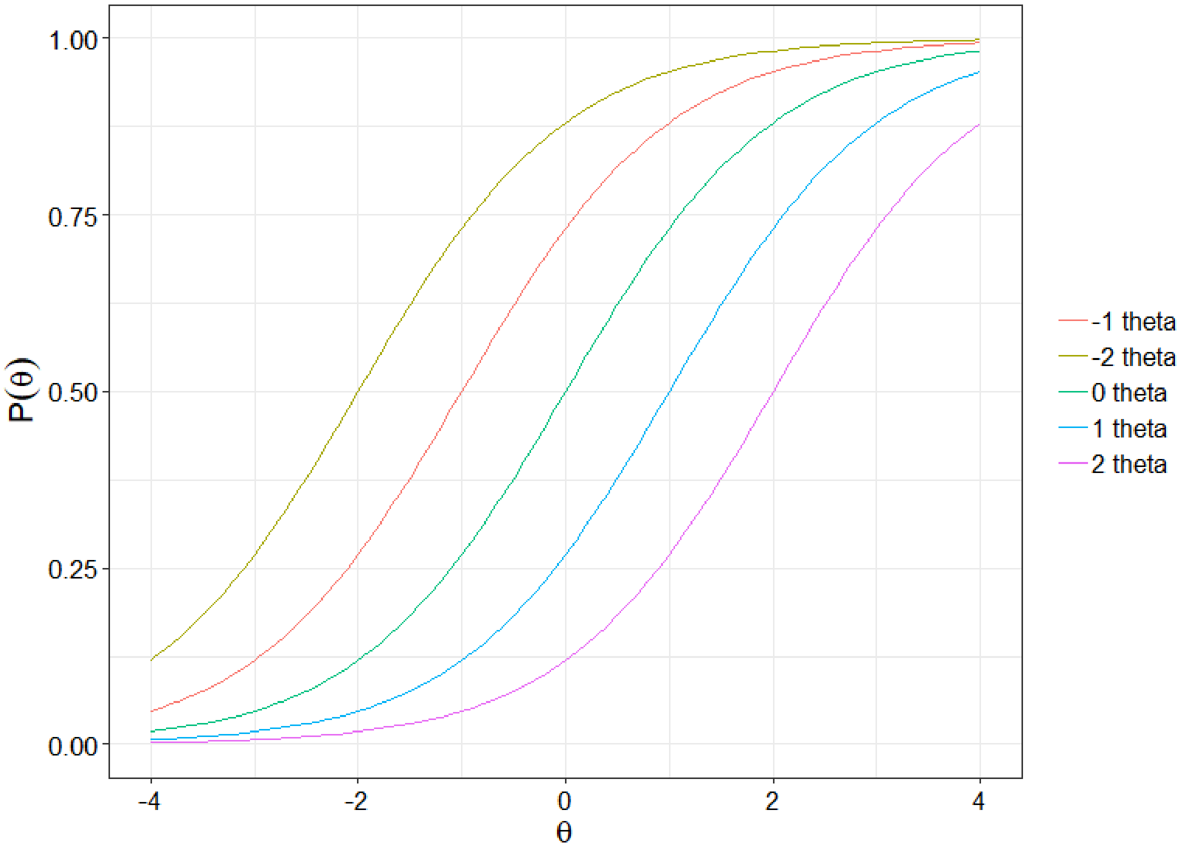
Mokken analysis was carried out using the *mokken v.2.8.11* package (Van der Ark, 2007, 2012) for R.

Straat et al. (2014) suggests that for scales with high quality items, Mokken scale analysis can be carried out if $N > 250$. Automatic item selection procedures (AISP) also perform well with small samples. It is proposed that as each item set is grounded in solid CHC theory, it is unlikely the Mokken scale analyses in this study misclassified items.

3.3.4.5 Rasch analysis. The data was analysed using the Rasch model (Rasch, 1960). The Rasch model provides difficulty parameters for *n* response categories (b_n; see Figure 3-8). That is, each level of theta (θ) is associated with a different probability of a person getting that question correct. Other parameters (i.e., guessing, discrimination) are held stable within the Rasch model. The diagnostics used within the Rasch modelling considered test-level and item-level measures, with fit statistics somewhat like that utilised in structural equation modelling. The model fit of individual items was evaluated with the M2 statistic (Maydeu-Olivares & Joe, 2006), the CFI (Bentler, 1990), the TLI (Tucker & Lewis, 1973), the RMSEA index (Browne & Cudeck, 1992) and the SRMR index (Hu & Bentler, 1999). Items that displayed a *p* value below .01 were removed from the item sets as this indicated they did not sufficiently fit the Rasch model.

Marginal reliability was also evaluated for each item set. This is a reliability estimate of the item set based on the standard error of measurement (SEM) of respondents given variable test lengths (Sireci et al., 1991). As all participants in this stage of the research completed a linear test comprising of all items, it is expected the marginal reliability would be like Cronbach's alpha.

Figure 3-8. Sample ICC with items of increasing trait level or difficulty



3.3.4.6 Local independence. Yen's Q3 method of correlated residuals (Yen, 1993) was used to test the local independence of items. While item residual correlations above .20 are usually considered indicative of local dependence between items, some authors suggest that no singular critical is appropriate for all situations (Christensen et al., 2017). Utilising the outcomes of these simulation studies, a more conservative cutoff of .1 was set. The information curves of items with local dependency were analysed, and items with lower information removed. Local independence was analysed using the *stats version 3.5.1* package for R (R Core Team, 2018).

3.3.4.7 Differential item functioning. Invariance in the item parameters across different sample characteristics was assessed using differential item functioning (DIF). DIF occurs when the probability of endorsing an item or getting an item correct is different for people with different demographic characteristics (Holland & Wainer, 1993; Thissen et al., 1993). Such demographic characteristics can include gender, age, occupation, income, and education. DIF analysis was carried out in the current sample using gender, occupation, and education.

As there were too few endorsements of each response category across all demographics, the DIF analysis utilised collapsed demographic categories. Employment was broken down into 'Not Employed' and 'Employed' while education was broken down into 'Non-Tertiary Education', 'Tertiary Education' and 'Postgraduate Education'. Gender remained as 'Female' and 'Male'.

The DIF analysis was conducted using *the lordif v.0.3-3* package for R (Choi et al., 2016).

3.3.4.8 Item analysis and removal sequence. For each construct, items were analysed and removed iteratively. First, reliability analysis was conducted. CFA and Mokken analyses were conducted together and items that appeared to violate the assumption of unidimensionality or monotonicity were removed. After this, Rasch modelling and local independence analyses were conducted together. Items that were found to be locally dependent or that did not fit the Rasch model were removed. This process was run a maximum of four times to avoid unnecessary levels of analysis for an ITOS. It is believed this was a conservative approach and provided an appropriate balance of CTT and IRT considerations. This provided evidence that the items fit the proposed model and measured varying levels of difficulty for each trait. Once a stable item set was identified, marginal reliability analysis and DIF were conducted.

3.4 Lexical Knowledge Results

3.4.1 Raw Score Outcomes

A high proportion of items were answered correctly by over 90% of the participants, with only five items answered correctly by less than 70% of participants (Figure 3-9). The mean time taken to complete the test was 460 seconds (*SD* = 120 seconds), and generally those that took over 300 seconds scored between 50 and 55 items correctly (Figure 3-10). The mean total score was 48.27 (*SD* = 9.19). Together these findings suggest the items were generally easy for participants, with only a couple of items considered moderate to high difficulty. As all items were presented in the same format, difficulty was purely a function of the item stimuli and the response options.

There was no significant difference between females and males on total score, *t*(991) = 0.75, *p* = .45, 95% CI [-0.69, 1.54], or time taken, *t*(1015) = 1.28, *p* = .20, 95% CI [-5.00, 23.91]. This is illustrated in Figure 3-11 and Figure 3-12. These results also suggest there were similar dropout rates among males and females.

3.4.2 Reliability

Cronbach's alpha for the Lexical Knowledge items was .97. This suggests the items have good internal consistency.

3.4.3 Item Analysis

Fit indices for the CFA, Loevinger's *H* values, and Rasch modelling fit indices are in Table 3-2. Item level statistics are in Appendix A and Appendix B. The item analysis began with the full 55-item set and items were iteratively removed based on outcomes of a CFA sequence, Mokken Analysis, Rasch modelling and local dependency analyses.

Figure 3-9. Gc:VL ITOS Percentage of Items Correct

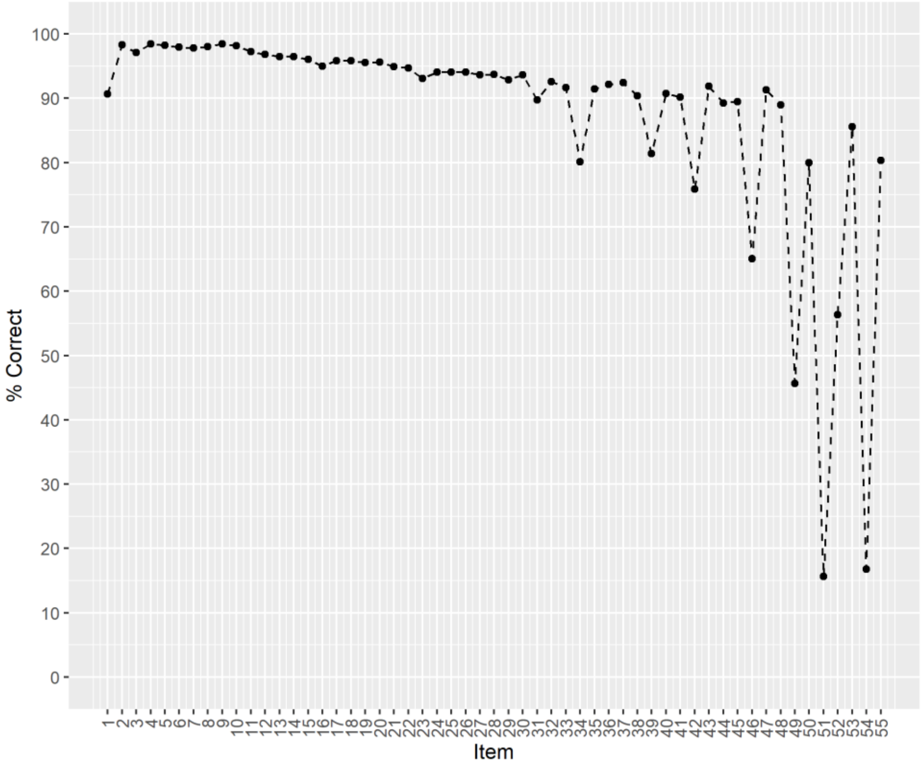


Figure 3-10. Gc:VL ITOS Total Score by Time Taken

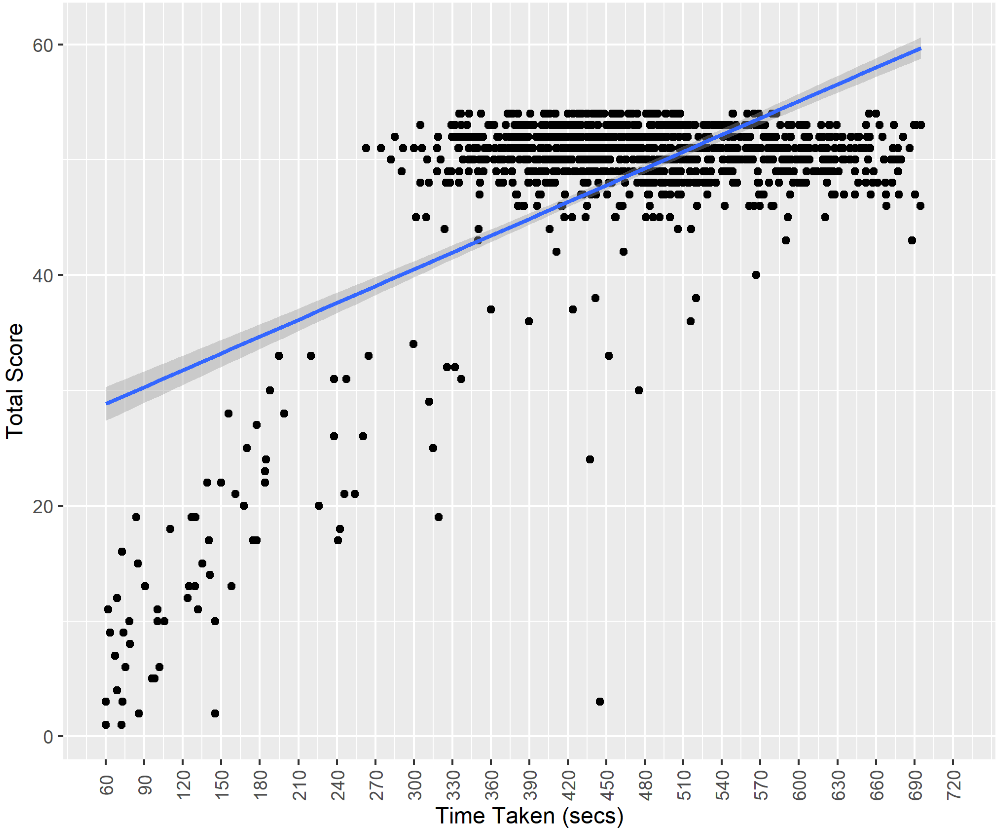


Figure 3-11. Gc:VL ITOS Gender by Total Score

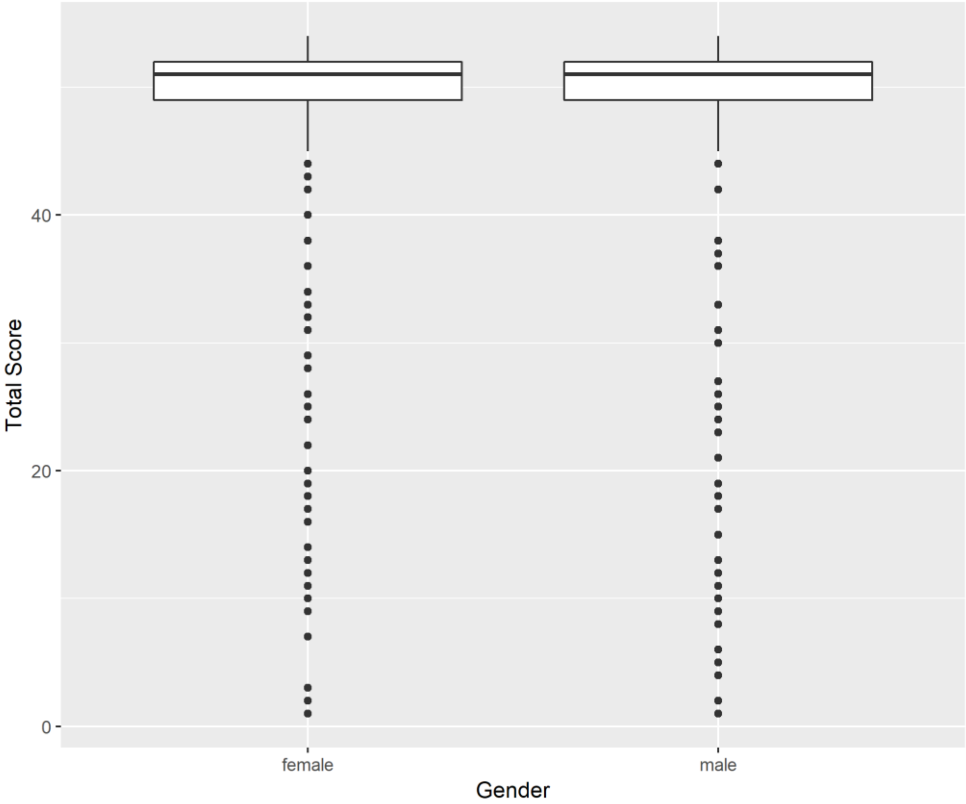
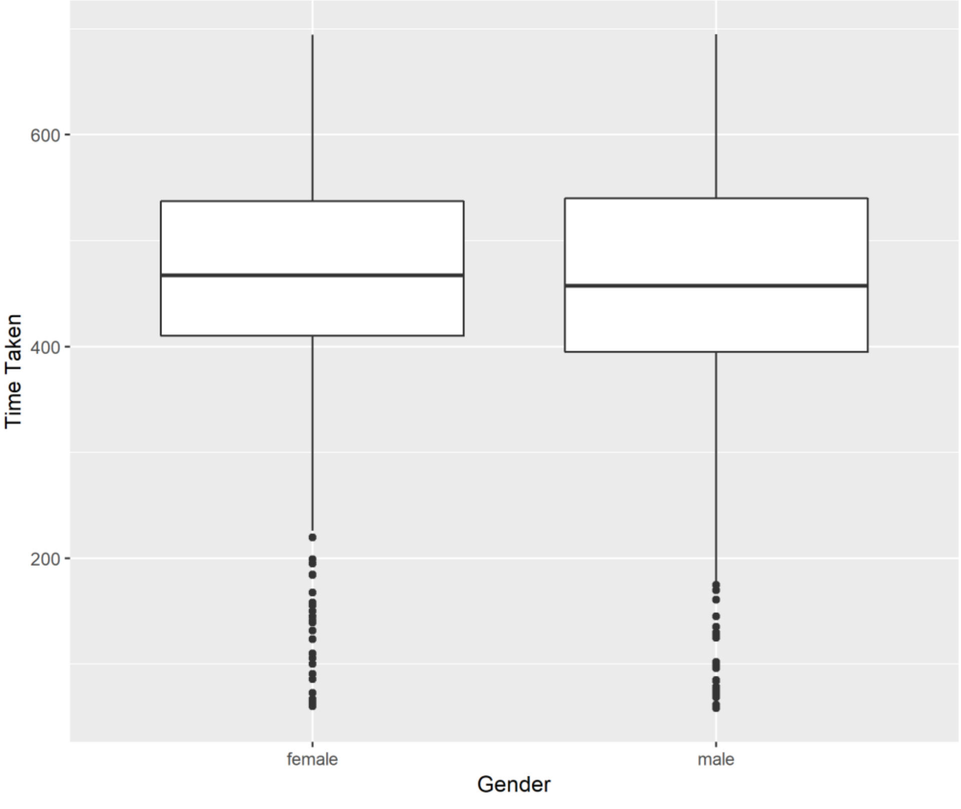


Figure 3-12. Gc:VL ITOS Gender by Time Taken



3.4.3.1 55-items. The fit indices for the initial 55-item model did not meet the threshold considered to be acceptable for any of the indices. At an item level the CFA suggested that all items load significantly on the same factor except items 1 and 3. There was large variation in the size of the standardized betas but only 8 items displayed standardized betas below .30. Items 1 and 3 were removed prior to Mokken analysis.

3.4.3.2 53-items. After the removal of items 1 and 3, there appeared to be little improvement in the CFA fit. The overall Loevinger's *H* value for the 53-item model suggested the item set had good ordering. Only items 2, 4, 5, 6 and 51 displayed Loevinger's *H* values below .50, with only item 2 being below .30. Items 2 and 52 were determined to be unscalable (i.e. received a 0 in the AISP) but no items were determined to be on different Mokken scales. A decision was made to remove items 2 and 52.

3.4.3.3 51-items. The CFA and Mokken analysis outcomes of the 51-item model were quite like that of the 53-item model. Subsequent Rasch modelling found that while RMSEA indicated a mediocre fit with the Rasch model, and the TLI and CFI indicated an acceptable fit, the SRMR index was well above the desired cutoff. Item level statistics revealed that 30 items did not fit the Rasch model. A high amount of local dependence was detected within the items. Due to the low difficulty of items identified earlier, 40 items were removed in a sequential order starting from the lowest item number. Five items were then removed for displaying misfit with the Rasch model.

3.4.3.4 6-items. For the 6-item model the CFA produced acceptable fit indices and all items displayed standardized beta values above .4. Mokken analysis continued to suggest there was good ordering of items; all items demonstrated appropriate Loevinger's *H* values and continued to remain on the same Mokken scale via AISP. One item continued to display low fit with the Rasch model; due to the low number of items remaining and the pre-determined limit of four iterations of analysis, this was not removed. No local dependency was evident in the final six items remaining. Marginal reliability was .41 (and Cronbach's alpha analysis resulted in an alpha of .76) suggesting that there may not be enough numbers of items remaining to reliably measure Lexical Knowledge in participants. Item characteristic curves (ICCs) are in Figure 3-13 and the Test Information Curve (TIC) is in Figure 3-14. Difficulty parameters are in Table 3-3.

3.4.4 Differential Item Functioning

DIF for gender was flagged for items 39 ("Subterranean") and 48 ("Artichoke"). This was determined to be negligible; only a small probability of difference was evident.

Table 3-2. Gc:VL CFA Fit, Mokken Analyses and Rasch Analyses

	55 items	**53 items**	**51 items**	**6 items**
CFA Fit Index				
χ^2	18477.01***	18252.99***	17764.77***	22.68**
CFI	.699	.701	.705	.991
TLI	.688	.689	.693	.985
RMSEA	.105	.109	.112	.038
SRMR	.089	.091	.092	.019
Mokken Analysis				
Loevinger's *H*	-	.728	.740	.544
Standard Error	-	.026	.026	.026
Rasch Analysis				
M2	-	-	6353.81***	116.85***
RMSEA	-	-	.061	.083
SRMR	-	-	.333	.119
TLI	-	-	.985	.953
CFI	-	-	.985	.956
Marginal Rxx	-	-	.692	.413

Note. **p* < .05; ***p* <0.01; ****p* < .001

Figure 3-13. Gc:VL ITOS Rasch ICCs

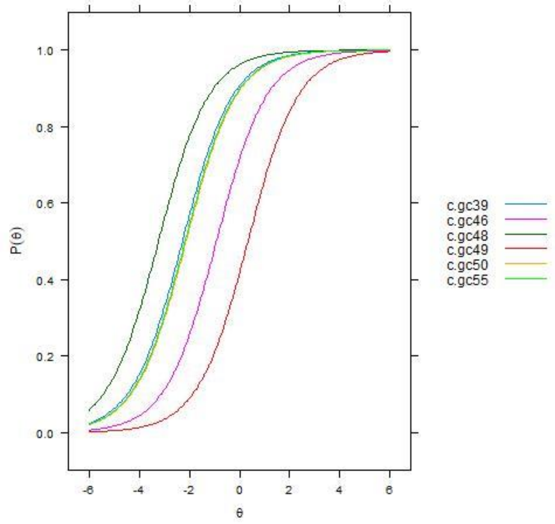
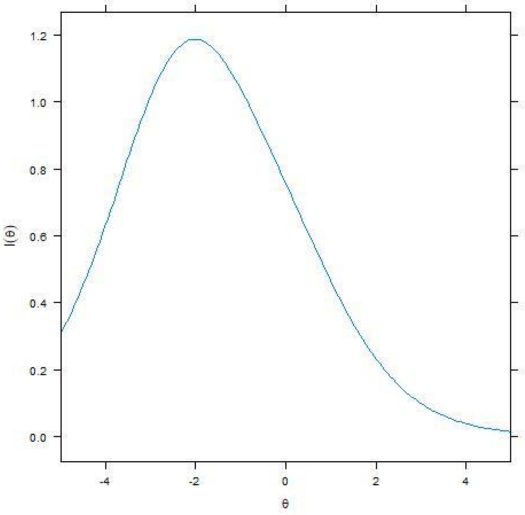


Figure 3-14. Gc:VL ITOS Rasch TIC



Note. Six items were retained for the Gc:VL test with a peak at approximately -2 theta.

Table 3-3. Gc:VL ITOS Rasch Item Parameters

Item	Difficulty (*b*)
39 (Subterranean)	-2.27
46 (Lintel)	-0.92
48 (Artichoke)	-3.22
49 (Ovine)	0.33
50 (Apiarist)	-2.13
55 (Thylacine)	-2.17

Note. These are traditional/classical IRT parameters

3.5 Induction Results

3.5.1 Raw Score Outcomes

The first 11 items were all answered correctly by over 90% of participants, while subsequent items showed a significant decline; 17 items were answered correctly less than 60% of the time. The mean score for the sample was 20.49 (*SD* = 4.09). This is best demonstrated by Figure 3-15. Generally, as people took more time to complete the items, their overall performance improved (Figure 3-16).

There was no significant difference between females and males on total score, *t*(507) = -0.69, *p* = .49, 95% CI [-0.91, 0.44] (Figure 3-17), or on time taken *t*(520) = 0.53, *p* = .59, 95% CI [-21.033, 36.60] (Figure 3-18). These results also suggest there were similar dropout rates among males and females.

Because of the significant drop in performance after item 11 two additional analyses were conducted to determine if there were differences in performance on certain *types* of Induction items. Items were classified by their matrix type (5 categories) and the amount of response options (3 categories). This is displayed in Table 3-4. Items were then additionally classified based on whether they were provided 30 seconds or 60 seconds (Table 3-5). Three approaches to determining possible differences in items were attempted; an ANOVA followed by two separate logistic regressions.

Table 3-4. Gf:I ITOS Item Classification

	4 Options	**5 Options**	**6 Options**
5 x 1 Matrix	6 items	3 items	0 items
2 x 2 Matrix	9 items	0 items	0 items
4x1 Matrix	4 items	0 items	0 items
3x3 Matrix	0 items	5 items	3 items
4x2 Matrix	0 items	3 items	0 items

Table 3-5. Gf:I ITOS Item Frequency Table

Item Type	**30 seconds**	**60 seconds**
5x1 Matrix 4 Response Options	4 items	2 items
2x2 Matrix 4 Response Options	7 items	2 items
4x1 Matrix 4 Response Options	3 items	1 item
3x3 Matrix 5 Response Options	1 item	4 items
3x3 Matrix 6 Response Options	0 items	3 items
4x2 Matrix 5 Response Options	0 items	3 items
5x1 Matrix 5 Response Options	0 items	3 items

Figure 3-15. Gf:I ITOS Percentage of Items Correct

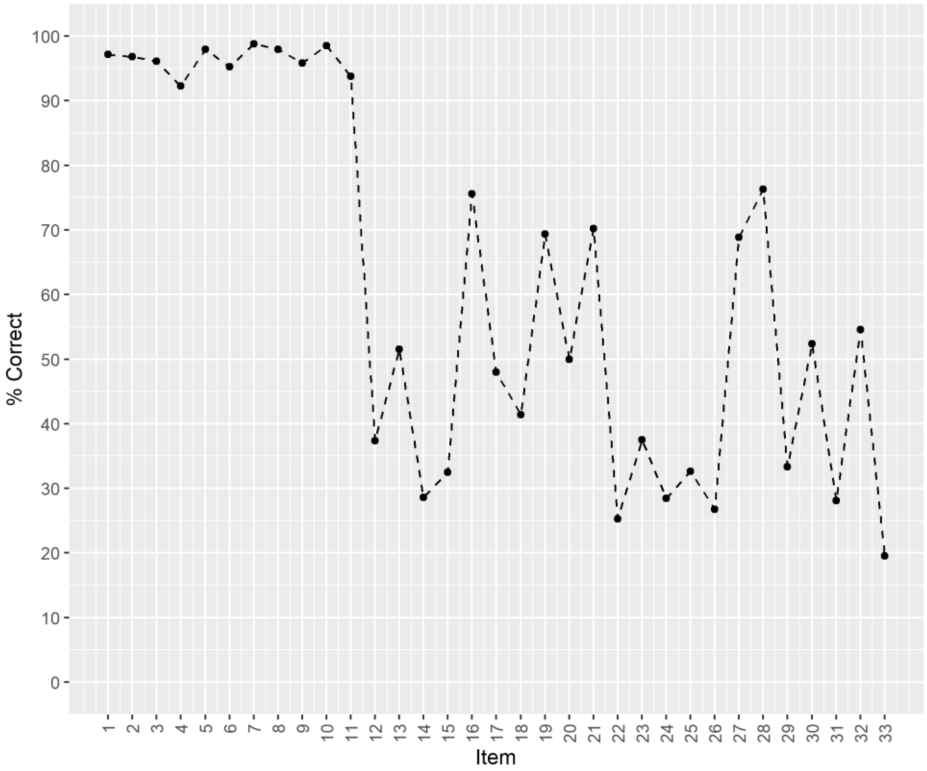


Figure 3-16. Gf:I ITOS Total Score by Time Taken

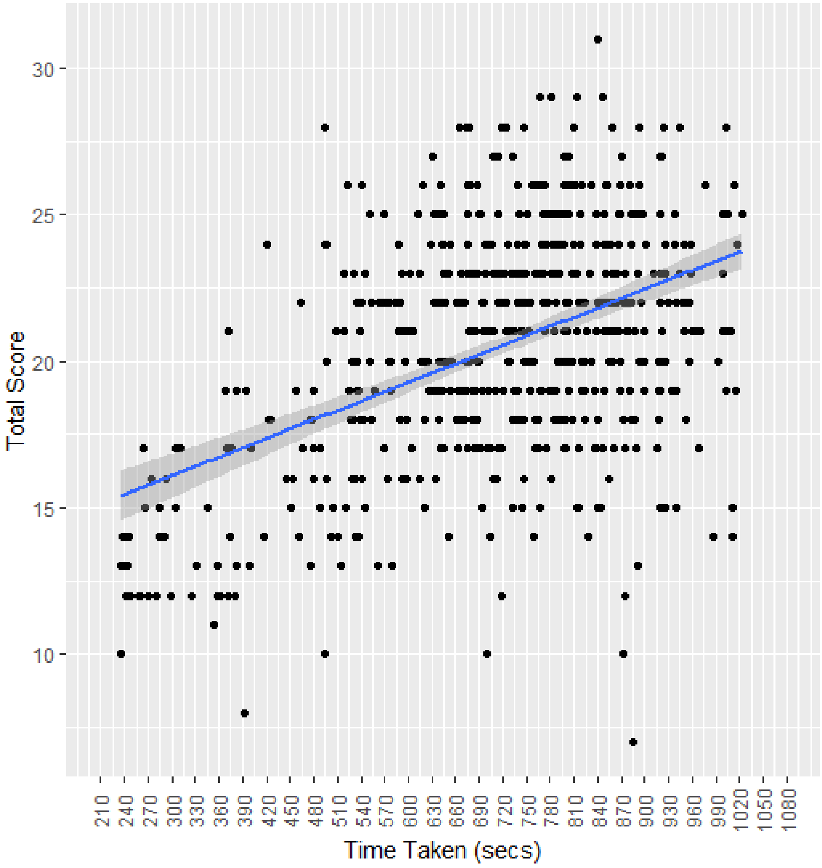


Figure 3-17. Gf:I ITOS Gender by Total Score

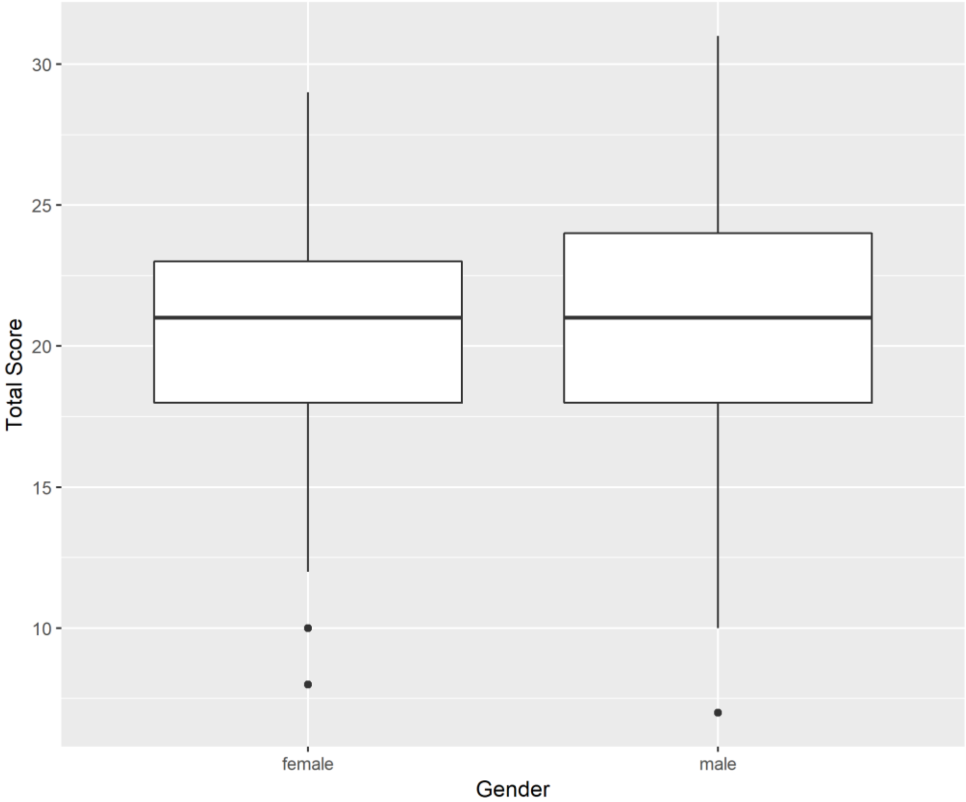
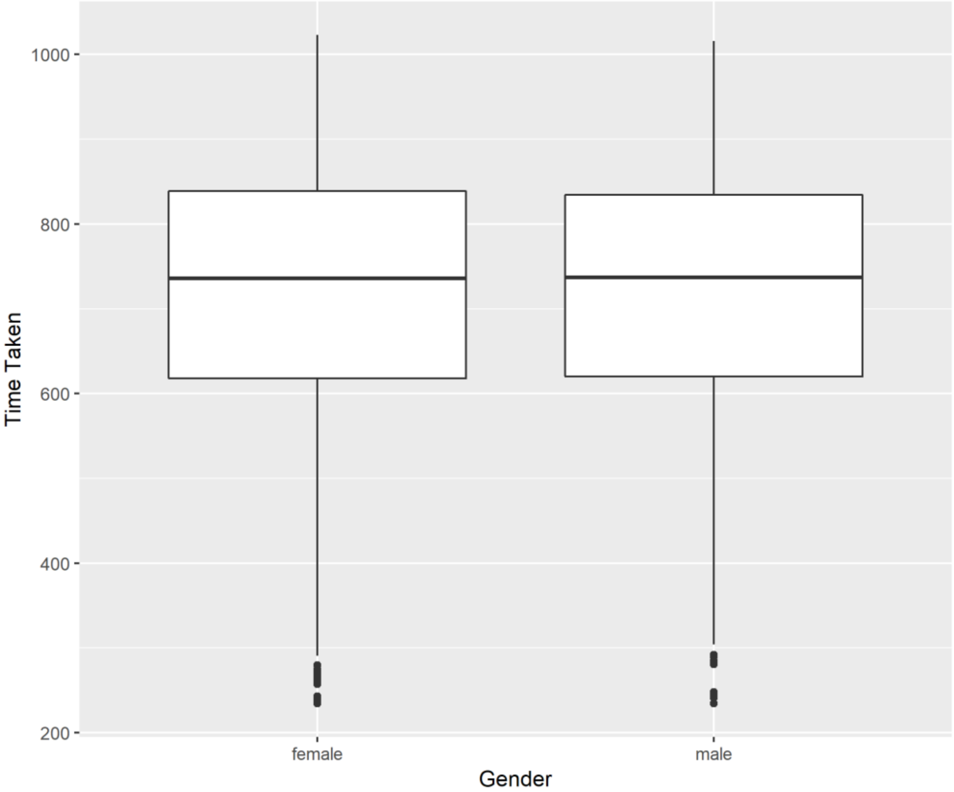


Figure 3-18. Gf:I ITOS Gender by Time Taken



An unbalanced Two-Way ANOVA using Type II (Yates method of fitting constants) was used (Langsrud, 2003) based on the percentage of participants that obtained a correct item as the dependent variable, and the combined item type and number of seconds allowed (as per Table 3-5) as independent variables. The analysis is considered unbalanced as there is an uneven number of items in each cell. This analysis was run utilising the ANOVA function in the *car v.3.0-2* package for R (Fox & Weisberg, 2011). The interaction effect between item type and time allowed was not significant, $F(3,22) = 1.90$, $p = .15$, nor was there a significant main effect for item type, $F(6,22) = .97$, $p = .46$, or time allowed, $F(1,22) = 3.46$, $p = .08$.

A factorial logistic regression (Table 3-6) was carried out with items categorised by matrix and response options separately (i.e., correct ~ matrixtype + responseoptions + timeallowed). The model includes dummy variables for each matrix type contrasted against the 5x1 Matrix item type, for each number of response options against the smallest amount (4 options), and for time allowed to respond to a question (30 seconds).

Table 3-6. Gf:I ITOS Item Format Logistic Regression

Variable	*B*	SE	Wald *Z* Statistic	Odds Ratio
Intercept	1.76***	.05	37.62	5.87
2x2 Matrix	-0.31***	.05	-5.74	0.73
4x1 Matrix	-0.43***	.06	-6.81	0.65
3x3 Matrix	0.06	.06	0.95	1.06
4x2 Matrix	-0.20**	.07	-2.83	0.82
Five Response Options	-1.27***	.07	-18.54	0.28
Six Response Options	-0.88***	.09	-9.38	0.41
60 Seconds Allowed	-0.98***	.04	-22.86	0.37

Note. The coefficients for matrix type are contrasts with the 5x1 matrix type. The response option coefficients are contrasted with the smallest number of response options (4 options). The time allowed coefficient is contrasted with the lowest amount of time allowed to respond to a question (30 seconds). *p < .05; **p < .01; ***p < .001

The logistic regression suggests that compared to a 5x1 matrix, a person is 27% less likely to get a 2x2 matrix correct, 35% less likely to get a 4x1 matrix correct, and 18% less likely to get a 4x2 matrix correct. In contrast, a person was 6% more likely to get a 3x3 matrix correct. There were more 3x3 matrices in the 60 second time allowances which may explain this. However, compared to a 30 second item, people were 63% less likely to get a 60 second item correct, likely due to 60 second items being at the more difficult end of the spectrum. These items included five or six response options which acted as additional distractors. In line with this finding, a person was 72% less likely to answer a five-response option item correctly, and 59% less likely to answer a six-response option item correctly when compared to an item with four response options. Items with five response options seemed to be more difficult than items with six response options, and there were insufficient numbers of items across each category to accurately measure the impact of time allowances. It is expected this may result in problems from an IRT perspective given that items are supposed to predictably increase in difficulty.

3.5.2 Reliability

Cronbach’s alpha for the Induction items was .71. This suggests that the items have good internal consistency.

3.5.3 Item Analysis

Fit indices for the CFA, Loevinger’s *H* values, and Rasch modelling are in Table 3-7. Item level statistics are in Appendix C and Appendix D. Item analysis began with the full 33-item set and items were iteratively removed based on the outcomes of a sequence of CFA, Mokken Analysis, Rasch modelling and local dependency analysis.

3.5.3.1 33-items. While the 33-item model could be described as a good fit based on the RMSEA and SRMR values, the CFI and TLI contradicts this. All items displayed standardized beta coefficients above .1, and showed significant factor loadings, $p < .05$ except for items 2 and 9 which were subsequently removed prior to Mokken Analysis.

3.5.3.2 31-items. A 31-item-set did not scale well. A high number of items demonstrated Loevinger's *H* values below 0.3. Seventeen items were determined to be unscalable (i.e. received a 0 in the AISP), while five items were determined to be on different Mokken scales (i.e. received a 2 or 3 in the AISP). This suggested that the item set may be multidimensional. Removal of these items resulted in a 9-item scale.

3.5.3.3 9-items. CFA of this item set did not improve the model fit; while CFI and TLI improved slightly, RMSEA and SRMR deteriorated. Despite this, all standardized betas were above .1. Although the 9-item-set did not display ideal CFA fit indices, the Mokken analysis resulted in a Leovinger's *H* at an acceptable level. All items had individual Leovinger's *H* values above .30 and the AISP placed all items on the same Mokken scale (Appendix A). Local dependence analysis using Yen's Q3 suggested there was a high amount of local dependency (0.35) between items 7 and 10. Item 7 had poor fit in the initial Rasch, and thus was removed from the item set.

3.5.3.4 8-items. A CFA using the remaining items demonstrated excellent fit compared to the 33- and 9-item sets. All items loaded on the same factor and the Mokken results were good, with Loevinger's *H* values above 0.3. The overall 8-item scale also demonstrated a Loevinger's *H* value above 0.3.

These items displayed good fit with the Rasch model, as did the overall item set. The ICCs (Figure 3-19) and TIC (Figure 3-20) demonstrate the remaining items measure from a very low difficulty of approximately -5 θ through to approximately 1.5 θ. The item parameters are in Table 3-8. No local dependency was evident in the final 8-items. Marginal reliability was .51 and additional Cronbach's alpha analysis resulted in an alpha of .57, suggesting an insufficient number of items remaining to reliably measure Induction. ICCs are shown in Figure 3-19, and a TIC in Figure 3-20.

3.5.4 Differential Item Functioning

An insufficient number of females obtained an incorrect answer on item 10 to enable full DIF analysis. However, after removal of item 10 DIF was flagged for item 27. This DIF was determined to be negligible based on the small change of probability.

Table 3-7. Gf:I ITOS CFA Fit, Mokken Analyses and Rasch Analyses

	33 items	**31 items**	**9 items**	**8 items**
CFA Fit Index				
χ2	954.88***	785.57***	137.29***	30.21
CFI	.67	.23	.76	.97
TLI	.65	.70	.68	.96
RMSEA	.04	.04	.08	.03
SRMR	.05	.05	.06	.03
Mokken Analysis				
Loevinger's H	-	0.15	0.37	0.37
Standard Error	-	.01	.03	.03
Rasch Analysis				
M2	-	-	74.97***	44.80*
RMSEA	-	-	.03	.03
SRMR	-	-	.07	.05
TLI	-	-	0.93	0.96
CFI	-	-	0.93	0.96
Marginal Reliability	-	-	.51	.51

Note. *p < .05; **p < .01; ***p < .001

Figure 3-19. Gf:I ITOS Rasch ICCs

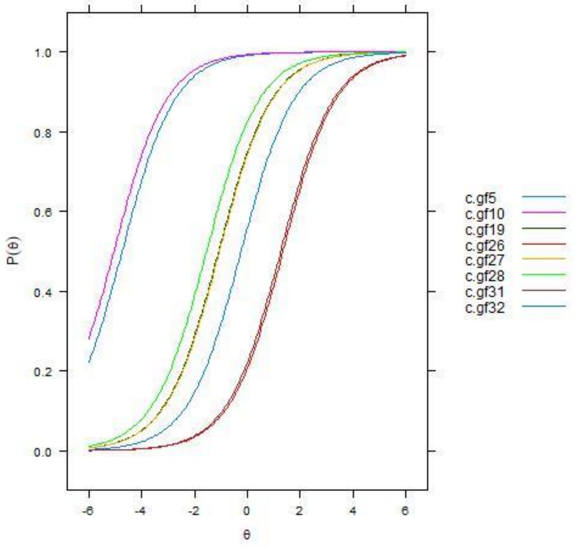
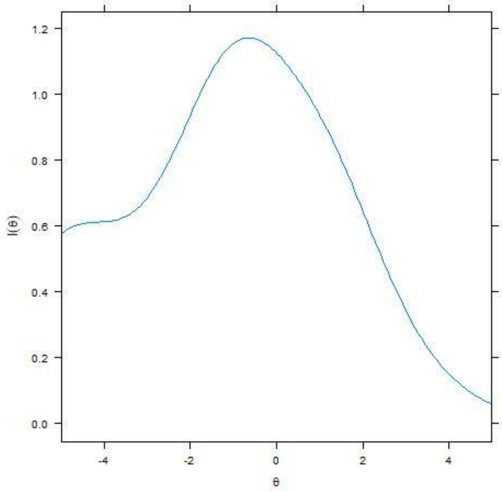


Figure 3-20. Gf:I ITOS Rasch TIC



Note. Eight items were retained from the Gf:I analysis with a peak at approximately 0 theta.

Table 3-8. Gf:I ITOS Rasch Item Parameters

Item	**Difficulty (*b*)**
5	-4.73
10	-5.05
19	-1.08
26	1.34
27	-1.05
28	-1.54
31	1.25
32	-0.23

Note. These are traditional/classical IRT parameters

3.6 Visualisation Results

3.6.1 Raw Score Outcomes

The mean score for the Visualisation items was 36.33 (*SD* = 10.21). Figure 3-21 shows these items had a good range of performance; most items were scored correctly by at least 30% of the sample. Generally, participants who spent longer on the test displayed a higher score improvement (Figure 3-22). The average time taken was 602.41 seconds (*SD* = 198.97 seconds).

There was no significant difference between females and males on total score, *t*(681) = 0.40, *p* = .68, 95% CI [-1.16, 1.77], or time taken, *t*(721) = 2.11, *p* = .03, 95% CI [2.10, 58.53]. This is illustrated in Figure 3-23 and Figure 3-24. These results also suggest similar dropout rates among males and females.

Differences in item performance based on item type (i.e., number of pieces or whether the shape had a border) were analysed further. Results in Table 3-9 indicate that as the number of puzzle pieces increased the likelihood of success decreased by 63%, while the presence of a border increased the chances of success by 158%.

Table 3-9. Gv:Vz ITOS Item Format Logistic Regression

Variable	***B***	**SE**	**Wald *Z* Statistic**	**Odds Ratio**
Intercept	2.87***	.06	47.37	17.80
Pieces	-0.97***	.02	-41.81	0.37
Border	0.95***	.02	41.35	2.58

3.6.2 Reliability

Cronbach's alpha for the Visualisation items was .93. This suggests the items have good internal consistency.

3.6.3 Item Analysis

Fit indices for the CFA, Loevinger's *H* values, and Rasch modelling are displayed in Table 3-10. Item level statistics are in Appendix E and Appendix F. The item analysis began with the full 52-item set and items were iteratively removed based on the outcomes of a sequence of CFA, Mokken Analysis, Rasch modelling and local dependency analysis.

Figure 3-21. Gv:Vz ITOS Percentage of Items Correct

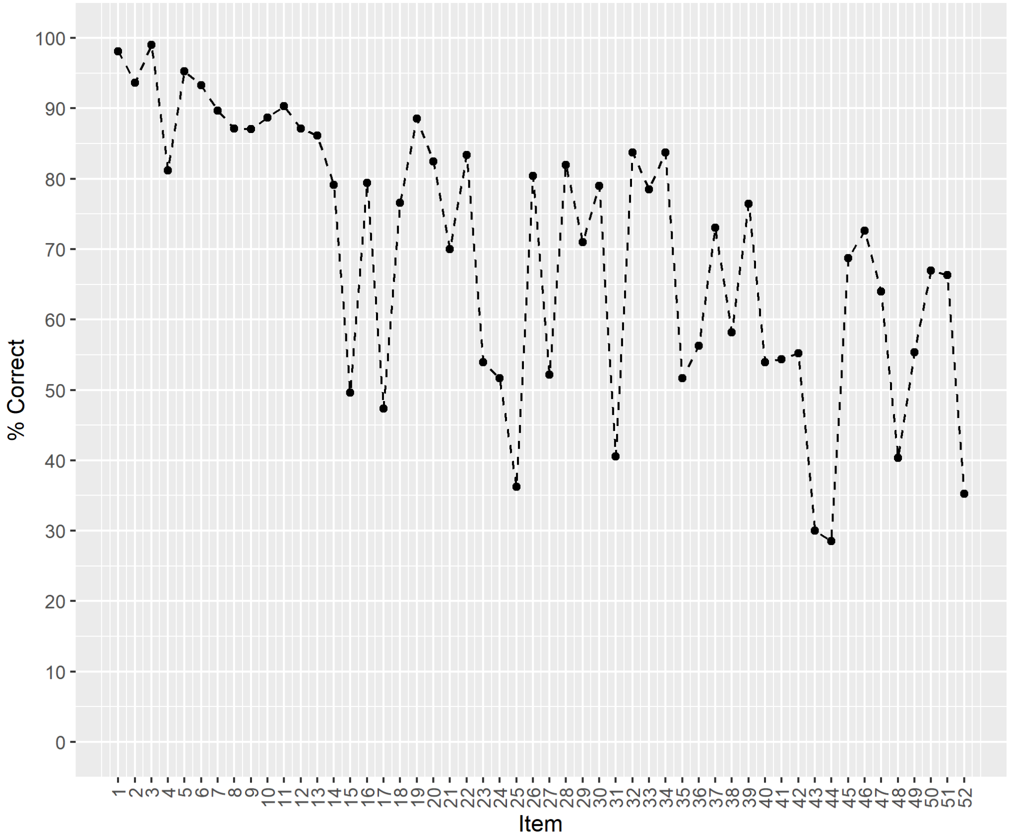


Figure 3-22. Gv:Vz ITOS Total Score by Time Taken

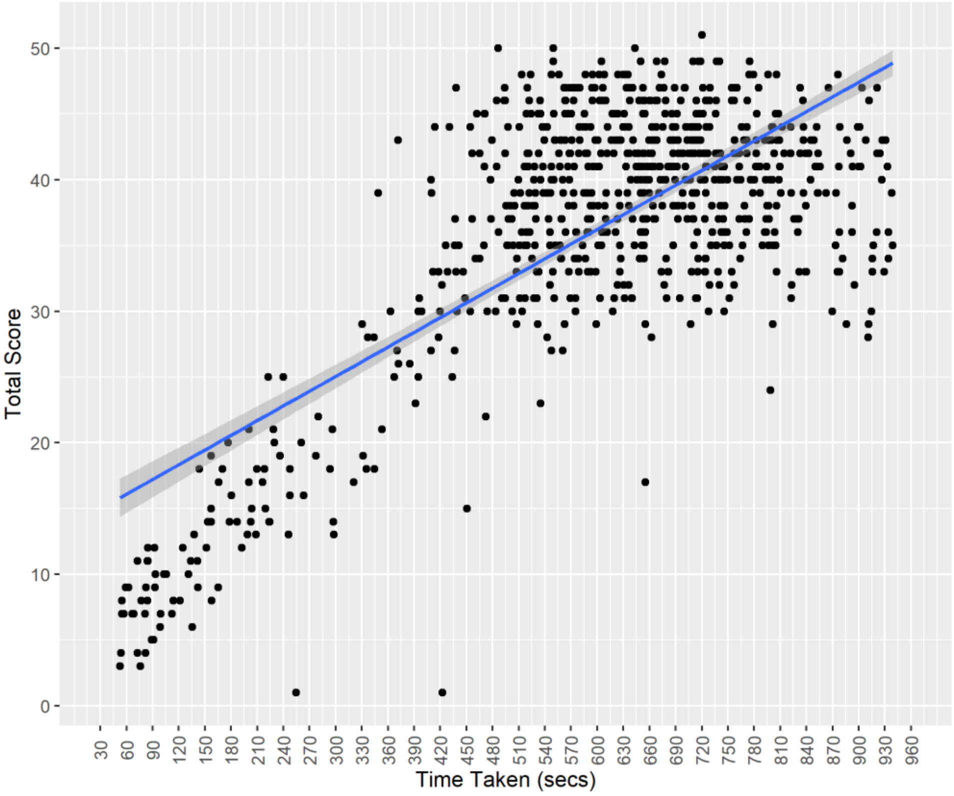


Figure 3-23. Gv:Vz ITOS Gender by Total Score

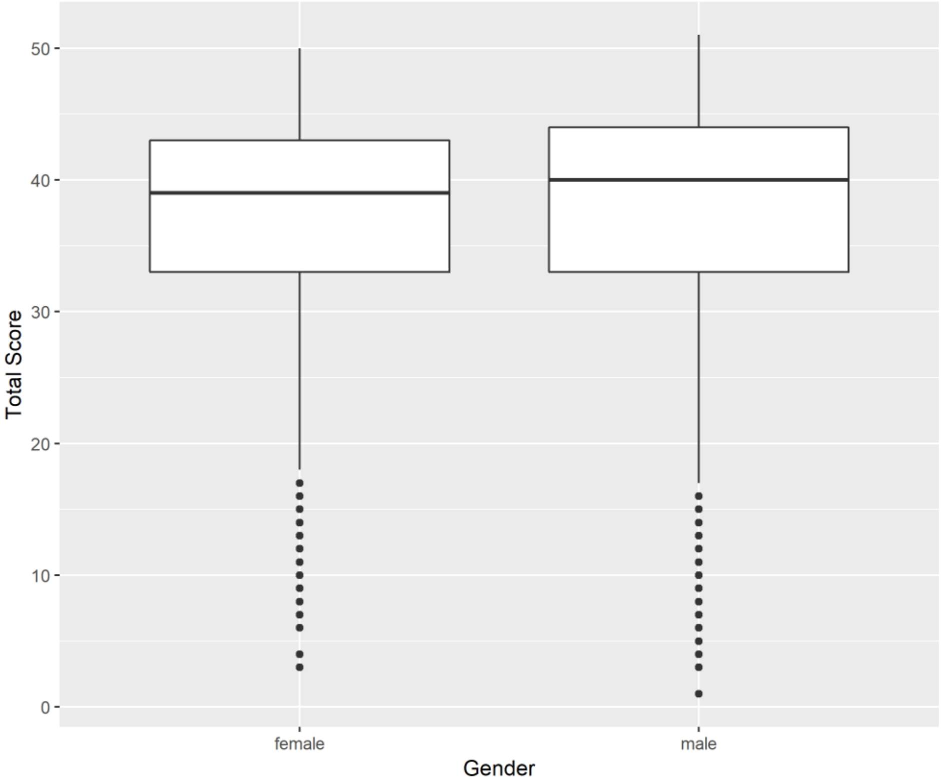
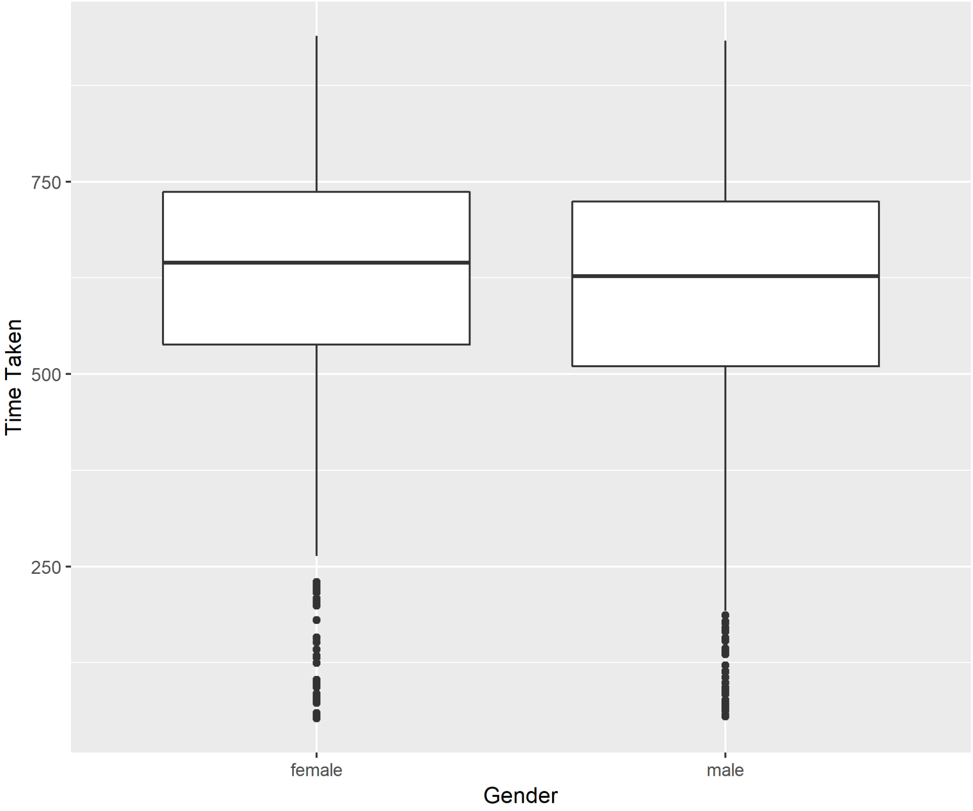


Figure 3-24. Gv:Vz ITOS Gender by Time Taken



3.6.3.1 52-items. Although the CFA showed poor fit, most items appeared to load on the same construct. The first 10 items, along with four items scattered through the item set, displayed standardized betas below .30. The overall Loevinger's *H* was .341, however 16 items displayed a Loevinger's *H* below .30. The AISP removed 15 items.

3.6.3.2 37-items. After removal of 15 items there was improved fit, implying unidimensionality. However, Rasch modelling suggested several items had poor fit. Additionally, items 11 and 13, items 19 and 20, items 32 and 34, and items 46 and 50, showed local dependence. At least one of each of these item pairs were determined not to fit the Rasch model, resulting in the removal of items that had poorer fit.

3.6.3.3 17-items. All items in this model showed standardized betas above .30 in a CFA. Improvements were noted in the Mokken analysis and all items appeared to be ordered. Although there was some improvement in the fit with the Rasch model it was only minor, with a reduction in marginal reliability due to the reduction in items. Items 19, 22 and 48 continued to display a poor fit with the Rasch model and were removed.

3.6.3.4 14-items. The remaining 14-item set displayed acceptable CFA fit indices and Mokken outcomes. Rasch analysis showed improved overall fit, although items 26 and 28 still showed poor fit. The items were retained to avoid further reduction in the item set, and to abide by the pre-established method of four iterations of item removal. No local dependency was evident in the final 14-items remaining. Marginal reliability was .69, and additional Cronbach's alpha analysis resulted in an alpha of .86. ICCs are in Figure 3-25 and the TIC is in Figure 3-26. The difficulty parameters are in Table 3-11.

3.6.4 Differential Item Functioning

DIF was flagged for items 40, 41 and 46, but was determined to be negligible.

Table 3-10. Gv:Vz ITOS CFA Fit, Mokken Analyses and Rasch Analyses

	52 items	37 items	17 items	14 items
CFA Fit Index				
χ^2	4519.62***	2515.35***	456.94***	307.68***
CFI	.75	.83	.92	.93
TLI	.74	.82	.91	.92
RMSEA	.06	.06	.06	.06
SRMR	.06	.05	.04	.04
Mokken Analysis				
Loevinger's *H*	.35	.44	.47	.44
Standard Error	.02	.02	.02	.02
Rasch Analysis				
M2	-	2900.98***	539.596***	360.97***
RMSEA	-	.06	.06	.06
SRMR	-	.12	.11	.10
TLI	-	0.96	0.97	0.97
CFI	-	0.96	0.97	0.97
Marginal Reliability	-	.84	.72	.69

Note. **p* < .05; ***p* < .01; ****p* < .001

Figure 3-25. Gv:Vz ITOS Rasch ICCs

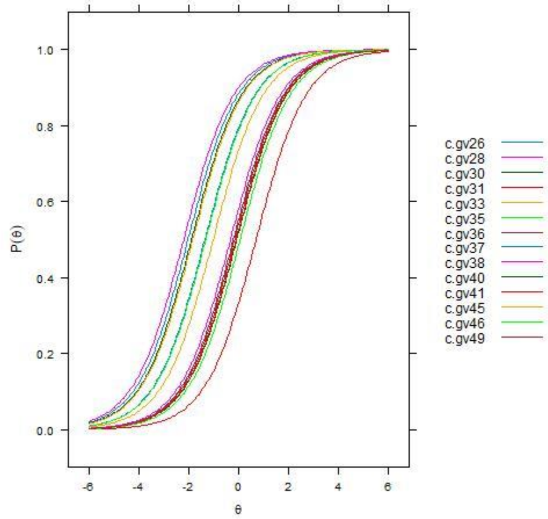
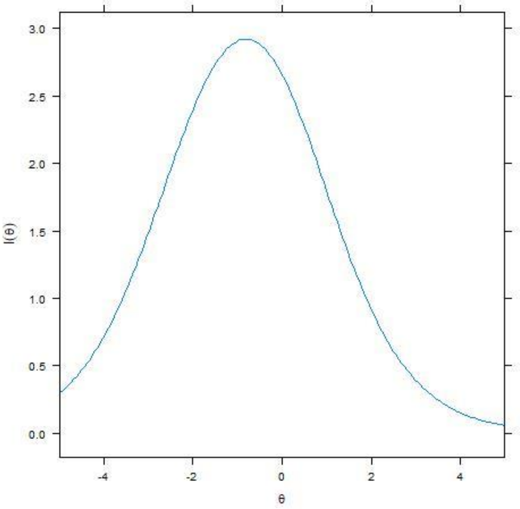


Figure 3-26. Gv:Vz ITOS Rasch TIC



Note. Fourteen items were retained from the Gv:Vz analysis which peaked at approximately -1 theta.

Table 3-11. Gv:Vz ITOS Rasch Item Parameters

Item	Difficulty (*b*)
26	-2.03
28	-2.20
30	-1.88
31	0.69
33	-1.83
35	0.56
36	-0.21
37	-1.35
38	-0.33
40	-0.08
41	-0.10
45	-1.02
46	-1.32
49	-0.16

Note. These are traditional/classical IRT parameters

3.7 Working Memory Results

3.7.1 Total Score Outcomes

The Working Memory items were with increasing number of stimuli to place more cognitive load on the participants as they progressed. The proportion of participants who answered items correctly generally followed this trend (Figure 3-27). Notable discrepancies occurred for items 8, 24, 34 and 35. Item 8 had a slightly different format ("What was the *last number, then* the first number?") compared to the other items surrounding it (e.g., "What were the two numbers?"). Otherwise, no qualitative differences could be identified for these discrepant items. Other than item 8, the first 16 items all had a success rate above 90%. While participants tended to perform better the more time taken, this pattern did not hold as strongly as for the other CHC narrow abilities beyond 10 minutes. The mean total score was 25.6 (*SD* = 5.52) and the mean time taken was 713.44 seconds (*SD* = 186.31 seconds).

There was no significant difference between females and males on total score, *t*(377) = -0.05, *p* = .96, 95% CI [-1.07, 1.02], or time taken, *t*(390) = 0.76, *p* = .44, 95% CI [-21.43, 48.23]. This is illustrated in Figure 3-29 and Figure 3-30. These results also suggest similar dropout rates among males and females.

3.7.2 Reliability

Cronbach's alpha for the Working Memory items was .84. This suggests the items have good internal consistency.

3.7.3 Item Analysis

Fit indices for the CFA, Loevinger's *H* values, and Rasch modelling fit indices are in Table 3-12. Item level statistics are in Appendix G and Appendix H. The item analysis began with the full 38-item set and items were iteratively removed based on the outcomes of a sequence of CFA, Mokken, Rasch and local dependency analyses.

3.7.3.1 38-items. The full item set showed poor fit via the CFA. The CFI and TFI were well below the required cutoff; the RSMEA was mediocre, but the SRMR also did not meet the required cutoff. Item 8 was shown to not fit the model, consistent with the unexpected low performance on this item based on raw data. This item was removed.

3.7.3.2 37-items. Removal of item 8 produced minimal improvement in overall item fit. Mokken analysis suggested that the overall scale is ordered, however several items displayed Loevinger's *H* values below the suggested .30 threshold. The AISP found that 7 items were unscalable and 7 items loaded onto different Mokken scales. These items were removed.

Figure 3-27. Gwm:Wc ITOS Percentage of Items Correct

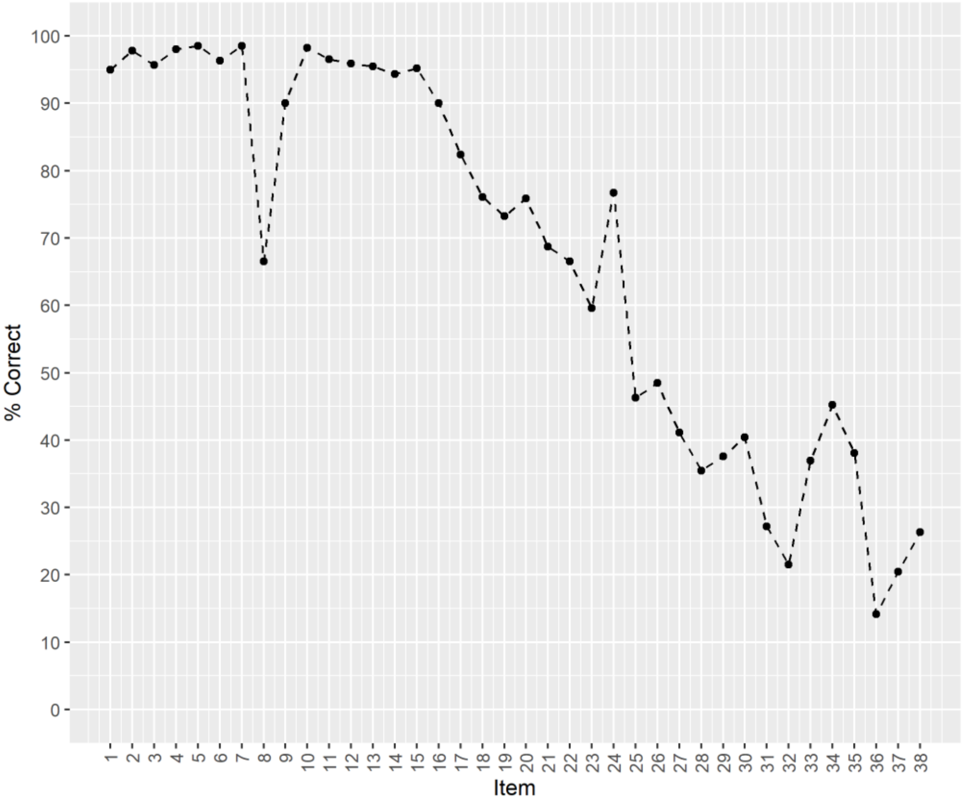


Figure 3-28. Gwm:Wc ITOS Total Score by Time Taken

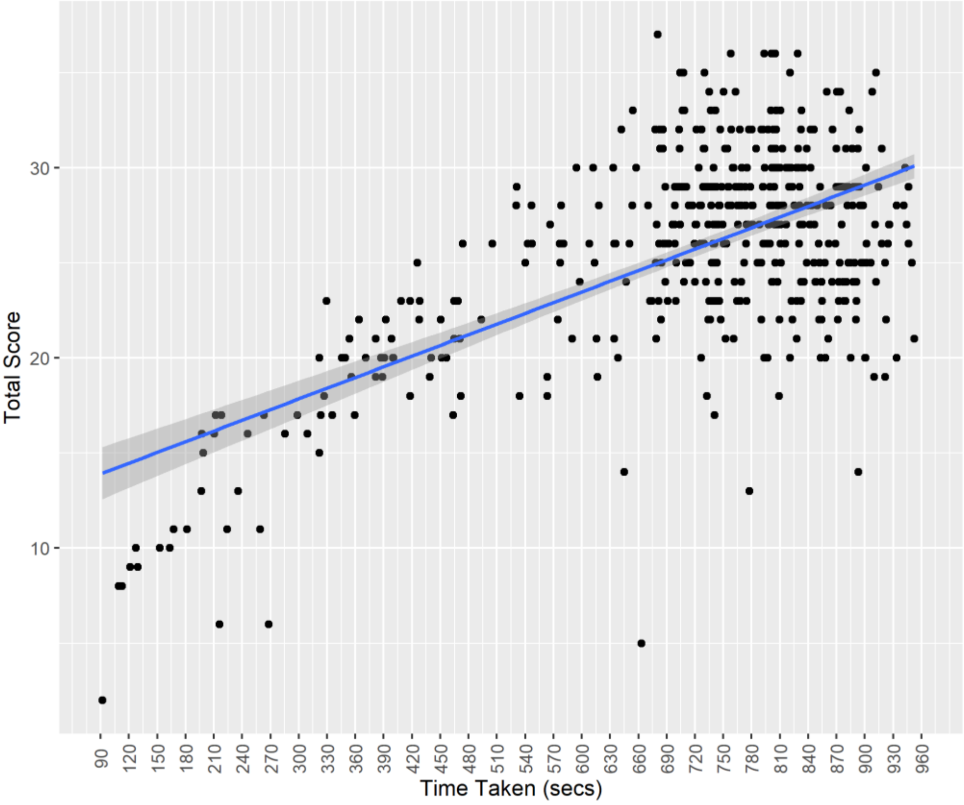


Figure 3-29. Gwm:Wc ITOS Gender by Total Score

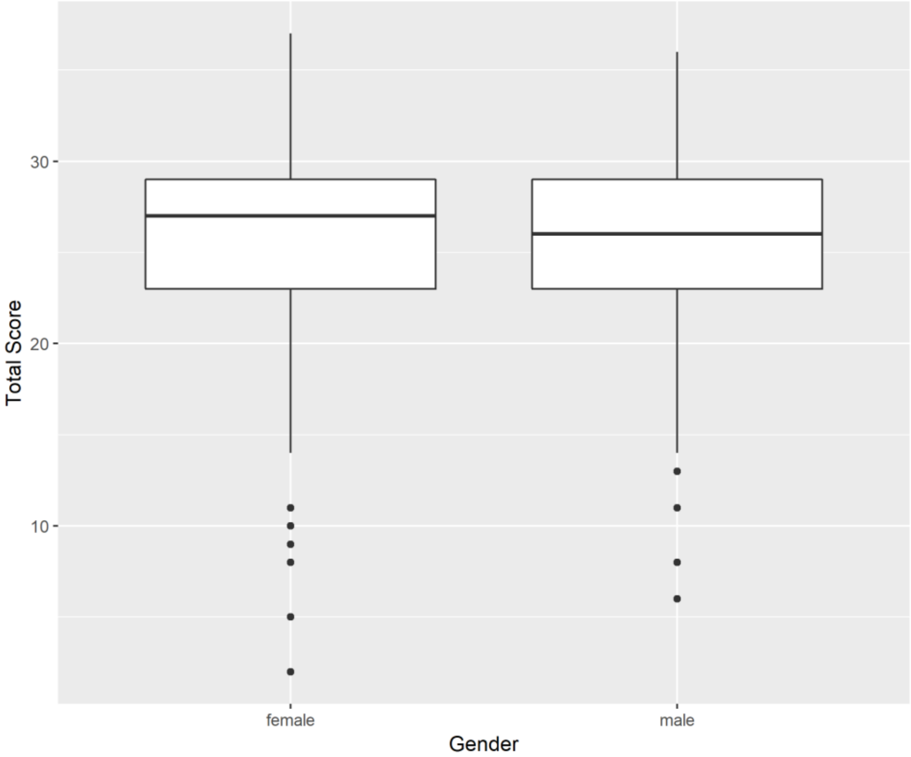
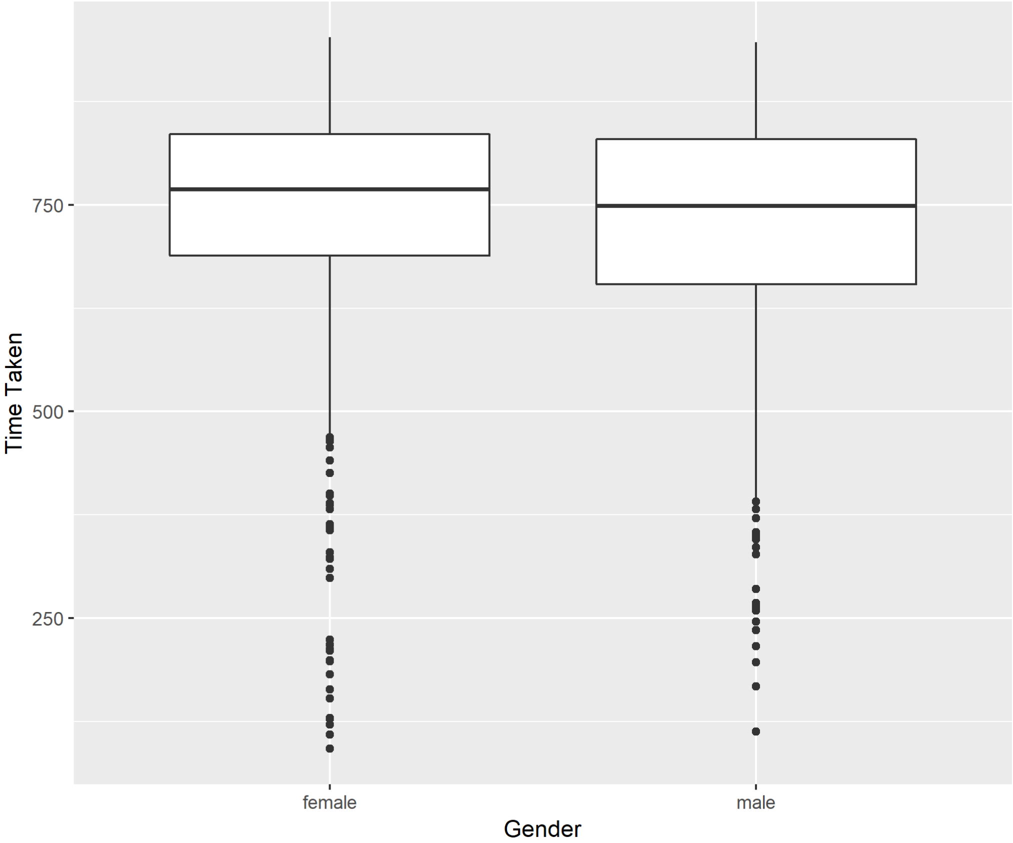


Figure 3-30. Gwm:Wc ITOS Gender by Time Taken



3.7.3.3 23-items. The CFA fit indices improved substantially, although not sufficiently. All items showed standardised betas above .1 and met the requirements via Mokken analysis. The item set sufficiently met the assumption of unidimensionality. Rasch showed that item 10 did not fill well, and Yen's Q3 showed local dependency between several items. Items 7, 10, 12 and 13 were removed.

3.7.3.4 19-items. A CFA using 19 items improved overall fit. Mokken and AISP analysis showed all items were unidimensional. Item 22 had a Loevinger's *H* value of .293 but met the .30 threshold when accounting for standard error. Rasch modelling showed good fit. As per the pre-determined methodology, item removal stopped here. No local dependency was evident in the final 19-items remaining. Marginal reliability was .706, and Cronbach's alpha analysis resulted in an alpha of .80. ICCs are in Figure 3-31 and a TIC in Figure 3-32. Difficulty parameters are shown in Table 3-13.

3.7.4 Differential Item Functioning

An insufficient number of all genders answered items 2 and 11 incorrectly to enable full DIF analysis. After removal of these items no items were flagged for DIF.

Figure 3-31. Gwm:Wc ITOS Rasch ICCs

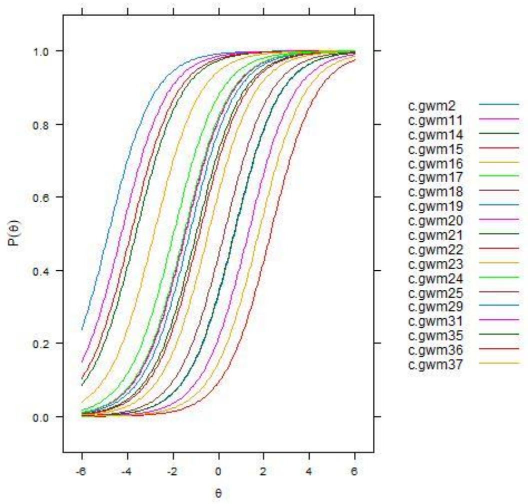
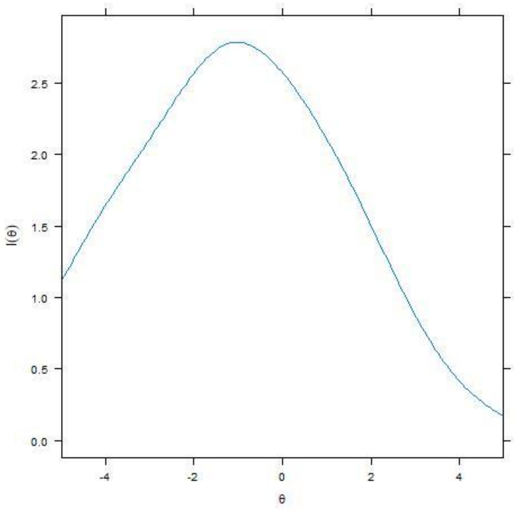


Figure 3-32. Gwm:Wc ITOS Rasch TIC



Note. Nineteen items were retained from the Gwm:Wc analysis with a peak at approximately -2 theta.

Table 3-12 Gwm:Wc ITOS CFA Fit, Mokken Analyses and Rasch Analyses

	38 items	**37 items**	**23 items**	**19 items**
CFA Fit Index				
χ^2	2183.88***	2123.31***	870.27***	437.00***
CFI	.59	.60	.74	.81
TLI	.57	.57	.72	.78
RMSEA	.07	.07	.08	.06
SRMR	.08	.08	.07	.06
Mokken Analysis				
Loevinger's *H*	-	.31	.39	.36
Standard Error	-	.03	.03	.03
Rasch Analysis				
M2	-	-	665.77***	380.40***
RMSEA	-	-	.06	.05
SRMR	-	-	.14	.10
TLI	-	-	0.93	0.94
CFI	-	-	0.93	.094
Marginal Rxx	-	-	.71	.71

Note. *p < .05; **p < .01; ***p < .001

Table 3-13. Gwm:Wc ITOS Rasch Item Parameters

Item	**Difficulty (*b*)**	**Item**	**Difficulty (*b*)**
2	-4.834	22	-0.840
11	-4.256	23	-0.448
14	-3.621	24	-1.506
15	-3.844	25	0.247
16	-2.824	29	0.705
17	-1.966	31	1.305
18	-1.459	35	0.681
19	-1.262	36	2.287
20	-1.443	37	1.760
21	-0.971		

Note. These are traditional/classical IRT parameters

3.8 Discussion

The development of a cognitive ability test with good reliability and validity requires items to demonstrate robust psychometrics. Knowing item psychometrics also allows the test developer to decide whether further items will be developed, and to understand the impact of item design decisions on the probability of success. The current ITOS was designed to pilot the items designed in four recent research projects. Rather than incorporating items from existing test tools, the current item sets are designed to fit within the taxonomy of CHC abilities. A relatively conservative approach was taken to demonstrate robust psychometrics for the item sets retained.

Overall, the current analysis retained different items than those retained in prior analyses, suggesting that items require further calibration and analysis to demonstrate their invariance across samples. Additionally, analysis resulted in the removal of a high number of items, further suggesting problems within the item sets. Maguire (2018) retained eight Lexical Knowledge items while the ITOS retained only six; in both cases this is a significant loss of items from the initial 55. Originally developing 33 items, Fleming (2018) retained seven Induction items while the ITOS retained eight. There were slight improvements in the number of items retained for Visualisation, with Heng (2018) retaining 15 items and the ITOS retaining 14 items. For Working Memory Kennedy (2018) retained 13 items while the ITOS retained 19 items.

While there was some overlap in the retention of Lexical Knowledge items, there were also some differences. Issues were noted with local dependency, possibly due to low variability in performance, and Rasch modelling. Analysis of items that fit poorly with the Rasch model suggest a potential difference in what some items are measuring (e.g., differences between General (Verbal) Information and Lexical Knowledge). Further item development was clearly required and occurs in the next chapter.

Different statistical issues were noted for the Induction items. While the removal of items largely resulted from violations of unidimensionality and monotonicity, the item difficulty appeared to vary substantially after item 11, and the time taken by participants was inconsistent. Items that were clearly more difficult from a qualitative and theoretical perspective were answered correctly by participants who answered easier items incorrectly. A more robust source of data is required than unproctored Internet participants, as increases in difficulty do not appear to be related to the rules utilised. The next chapter takes into consideration the rules within each item.

Relative to the other CHC narrow abilities Visualisation displayed more overlap in the items retained between the previous projects and the current ITOS relative to the other CHC narrow abilities. Only one item retained by Heng was omitted in the current study. Difficulties in Mokken scaling and Rasch modelling may be due to earlier items possibly testing Speed Rotation rather than Visualisation. Additionally, the difficulty increase throughout items was not consistent with previous research on internal cues and number of pieces. Inspection of items that were retained and removed suggests that piece similarity rather than internal cues and number of pieces impacted difficulty.

Finally, Working Memory items appeared to follow an expected difficulty increase based on the number of stimulus chunks. Item 8 was clearly violating the assumption of unidimensionality, and items prior to this appeared to be too easy. Problematically, due to the way items were originally designed and the way in which they were implemented in Concerto, response options were available to participants during the administration of the item. Based on these findings, more difficult items were required, as well as changes to the item design so that answer options are not visible. This is addressed in the following chapter.

The items in this analysis were largely developed to measure the ability levels of children. Unfortunately, due to ethical limitations (which are later addressed in chapter 4), data collection was focused on adults. As abilities should theoretically develop in a sequential manner it was initially believed that the items would still demonstrate fit with the Rasch model. However, problematically many items were far too easy. Therefore, while the above paragraphs demonstrate that there are some psychometric and design problems to be addressed in the items, there was also problems with the targeting of the items for the sample ultimately collected for this analysis.

Overall, the ITOS resulted in items with good psychometrics for four core cognitive abilities known to be important for learning. Unfortunately, insufficient items were retained to be useful in a CAT. After completion of the ITOS, several improvements were identified to enhance the item sets. Firstly, more items needed to be developed for all four item sets to increase the content range (i.e., difficulty). Secondly, further item development focused on the characteristics of Induction and Visualisation items was required to address the apparent discrepancy between suggestions in the literature regarding how to vary difficulty in such items, and what the ITOS analysis found. Thirdly, an evaluation of the items by education and assessment professionals was deemed necessary to classify items that may have construct irrelevant variance. Finally, collection of data from those with innately lower abilities (such as children) would be beneficial in assessing the difficulty range of the item sets; unfortunately, many items were removed likely due to low variability in the data set. The next chapter aims to address these issues.

Chapter 4: Item Calibration Study

4.1 Introduction

To implement a computer adaptive test (CAT), a set of items with known parameters is required. While the Item Tryout Study (ITOS) established a set of difficulty parameters and strong item psychometrics, an overly conservative statistical approach was employed. The ITOS demonstrated that items retained under such statistical analysis are variable, and only few items meet the assumptions when using strict cutoffs that may infringe on theory-based item design decisions. Across the four item sets there were some issues of poor fit with the Rasch model, local dependency, Mokken scaling and items being too easy. The current chapter details an Item Calibration Study (ICS) which aimed to increase the number of items available for a CAT, with items designed with Cattell-Horn-Carroll (CHC) theory in mind. The outcome of the current study aimed to achieve sets of item parameters that measure a wide range of the theoretical spectrum of each CHC ability under investigation.

4.2 Chapter Aim

To establish useful item difficulty parameters, improvements needed to be implemented following the ITOS. Several problems were identified with the initial items developed for each CHC scale as part of the ITOS. These problems included psychometric approaches to item selection resulting in small sets of items being retained, identifying items of extremely low difficulty, and only a narrow range of difficulty being measured by the remaining item set. This was believed to be largely symptomatic of a high mean age within the sample and a low ceiling within the difficulty of the item sets. These issues resulted in low variance within the data set (i.e., a high number of individuals achieved very high scores). New items were developed for the ICS to address this.

The ICS also aimed to extend the methodology utilised in the ITOS. This included the use of multivariate imputation by chained equations (MICE) to estimate missing data, rather than biasing the data by assuming all unattempted items were incorrect. As part of this ICS, additional items were developed based on the CHC theory discussed in earlier chapters as an attempt to measure the constructs of interest more broadly. The goal of the analysis in the current chapter was to establish a set of items that measure a wide range of difficulty for each CHC ability, to be utilised in a CAT simulation. The current chapter places a focus on Rasch analysis relative to other statistical methods used in the previous chapter.

4.3 Method

4.3.1 Participants

Data for this study was collected from three sources. The first source was the data from the 2,776 participants of the ITOS. The second source was additional recruitment of participants via Facebook (ICS Adult; ICS-A). As with the ITOS, these participants were recruited through paid Facebook advertisements, distribution of a link (chctest.com.au) on social media websites, a Facebook page (facebook.com/chctest) and snowball recruitment. The paid Facebook advertisement was completed in two stages, with the first focused on 18- to 90-year-olds for one week and subsequently 18- to 25-year-olds for two weeks. While 2,185 users completed the demographic information, only 1,929 participants proceeded any further. The third data source was face-to-face administration to school aged children which consisted of 144 participants (ICS School Aged; *ICS-U*).

Data was cleaned, merged, and recoded utilising the *dplyr v.0.8.5* (Wickham et al., 2020), *tidyr v.1.0.2* (Wickham & Henry, 2020) and *stringr v.1.4.0* (Wickham, 2019) packages, via *R v.3.6.3* (R Core Team, 2019) within the *R Studio Integrated Development Environment v.1.2.5033* (R Studio Team, 2019).

Participants that received a score of zero across all four items sets were removed (Table 4-1). It was technically possible for a participant to enter one item set and decide not to proceed, which would have resulted in a score of zero for that item set and resulting in a score of zero across all four item sets. Such participants are not useful for any stage of analysis in this chapter and will subsequently be referred to as 'non-responders'.

Removal of participants above impacted demographic data but not the number of participants that completed each CHC test; if a participant received a score of zero then they are assumed to have not completed the test.

Table 4-3, displays the number of females and males within each study, before and after removal of participants that were identified as non-responders. Prior to removal of non-responders, there were 2,473 (51.00%) participants that identified as females, 2,260 (46.65%) participants that identified as males, and 80 (1.65%) participants that identified as another gender. After removal of non-responders, there were 1,762 (51.34%) participants that identified as females, 1,587 (46.24%) participants that identified as males, and 49 (1.43%) participants that identified as another gender. The proportion of each gender across the sample before and after cleaning suggests that a similar rate of each gender dropped out of the study, which is consistent with dropout rate in the ITOS. While there was no missing gender data, a total of 34 participants explicitly selected Prefer Not to Say and these were retained after cleaning the sample.

Whether a person was untruthful or preferred not to reveal their age does not impact the underlying probability of whether they would obtain a correct answer. Any invalid ages were recoded as ‘NA’ (e.g., above 90, those in the ICS Adult sample stating ages below 18). The distribution of ages by group was relatively similar before (Figure 4-1) and after cleaning (Figure 4-2). After cleaning, there were 3,401 participants with valid ages (*M* = 39.79 years, *SD* = 19.41 years) with a range of age of 6 to 90 years old (Figure 4-3 and Figure 4-4).

Table 4-4 displays the nationality of participants within each study, before and after removal of non-responders. Any missing data for this variable was recoded as Prefer Not to Say (PNTS).

Unlike the initial analysis for the ITOS, participants at the highest and lowest ends of time taken and total score were not removed. In the initial analysis there was an attempt to balance Classical Test Theory (CTT) and Item Response Theory (IRT); CTT approaches focus on test level performance rather than item level performance, which means that individuals who performed quite low or quite high have an undue influence on the outcomes of psychometric analyses. In the current analysis, with a greater focus on IRT, any attempt at an item is valuable data, as it assists in understanding the probability that a person may or may not obtain a correct answer. However, retaining such data needed to be balanced with avoidance of biasing any strategy utilised to address missing data. Missing data was addressed using MICE, rather than full information maximum likelihood (FIML), or scoring unattempted items incorrect as in the ITOS. This is discussed in further detail below. The number of participants in each stage of analysis is shown in Figure 4-5.

Table 4-1. ICS Participants by Data Collection Phase

Sample	ITOS	ICS Adult	ICS School Aged	Total
Initial Sample	2776	1929	144	4849
Remaining Sample	1376	1913	143	3432

Note. Remaining sample after non-responders removed

Table 4-2. ICS Subtest Participant Counts by Data Collection Phase

Phase of Data Collection	G*c:*VL[1]	G*v:*Vz[2]	G*f*:I[3]	Gwm:Wc[4]
ITOS (*n* = 1376)	1203	876	670	521
ICS Adult (*n* = 1913)	1130	648	877	315
ICS School Aged (*n* = 143)	95	103	97	88
Total Sample (*n* = 3432)	2428	1644	1627	924

Note. [1]Lexical Knowledge; [2]Visualisation; [3]Induction; [4]Working Memory

Table 4-3. ICS Participants by Gender

Sample	Female	Male	Other	Prefer Not to Say
Initial Sample				
ITOS	1439	1290	47	0
ICS Adult	965	897	33	34
ICS School Aged	69	75	0	0
Total	2473	2262	80	34
Cleaned Sample				
ITOS	735	624	17	0
ICS Adult	958	889	32	34
ICS School Aged	69	74	0	0
Total	1762	1587	49	34

Table 4-4. ICS Nationality of by Data Collection Phase

Nationality	ITOS	ICS Adult	ICS School Aged	Total
Initial Sample				
Australia	2195	1723	144	4062
Non-Australian	498	189	0	687
PNTS	83	17	0	100
Cleaned Sample				
Australia	1126	1708	143	2977
Non-Australian	221	188	0	409
PNTS	29	17	0	46

Figure 4-1. ICS Participants by Age Grouping before Data Cleaning

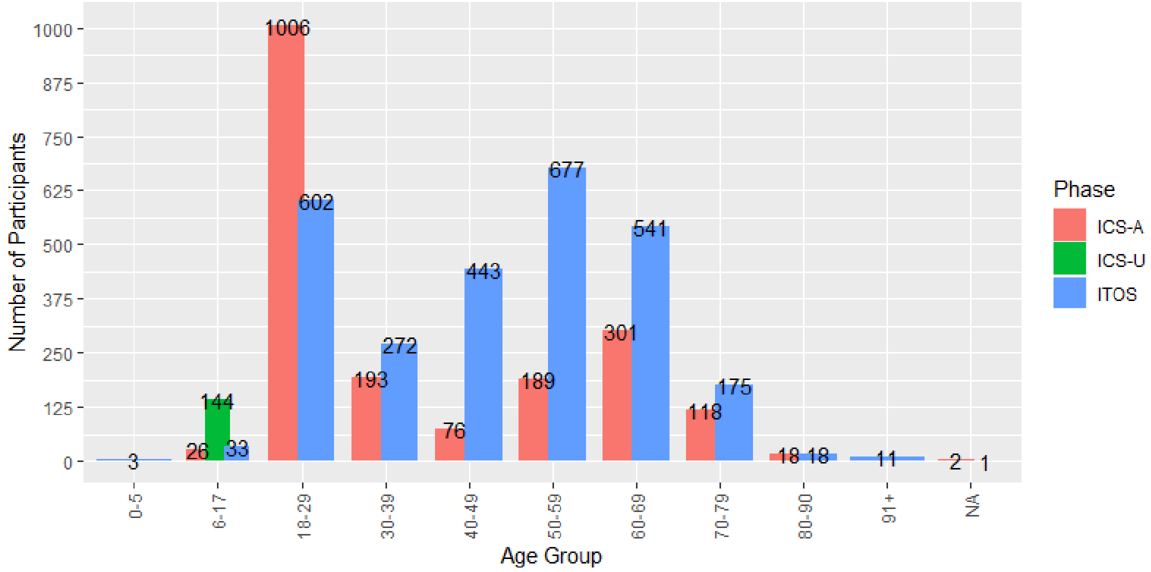
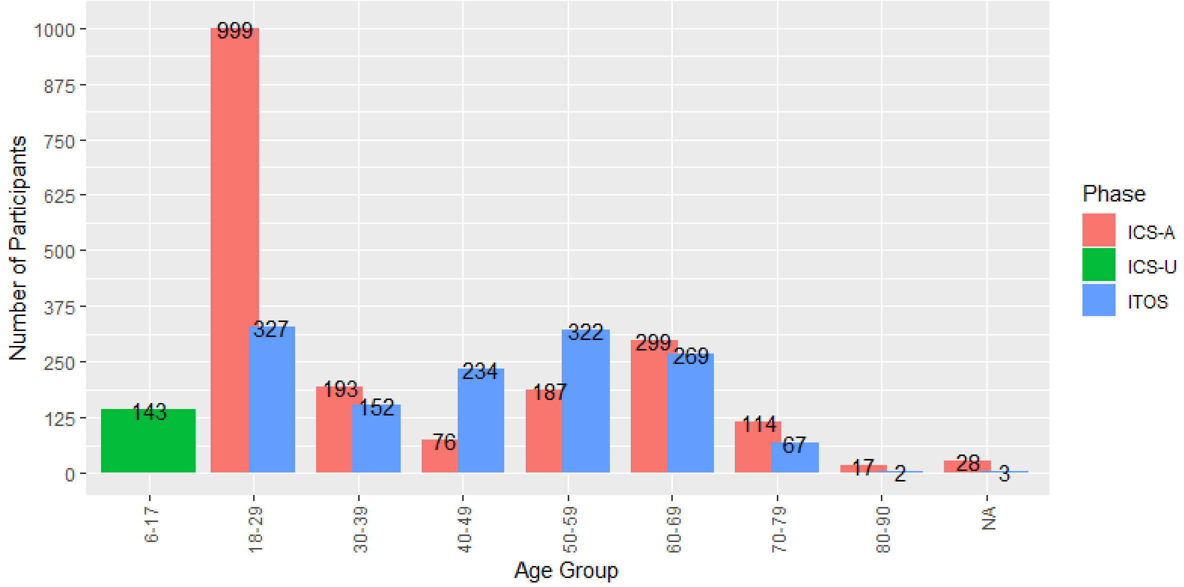


Figure 4-2. ICS Participants by Age Grouping after Data Cleaning



Note. Due to ethical factors and the design of the studies there were more participants in older age groups in the ITOS, followed by more young adults in the ICS-A, and lastly only children in the in the ICS-U.

Figure 4-3. ICS Age of Participants by Phase

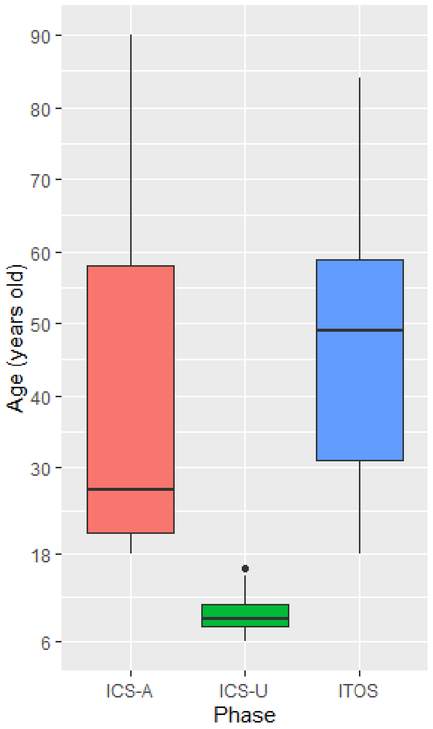
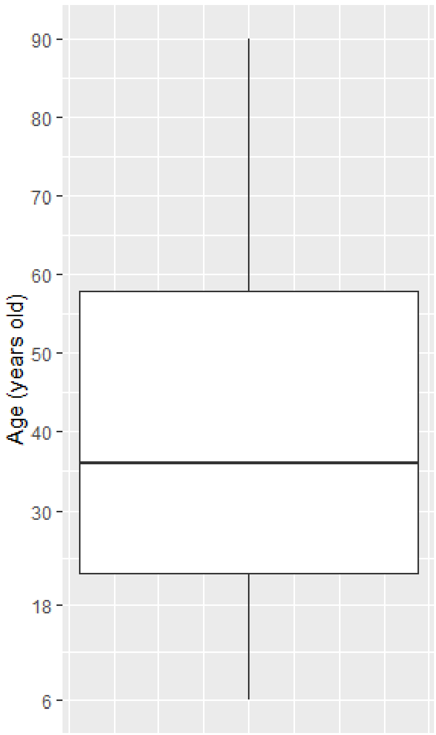
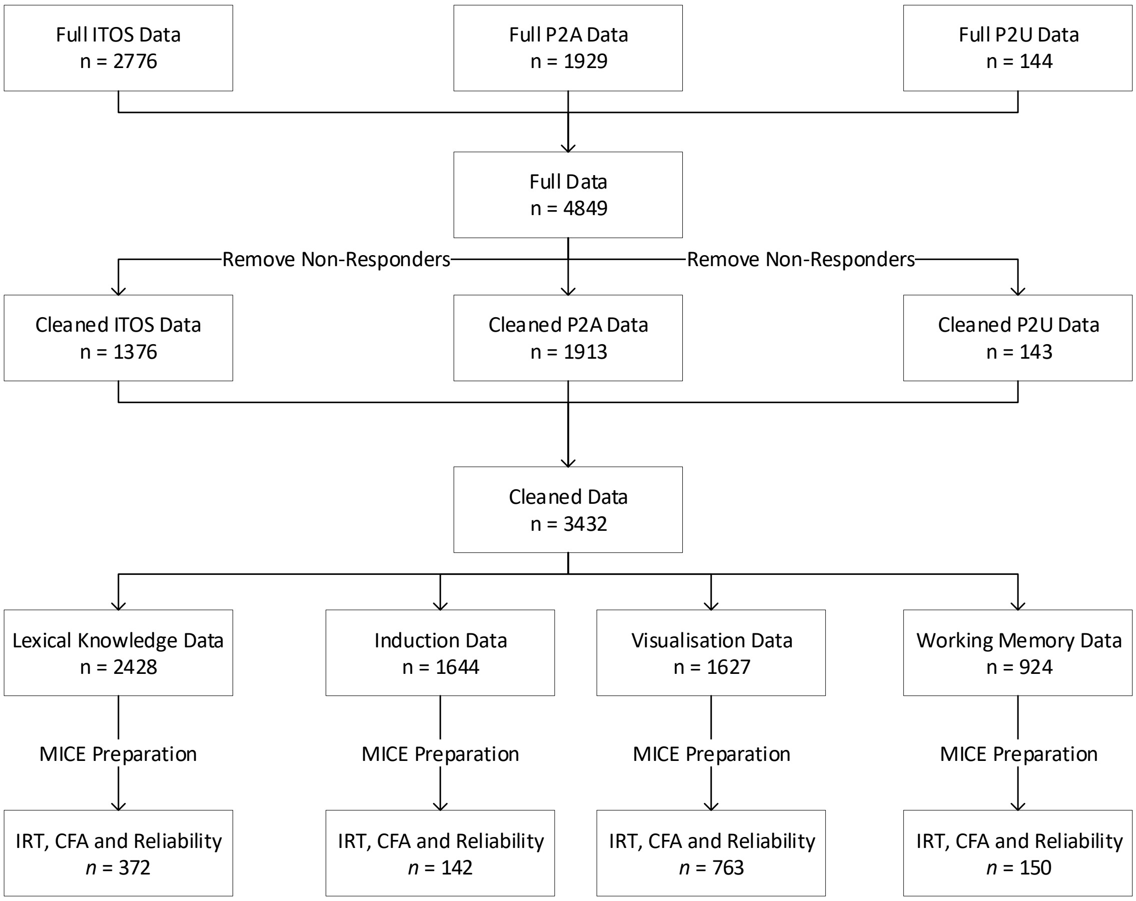


Figure 4-4. ICS Age of All Participants



Note. Due to ethical factors and the design of the studies there were more participants in older age groups in the ITOS, followed by more young adults in the ICS-A, and lastly only children in the in the ICS-U.

Figure 4-5. ICS Participants by Stage of Analysis



4.3.2 Materials

4.3.2.1 Test Platform. The platforms used to host the ICS School Aged and the ICS Adult phases of data collection were Concerto v.5.0.0 and Concerto v.5.0.9 respectively. A changelog between versions is available at Lis (2020). The justifications for the use of the Concerto platform are reflected in the materials section of Chapter 3.

4.3.2.2 Lexical Knowledge Lexical Knowledge (G*c:*VL) is a major narrow ability subsumed by the Comprehension Knowledge broad ability, and is the knowledge of words, their definitions and their related concepts (Schneider & McGrew, 2018). Several issues were noted in both Maguire (2018) and the ITOS analysis with the Lexical Knowledge item set developed for the CHC-CAT. Maguire retained items 38 (Fortnight), 39 (Spanner), 45 (Raptor), 46 (Lintel), 48 (Artichoke), 49 (Ovine), 50 (Apiarist) and 53 (Caucus), while the ITOS analysis omitted items 45 and 53 in favour of 55 (Thylacine). There was an extreme amount of local dependency found throughout the data set, likely due to a low amount of variance. Additionally, given that the English language often contains common etymologies, it is possible for people to derive the meaning of new words based on their knowledge of other words. Local dependency was somewhat unsurprising given that the average English-speaking adult knows more than 42,000 words (Brysbaert et al., 2016). In order to reach a higher level of reliability, it was noted that the finalised Lexical Knowledge part of the CHC-CAT would require items with a varied level of difficulty.

Some items were shown to violate the assumptions of IRT. Four items were suspected to violate the assumption of unidimensionality; 1 (Running), 2 (Circle), 3 (Knee), and 52 (Corpulent). It was suspected that the first three items are such basic knowledge that they are measures of G*c*:K0 rather than G*c*:VL. It is unknown why item 52 was identified as measuring more constructs than the theorised Lexical Knowledge. All four items were removed for subsequent iterations of the CHC-CAT. Additionally, four items were found to have poor fit with the Rasch model (after other item calibrations), 34 (Quadrilateral), 51 (Phlegmatic), 54 (Pernicious) and 55 (Thylacine). Given the large number of items removed prior to Rasch modelling, these items were retained for the ICS to evaluate them when a larger set of items was used.

In order to incorporate the data obtained in the original ITOS in conjunction with additional data obtained in the current study, existing item stimuli could not be significantly altered as this would potentially impact the response style or pattern of performance by participants. It is argued that resizing of images was not enough cause for removal of an item as it retained the original response options; participants that did have problems (outside their innate trait level) due to image size could not be isolated within the original data set. Table 4-5 details the steps taken to improve Lexical Knowledge items.

In developing new items, the concept of vocabulary difficulty was considered. While on the surface there appeared to be a breadth of literature on vocabulary difficulty, it seemed to be a somewhat undefined concept in many studies (e.g. Stahl & Jacobson, 1986). As a result, researchers seem to subjectively choose words based on what they perceive to be difficult or easy or rely on the number of syllables. Twinword, a company that builds Natural Language Analysis tools and software, argues that there "are no scientific [or] mathematic ways to measure vocabulary difficulty" (Twinword, 2016, para. 4). Thus, deciding on which words to include in a test of Lexical Knowledge can be somewhat subjective.

Table 4-5. Updates to Gc:VL for ICS

Problem	Plan	Actions
Breaches of unidimensionality	Remove items	Remove items 1, 2, 3, 52
Poor fit with the Rasch model (unexpected response pattern)	Compare and contrast IRT models in next study	Monitor items 34, 51, 54, 55 during next analysis
Poor presentation on non-desktop devices	Resize all images to width of 300px and maintain aspect ratio Remove items with poor image quality after resizing	Batch resize all images using Photoshop CC Remove items 20, 35 and 47 based on qualitative evaluation of items
Participants with slow internet connection having trouble loading large item files	Set file size (150KB) limit Remove items that exceed this requirement	Due to batch resize, all original images now met this new requirement
Low item difficulty for adults (possible cause of Local Dependency)	Created 62 new items Split item set into three sets	62 items created, with 10 items overlapping in the middle 'difficulty'

The Twinwords (2016) Language Scoring API rates vocabulary difficulty based on its frequency of use as well as how commonly it is used in tests and exams. This is somewhat reflective of the process of teaching English, as we generally use the most common words as beginning words for children to learn (Hinzman & Reed, 2018). Unfortunately, words that appear to be culturally specific (e.g. Koala) or more complex or rare (e.g. Ovine) do not produce a difficulty score via the Twinword Vocabulary Difficulty tool. This is likely due to their algorithms being trained on databases that are not large enough to include these words or are US centric. It is not surprising given that common words are less complex than rare words; they differ in phonemic and graphemic composition, and rare words contain phonemes that are less intelligible (Landauer & Streeter, 1973). The mean and median difficulty of the original item set, as evaluated by Twinword's 10 levels of difficulty, was 4.93 and 4, respectively. This suggests the words were at the easier end of the spectrum.

Beyond increasing the innate difficulty of the word, the difficulty of an item can be increased or decreased by adjusting distractor items. To make an item easier, incorrect answers can be visuals that are clearly wrong. To make items harder, images can be used that represent somewhat related subject matter. However, care needs to be taken as this can create ambiguity, which has implications for Mokken scaling and Rasch analysis because people may start guessing on a multiple-choice test. Overall, given the ease that adults had with ITOS items, the difficulty of words was increased by choosing those that scored above 5 on via Twinword evaluation, and were far less frequent in the Corpus COCA (Davies, 2008) and Corpus iWeb (Davies, 2018) databases of word frequencies. Additional difficulty was added for later items by incorporating increasingly difficult distractors. For example, the word "incensed" (which means "to be very angry") had pictures of incense (a substance that is burned for a fragrant smell), someone very angry (correct answer), a calm person (an antonym) and a flower (also a fragrant object). These strategies were implemented to identify participants that truly understand the definition of a word.

Fifty-two new items were developed. While not all words had a Twinword10 difficulty value or a Corpus COCA and Corpus iWeb frequency count, a comparison between the original items and new items is in Table 4-6. For new items, the Twinword10 difficulty increased and the Corpus COCA and Corpus iWeb frequency decreased, suggesting new words were more difficult than old words.

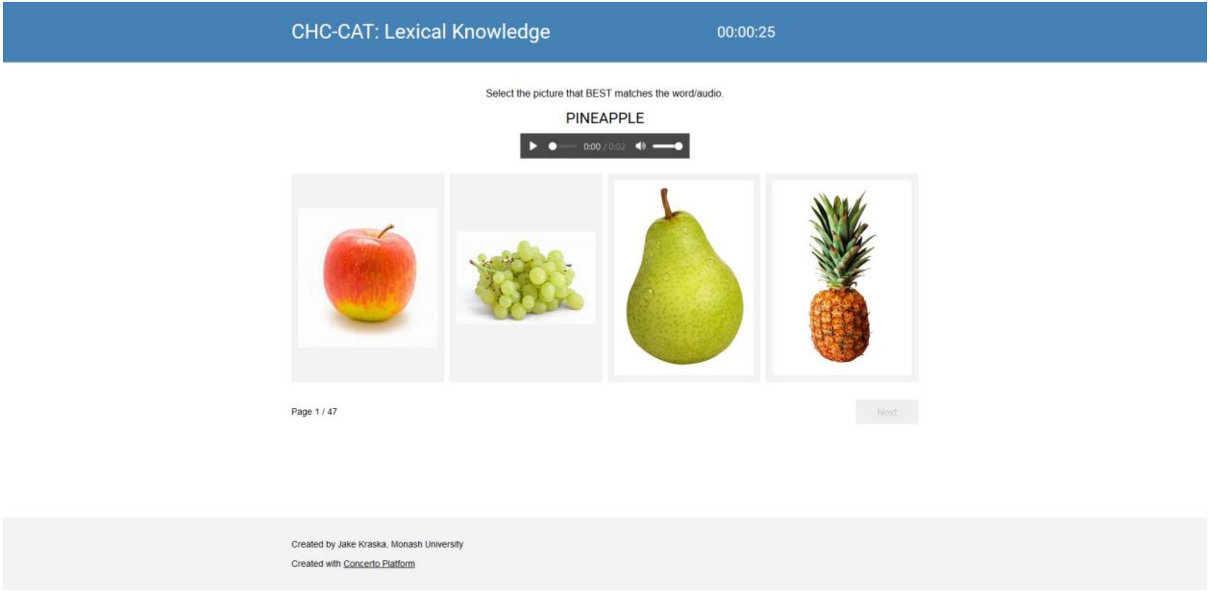
Table 4-6. Average Difficulty and Frequency for Item Development

Source	**Old Items (*n* = 55)**	**New Items (*n* = 52)**	**All Items (*n* = 107)**
Twinword10	4.93	6.74	5.78
Corpus COCA	15,930	3,148	9,722
Corpus iWeb	364,247	92,797	228,522

Consistent with original items, all new items had a monotone female voice recording associated to reduce the reading requirements of the task. This made the task like the *Wechsler Receptive Vocabulary* task, which has been classified by *Cross-Battery Assessment* methodology as a measure of Lexical Knowledge (Flanagan et al., 2013).

Two items were removed from the new set based on the evaluation of a panel of experts (described in the procedure): items 61 and 68. Concerns were raised regarding the ease of the initial three items (for which problems had already been noted from a psychometric perspective in the ITOS). However, concerns were also raised regarding the ability to pictographically represent items 68 (Decriminalise) and 61 (Original) in a non-ambiguous way. These were therefore not included in the ICS data collection for either adults or school aged individuals. Figure 4-6 demonstrates an example the Lexical Knowledge subtest.

Figure 4-6. Example of Gc:VL ICS Item



CHC-CAT: Lexical Knowledge 00:00:25

Select the picture that BEST matches the word/audio.

PINEAPPLE

Page 1 / 47

Created by Jake Kraska, Monash University

Created with Concerto Platform

4.3.2.3 Induction. Induction (G*f*:I) is a major narrow ability subsumed by the Fluid Reasoning broad ability that is considered to be the ability to infer rules, or observe a problem and detect the underlying principles of the problem (Schneider & McGrew, 2018). For the 33 Induction items developed for the CHC-CAT, Fleming (2018) conducted analysis that resulted in retaining seven (items 5, 8, 19, 23, 26, 27 and 32) and the ITOS data analysis retained eight items (items 5, 10, 19, 26, 27, 28, 31 and 32). Both analyses revealed that several items were items were either extremely easy or moderately to significantly difficult. The first 11 items were all answered correctly by over 90% of participants, before a significant drop off with 17 items answered correctly less than 60% of the time. This suggests that the item design may not align well with theory or increase in difficulty in a predictable manner.

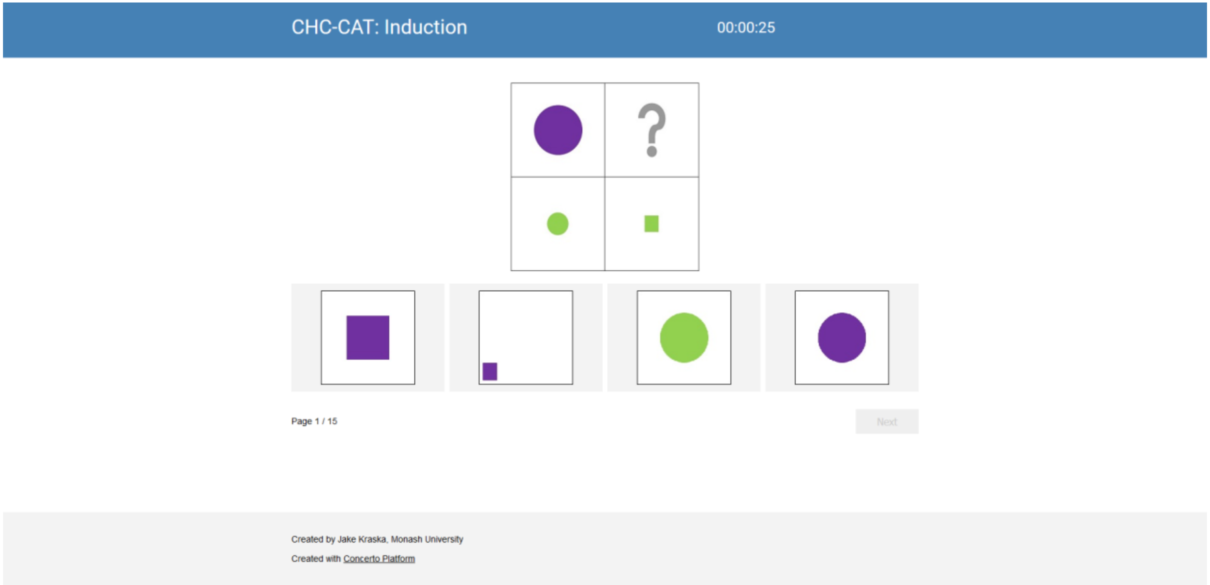
Problematically, the unbalanced execution of matrix size, time allowances and number of response options complicated the analysis of the Induction items from a CTT perspective. The Induction item set for the ITOS had the lowest reliability result of the four CHC abilities of interest (although this may reflect a wider difficulty range). To further evaluate the impact of these variables on the difficulty of items, a high number of additional items needed to be developed. Forty-seven new items were developed for the ICS. Table 4-7 show the number of items per variable under consideration demonstrating a more balanced implementation of these variables compared to Table 3-4 and Table 3-5.

Table 4-7. Gf:I Item Classification for ICS

	4 Options		**5 Options**		**6 Options**	
Matrix Size	**30s**	**60s**	**30s**	**60s**	**30s**	**60s**
5 x 1 Matrix	4	2	2	3	3	2
2 x 2 Matrix	7	2	3	2	3	2
4x1 Matrix	3	2	3	2	3	2
3x3 Matrix	3	2	1	4	2	3
4x2 Matrix	3	2	2	3	3	2

All Induction items (original and newly developed) were reviewed by the same panel of experts as per the review of Lexical Knowledge items. Items 22, 24, 25, 30, 31 and 33 (all from the original item set) were rated as concerning items due to perceived poorly executed rules. While these were included in the ICS data collection, their impact on the subsequent analyses was monitored. An example of Induction subtest is displayed in Figure 4-7.

Figure 4-7. Example of Gf:I ICS Item



Created by Jake Kraska, Monash University
Created with Concerto Platform

4.3.2.4 Visualisation. Visualisation (Gv:Vz) is a major narrow ability within the Visual-Spatial processing broad ability and is considered to represent the ability to perceive, manipulate, and mentally simulate the transformation of complex visual patterns and shapes (Schneider & McGrew, 2018). Fifty-two items were developed for the ITOS. After tests for psychometric robustness and Rasch model analysis, Heng (2018) retained 17 items while the ITOS analysis retained 16 items. There was a high amount of overlap between the items retained in both analyses. In the ITOS analysis, the items retained were a good fit to the Rasch model, met the assumptions of monotonicity and unidimensionality. However, consistent with the other CHC abilities in the ITOS, the items were low in difficulty and information, requiring administration of the whole set to obtain a marginal reliability of .71.

Thirty-five items were removed due to various violations of IRT assumptions. Sixteen items were removed due to violation of unidimensionality and monotonicity, likely due to the ease of their completion and possibility of testing Speeded Rotation rather than Visualisation. As all items had the same time limit, it is likely earlier items were easily answered within the time limit whereas later items may have discriminated better between adults with good and bad visualisation, consistent with the (Carroll, 1993) discussion of Speeded Rotation versus Visualisation. Burton and Fogarty (2003) conducted CFA studies of Visual Processing subtests and found a correlation of .77 between Visualisation and Speeded Rotation tasks (the highest factor correlation in their study) showing the difficulty in separating these narrow aspects of Visual Processing. Eighteen items were removed due to poor fit with the Rasch model, and two items were removed due to local dependency, however qualitative analysis of these items did not identify the cause of these violations. As the next stage of the research in this thesis focused on recruitment of *both* children and adults, as well as engage in non-random offline testing, there was more likely to be more variability in the performance of participants on the easier items. Therefore, for the ICS, all items removed in the ITOS analysis were retained for further analysis.

While all items were retained, item difficulty did not seem to follow a pattern consistent with previous research (i.e., increases in difficulty related to internal cues or number of pieces requiring rotation). The addition of items was anticipated to increase the available range of θ being tested, as well as provide additional items in which reliability can be increased. It is possible the number and similarity of puzzle pieces resulted in changes in difficulty and was therefore the focus of additional item development. An additional set of items that required the rotation of four pieces, as well as a set of items (with two, three and four pieces) that included highly similar pieces, were added (Table 4-8). All items had a time limit of 30 seconds.

Following the addition of items, minor changes to the training items were required. Firstly, training item 3 showed the examinee they would be required to mentally rotate puzzle pieces, however the instruction in the original training GIF (Graphic Interchange Format; a lossless format for image files that supports both animated and static images) was simply “rotate” (Figure 4-8); further clarification that this rotation was required mentally was achieved by changing the text to “rotate in your mind” (Figure 4-9).

Training item 4 aimed to show that examinees could rotate but not flip pieces. In the original version, one puzzle piece had the text “cannot flip” (Figure 4-10) before an attempt to place the pieces together (Figure 4-11). To clarify this process, the new sequence of images included a step between; the sequence showed the examinee they could still rotate the piece in their mind (Figure 4-12), and putting them together at the end showed they are different (Figure 4-13).

All visualisation items (original and newly developed) were reviewed by the same panel of experts called upon throughout this stage of the study. No items were identified as concerns for administration in a screening tool as a measure of an individual’s ability to mentally rotate shapes. The Visualisation subtest as displayed in the CHC-CAT is displayed in Figure 4-14.

Table 4-8. Number of New Gv:Vz Items for ICS

# Pieces	Internal Cues Included		Internal Clues Excluded	
	New Items	ICS	New Items	ICS
2	1	13	1	14
3	1	16	1	13
4	7	7	7	7
5	2	2	0	0
Total	11	38	9	34

Figure 4-8. Training item 3 frame 5

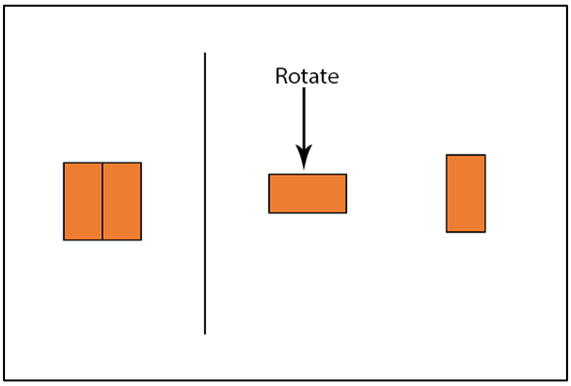


Figure 4-9. New training item 3 frame 5

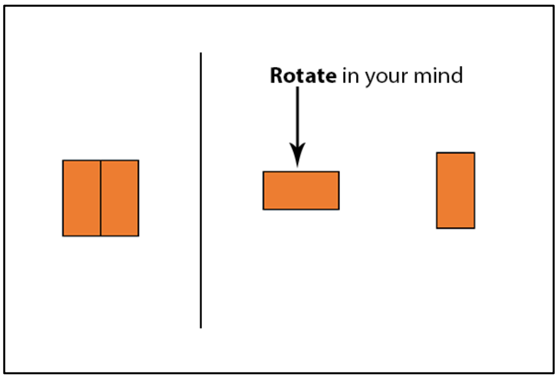


Figure 4-10. Training item 4 frame 5

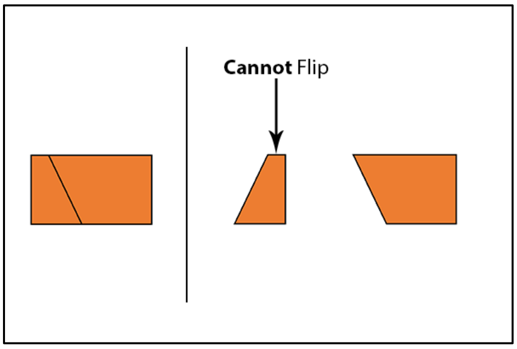


Figure 4-11. Training item 4 frame 6

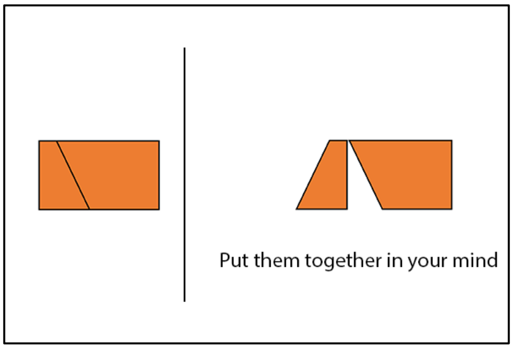


Figure 4-12. Training Item 4 new frame

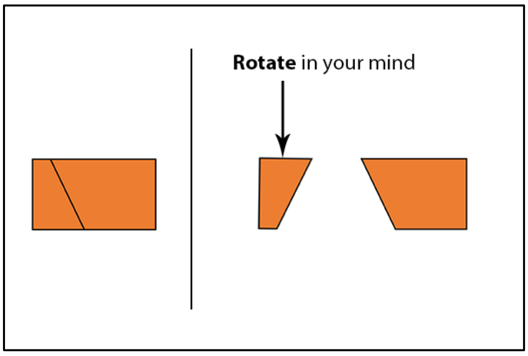


Figure 4-13. Training Item 4 Conclusion

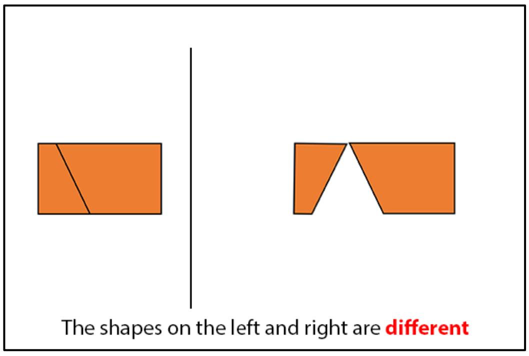
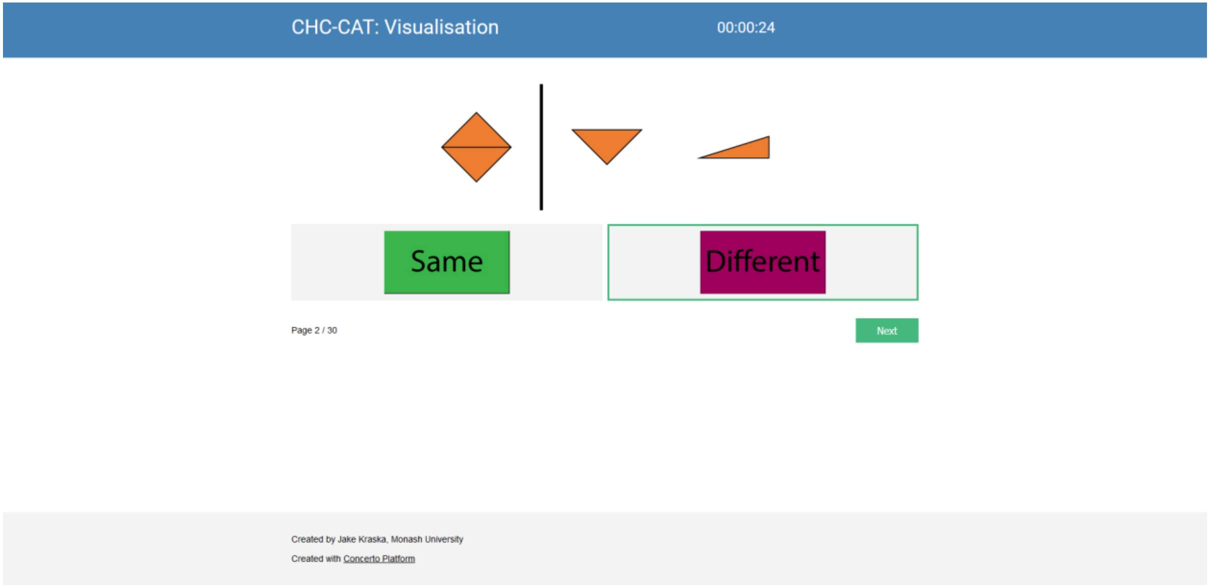


Figure 4-14. Example of Gv:Vz Item



Created by Jake Kraska, Monash University

Created with Concerto Platform

Figure 4-15. Example of original Gwm:Wc stimuli sequence

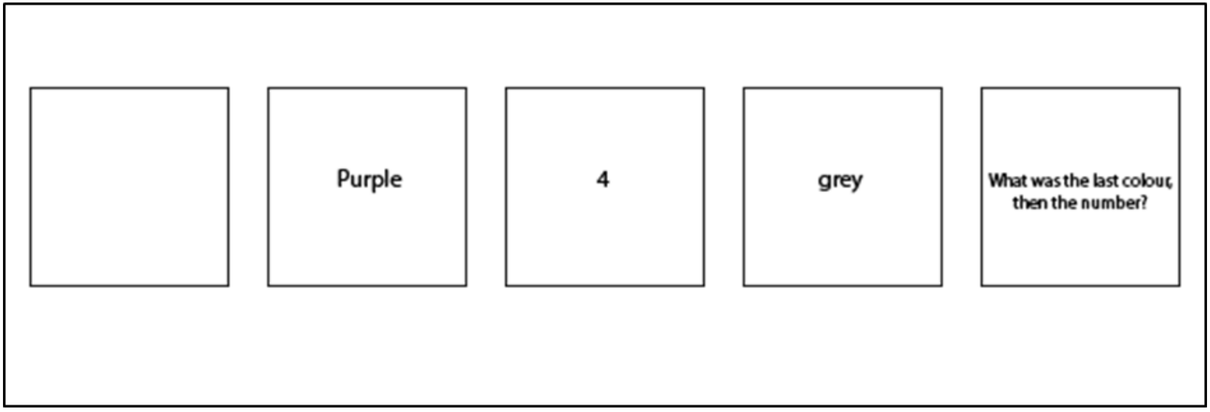
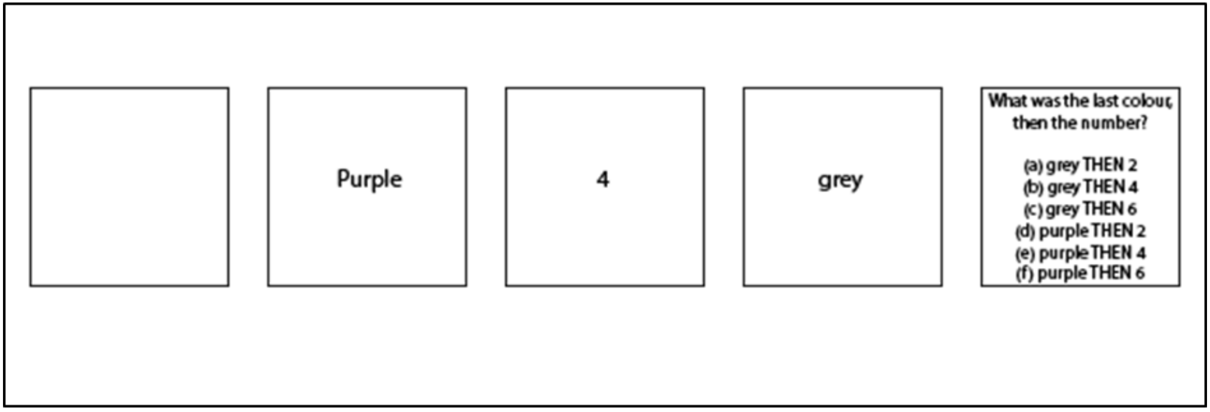


Figure 4-16. Example of altered Gwm:Wc stimuli sequence



Note. The figures above demonstrate the change between the original ITOS Gwm:Wc stimuli and the ICS stimuli. The goal was to remove the response options from being available during item administration.

4.3.2.5 Working Memory Working Memory (Gwm:Wc) is defined as the ability to manipulate information in primary memory and is a combination of attentional control with either visual or auditory short-term storage (Schneider & McGrew, 2018). Kennedy (2018) developed 38 items that measured Working Memory. Kennedy retained 13 items while the ITOS analysis retained 19 items, with significant overlap across the two analyses. The 19-item scale measured from extremely easy (-4.8 θ) to moderately difficult (1.8 θ). Items were removed due to being identified on a different Mokken scale or unscalable (i.e., did not progress in difficulty in an orderly way). Some items were removed due to local dependency, largely in the first half of the test, likely due low variability in the performance of the participants. Also, the ability to recall any amount of information is dependent on being able to retain a smaller amount of information – given that working memory is a complex theoretical construct, there is likely to be some innate interdependence between probability of success on harder items and on easier items. Over 90% of participants obtained the correct answer on the first 16 items (except for item 9). Consideration of item design factors appeared important following the ITOS.

Items were designed so the participant would not know in advance what information that they would be required to recall. Importantly, this item design balanced deliberate processing demands and storage demands. The items developed for the ITOS tap into the Visual Spatial Short-Term Storage (G*wm*:Wv) and Attentional Control (G*wm*:AC) narrow abilities, and thus are a measure of Working Memory (Gwm:Wc). Schneider and McGrew (2018) argue "for understanding academic problems, working memory capacity tests that require simultaneous storage and processing are most important". While a test that also tapped into Auditory Short-Term

Storage (G*wm*:Wa) would be preferred, each participants' audio setup could not be anticipated.

One problem with the original implementation of Working Memory items was the response options (i.e., "(a) grey THEN 2; (*b*) grey THEN 4" etc) were viewable during item administration. That is, the participant could see six response options below the GIF that played the stimuli sequence. To improve this for the next phase of testing only options 'a' to 'f' would be displayed below the item, and the response options would be included in the last frame of the sequence (Figure 4-15 and Figure 4-16).

In addition to changes regarding the presentation of response options, changes were made to the method of presenting each frame of the stimuli. In the ITOS, items were presented using GIFs that started with a play button. The length of each GIF was longer than that of the item so the GIF would not loop prior to moving to the next item. As all Working Memory ITOS data was discarded for the ICS analyses, the opportunity was taken to develop a more reliable methodology for a range of devices. JavaScript was used to present each frame as an individual image, with the final frame remaining regardless of how long the item was available. This meant frames could be implemented in the more versatile .jpg or .png format. If any future changes were required as to stimuli presentation, they could be easily implemented without manipulating GIFs.

Standardised changes were also made to each of the items to accommodate these formatting changes. To do this the question was moved to 0.5 inches from the top of the frame, and each question was aligned by horizontal centres with the canvas (whereas they were previously aligned by both horizontal and vertical centres). All colours within the slides were standardised as per Table 4-9. Any future item development within the CHC-CAT for Working Memory items should adhere to these standards.

Code Snippet 4-1. JavaScript Code for Working Memory Item Presentation

```
var currentSlideIndex = 0;
clearTimeout(showNextSlideTimeout);
var showNextSlideTimeout;
var showNextSlide = function() {
 var slideDivs = $('.gwmImages > div');
 currentSlideIndex++;
 slideDivs.eq(currentSlideIndex - 1).hide();
 slideDivs.eq(currentSlideIndex).show();
 if(currentSlideIndex < slideDivs.length - 1) {
  showNextSlideTimeout = setTimeout(showNextSlide, 1500);
 }
 var slideDivs = $('.gvtImages > div');
 currentSlideIndex++;
 slideDivs.eq(currentSlideIndex - 1).hide();
 slideDivs.eq(currentSlideIndex).show();
 if(currentSlideIndex < slideDivs.length - 1) {
  showNextSlideTimeout = setTimeout(showNextSlide, 3000);
 }
};

var documentReadyHandler = function() {
 var slideDivs = $('.gwmImages > div');
 slideDivs.hide();
 slideDivs.eq(0).show();
 slideDivs.eq(0).on("click", showNextSlide);
};
$(documentReadyHandler);
```

Note. The code above uses JavaScript to retrieve the image frames for the Gwm:Wc items then present it in a sequential and well-timed manner. This code is added to the Concerto platform so that it is utilised during each iteration of the CHC-CAT.

Table 4-9. Colour Standardisation for ICS

Colour	Hex Code	Colour	Hex Code	Colour	Hex Code
Grey	#B7B7B7	Pink	#ff00ff	Brown	#7d4900
Purple	#440e62	Green	#00ff00	Yellow	#fff200
Red	#ff0000	Blue	#2e3192	Orange	#f7941d

Analysis of the deleted versus retained items revealed no commonalities. This made deciding which items to retain difficult. Given the identified problems with item presentation as well as the high performance of participants, recruiting a wider age range of participants was warranted before permanently removing more items.

Ultimately it was decided to retain items 11 onwards, with the addition of 6 new items. The first 10 items were deemed too easy. Due to 19.95% of participants still obtaining a correct score for item 37, 6 new items were added. Item 37 contains 11 'chunks' and was the most difficult. While this may turn out differently in the next study phase due to changes in item presentation, adding items to the extreme range of difficulty would be helpful for the purposes of trialling them. Given that items 37 and 38 included 11 chunks of information, the new items would progress from that point (i.e. two items with 12 chunks, two items with 13 chunks, and two items with 14 chunks).

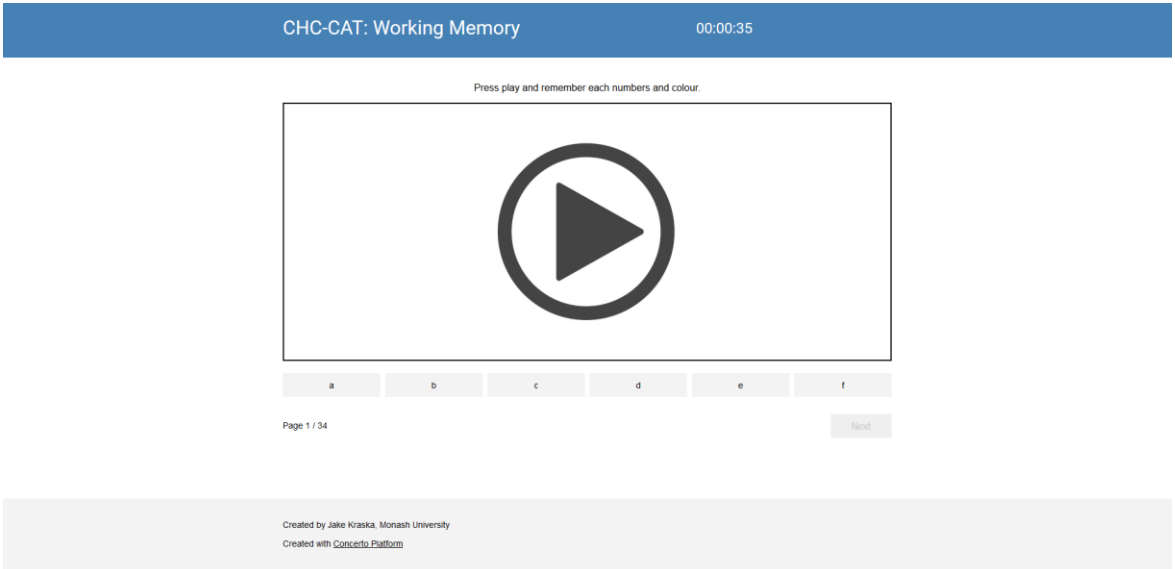
The same time limit of 40 seconds remained for the whole item. This ensured the testing was not overly onerous on participants' time. The final item (item 44) showed the last frame within 18 seconds, allowing 22 seconds to decide. This timing was informed by previous research suggesting that information degrades from memory within 30 seconds (Atkinson & Shiffrin, 1971; Peterson & Peterson, 1959; Revlin, 2012).

All Working Memory items (original and new) were reviewed by the chosen panel of experts. No items were identified as concerns for administration.

Unfortunately, due to the planned changes to include the response options at the end of the stimuli sequence, the existing ITOS data were only included in the exploration phase and not the multivariate imputation (MI), CFA, or IRT analyses.

Examples of the Working Memory subtest are displayed in Figure 4-17, Figure 4-18 and Figure 4-19.

Figure 4-17. Example of Gwm:Wc ICS Start



CHC-CAT: Working Memory 00:00:35

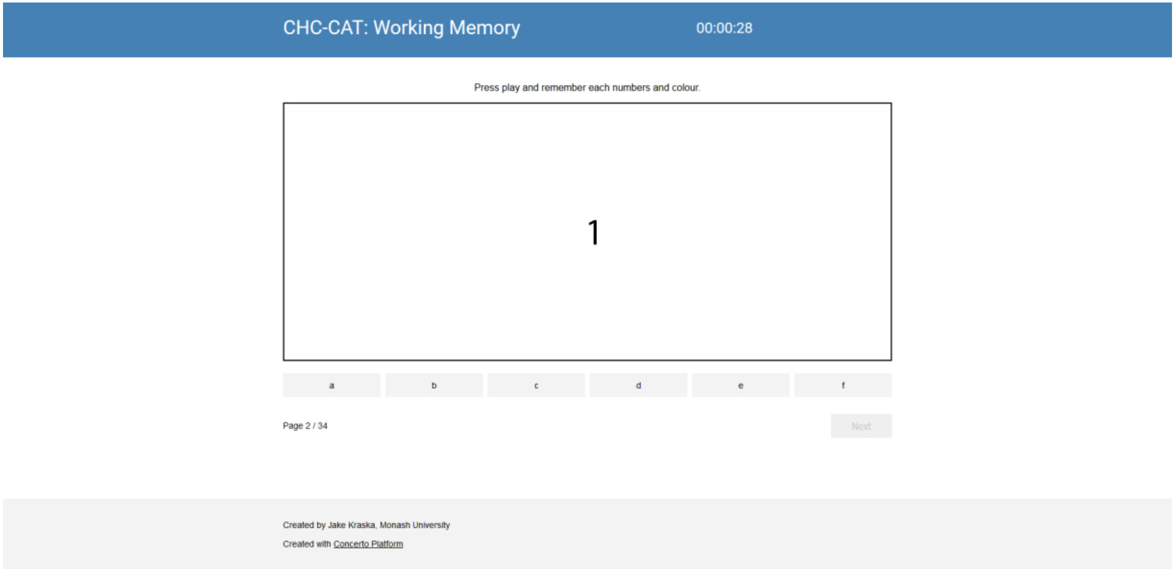
Press play and remember each numbers and colour.

Page 1 / 34

Created by Jake Kraska, Monash University

Created with Concerto Platform

Figure 4-18. Example of Gwm:Wc ICS Stimulus



CHC-CAT: Working Memory 00:00:28

Press play and remember each numbers and colour.

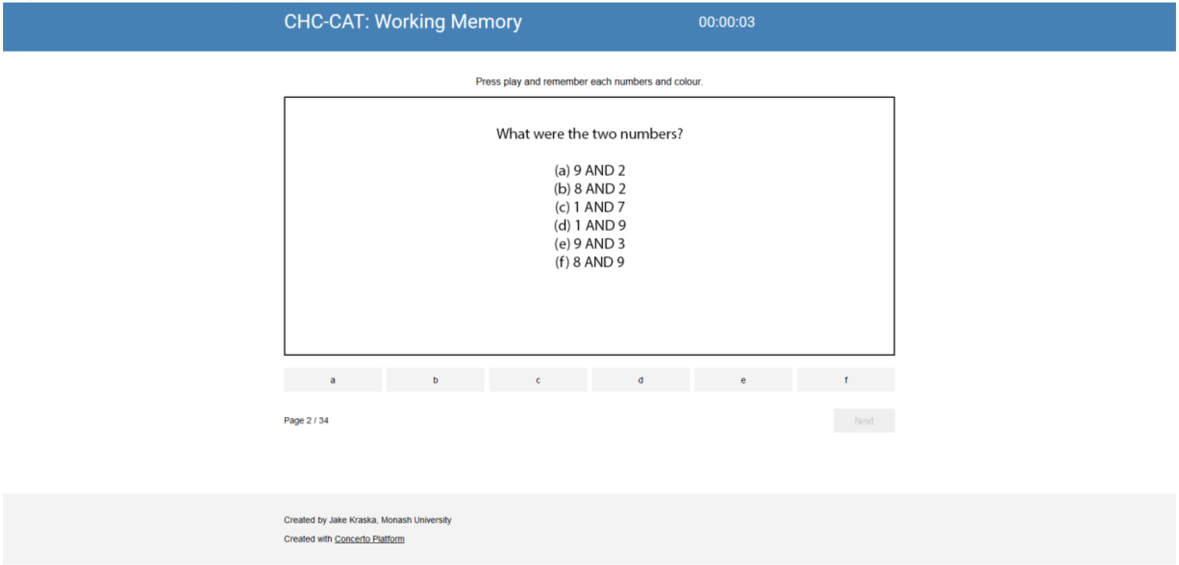
1

Page 2 / 34

Created by Jake Kraska, Monash University

Created with Concerto Platform

Figure 4-19. Example of Gwm:Wc Final Frame



CHC-CAT: Working Memory 00:00:03

Press play and remember each numbers and colour.

What were the two numbers?

(a) 9 AND 2
(b) 8 AND 2
(c) 1 AND 7
(d) 1 AND 9
(e) 9 AND 3
(f) 8 AND 9

Page 2 / 34

Created by Jake Kraska, Monash University

Created with Concerto Platform

4.3.2.6 Item Groups. To reduce the time requirements for participants items were ordered based on item sets. While the ITOS showed items did not increase in difficulty in a strictly linear fashion based on item number (as it is challenging to predict the psychometric difficulty of an item during development), it was generally shown that higher numbered items were more difficult than lower numbered items. As new items were developed for the second phase of data collection, it was important to have 'anchor' or 'linking' items (see: Meade & Wright, 2012) to place the new items and old items from each item set on the same latent trait scale in the IRT analysis. For the Confirmatory Factor Analysis (CFA) (and underlying missing data analysis yet to be conducted) the anchor items were known as the 'X set' which has shown importance regarding the efficiency of a multiform design missing data analysis (Rhemtulla & Hancock, 2016). It was also recognised in the ITOS analysis that the original items were too easy for adults and thus more difficult items were required, but that the original items were more appropriate for children and adolescents. Table 4-10 displays the breakdown of each subtest into item sets, and the next section describes the order of these item sets based on test time limit considerations and age of participants.

Table 4-10. Item Sets and Linking Items for ICS Data Collection

Subtest	ITOS	ICS Item Set A	ICS Anchor Items	ICS Item Set B
G*c:*VL	1-55	4-19; 21-34	36-46; 48-51; 53-55	56-69;62-67; 69-107
G*f:*I[1]	1-33	30 seconds: 1-10	30 seconds: 11-15	30 seconds: 37; 40-42;45-49;52-54;57-59;62-64;67-69;72-78
		60 seconds: 16-28	60 seconds: 29-33	60 seconds: 34;38-39;43-44;50-51;55-56;60-61;65-66;70-71;79-80
G*v:*Vz	1-52	1-21	22-52	53-72
G*wm:*Wc	1-38	11-21	22-32	33-44

Note. Missing items are a result of post-ITOS or pre-ICS item deletion as described above. [1]Induction items were administered in 6 sets, starting with 3 30 second sets followed by 3 60 second sets.

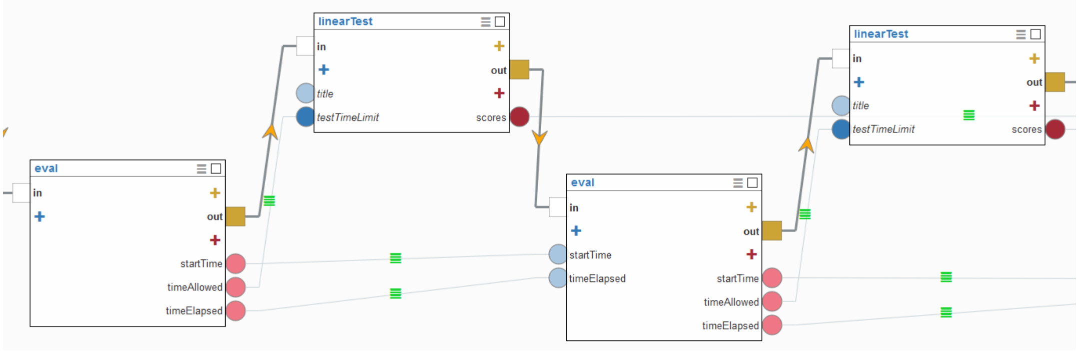
4.3.2.7 Test Time Limits. Due to the increased number of items for each CHC factor, test time limits were implemented due to ethical considerations and reducing the number of items administered to participants to avoid disengagement. Integration of these time limits aimed to reduce the impact of missing data based on age groups, however, was not always successful (discussed later). Attempts were also made to set these time limits based on research regarding attention spans. Memory and attention span generally develop and degrade as a function of age (De Luca et al., 2003; Gomes et al., 2000; Klenberg et al., 2001), although the speed and continuity of this development is uncertain (Gómez-Pérez & Ostrosky-Solís, 2006).There is ongoing debate about the potential impact of a wide range of factors on the development of such executive functions (Lillard & Peterson, 2011). Some preliminary research suggested that by four years old children are capable of sustained attention for at least three minutes (Ruff & Lawson, 1990). Other early authors argued that two-year-old children can sustain attention for about five minutes, while older children and adolescents can sustain attention for up to 20 minutes (Cornish & Dukette, 2009). However, recent contentions suggest that perpetuations of a “10- to 15-minute limit” do not account for individual differences, are based on outdated research and are dependent on the type of task under consideration (Wilson & Korn, 2007). Other research suggests attention tends to fluctuate throughout tasks and “active learning moments” can increase the attentional system (Hlas et al., 2017). Time limit decisions are consequently difficult to justify.

Therefore, rather than implementing a single time limit for all adult participants, three options were offered to participants for each CHC factor subtest that they attempted: “10-minute time limit”, “15-minute time limit”, “I would like to try all items available” (i.e., no time limit). It is believed that this would account for individual differences in attention spans as well as allow online participants to select the amount of time they were available to participate in the study. For the school aged sample, due to ethical considerations about total testing time (discussed below), all participants were asked to complete three CHC factor subtests with a time limit of 15 minutes each under the direct supervision of provisional registered or general registered psychologists, resulting in a maximum test time of 45 minutes – much less than a face-to-face comprehensive cognitive ability which may take 60 to 90 minutes.

For participants under the age of 18, items were administered in the order of anchor, Item Set A, then Item Set B, with as many items completed as possible within 15 minutes. Total test time limits were put in place due to ethical considerations of working with young people within the school environment; there was a need to reduce the amount of time children were outside of class. There were also concerns given the uncertainty about the difficulty of items, how challenging some children may find the items as adaptive functionality had not been put in place yet there were concerns that without time limits children with low abilities may have been required to undergo testing for significant periods of time to finish item sets.

Version 5.0.0 of the Concerto Platform was utilised for this data collection. This version did not yet have the functional ability to output total time taken from each individual ‘node’, therefore evaluation nodes were added prior to, and after, each individual subcomponent of a test (Figure 4-20). Each evaluation node included custom R code to ensure total test time limits worked correctly (Code Snippet 4-2).

Figure 4-20. Example of Node Setup to Limit Overall Test Time



Note. Each "eval" node includes the R code used to execute the code below, passing the time allowances between each eval node.

Code Snippet 4-2. R Code for Test Time Calculation

```
timeAllowed <- 900
if (!exists("startTime")) {
 startTime <- Sys.time()
 timeElapsed <- 0
} else {
 if (!exists("timeElapsed")) {
  timeElapsed <- as.numeric(startTime - Sys.time())
 } else {
  timeElapsed <- as.numeric(startTime - Sys.time()) + timeElapsed
 }
 startTime <- Sys.time()
}
if (abs(timeElapsed) > timeAllowed) {
 timeAllowed <- toString(30)
} else {
 timeAllowed <- toString(timeAllowed + timeElapsed)
}
```

Note. The code above uses R to calculate the remaining time allowed for each set of G*f*:I items to ensure that total test time limits are not exceeded.

For people over the age of 18, items were administered in the order of anchor items, Item Set B then Item Set A. Due to the online self-selection recruitment methodology, and not needing the same ethical considerations as for those under the age of 18, people were offered the opportunity to place their own time limits or to have no time limit at all. As this data collection was conducted after the collection of school aged data, an updated version of Concerto (v.5.0.9) was used. This version had integrated 'test time' functionality where previous versions only included 'item time' function meaning that solutions such as Figure 4-20 were not necessary. Test design decisions such as time limits and item set predictably resulted in missing data.

4.3.3 Procedure

Prior to recruitment and administration of the test to participants, items were reviewed by a panel. The panel included a special education teacher familiar with CHC, a primary school teacher familiar with screening tools in the classroom and an Educational Psychologist familiar with CHC evaluation. The panel were instructed to identify items that they believed were ambiguous in question or answer, deviated from CHC theory, or did not do implement rules in a logical or consistent fashion. Items that were identified for removal are described above in the Materials section.

For the current study two samples of data were collected: school aged participants and adult participants. The first sample consisted of individuals under the age of 18. These individuals were recruited via snowball sampling and completed both the *Wechsler Intelligence Scale for Children – Fifth Edition* (Wechsler, 2016) and three of the four CHC-CAT ability tests in a linear fashion. This testing was conducted by provisional and general registered psychologists on school grounds with children whose parents contacted the chief investigator upon hearing about the research (i.e., via snowball sampling). Breaks were given between each CHC-CAT subtest but were not given to adults, as adults could self-impose breaks between subtests as required.

Adults were recruited online (as per the Participants section) via both social media advertisements and snowball recruitment. Adults participated at a time of their convenience. For both the school aged and adult samples the participants accessed the Concerto platform via a link. For school aged participants this was completed predominantly on iPads, and for adults on a range of devices (Table 4-11).

The Concerto platform was hosted via a National eResearch Collaboration Tools and Resources (Nectar) cloud allocation (Australian Research Data Commons, 2020). The server utilised *Ubuntu v18.04* on a 'm3.small' allocation and Concerto was installed via the recommended instructions provided by the Concerto developers.

Table 4-11. Devices Used to Access Concerto Platform

Device	ITOS	ICS Adult	ICS School Aged
Personal Computer	0	366	0
Laptop	0	434	10
Samsung Galaxy	0	203	0
Google Pixel	0	34	0
Microsoft Surface	0	9	0
Apple iPad	0	149	131
Apple iPhone	0	547	0
Other Phone	0	130	0
Other Tablet	0	10	0
Prefer Not to Say	0	9	0
Unknown	1376	0	0

4.3.4 Data Analysis

The data analysis for this study can be found at github.com/jakekraska/phd. For the ICS approximately 3,100 lines of R code were written.

4.3.4.1 Software. The analysis for this study was conducted using R Version 3.6.3 (R Core Team, 2020) within the R Studio Integrated Development Environment Version 1.2.5033 (R Studio Team, 2019). Packages used are discussed below.

4.3.4.2 Missing Data. While IRT is relatively robust to missing data, not all other statistical procedures are as flexible. Missing data was analysed using the *mice v. 3.8.0* (Burren & Groothuis-Oudshoorn, 2011), *psych v.1.9.12* (Revelle, 2019), *tibble v.2.1.3*

(Muller & Wickham, 2019), *VIM v.5.1.1* (Kowarik & Templ, 2016), *BaylorEdPsych v.0.5* (Beaujean, 2012) and *reshape2 v.1.43* (Wickham, 2007) packages for *R*.

Understanding and dealing with missing data can be a complex task with many methods available. In the ITOS analysis an assumption was made that missing data was a result of inability to complete items and thus coded as a 0; given the addition of new items in the data set that different participants did not have a chance to attempt, this approach was no longer appropriate. Little and Rubin (2020) define missing data as "unobserved values that would be meaningful for analysis if observed" (p. 4). Some causes of missing data are univariate missing data, item nonresponse, attrition in longitudinal studies, two sets of variables never jointly observed, latent variables that are never observed, and missing data in clinical trials (Little & Rubin, 2020). Traditional methods of addressing missing data such as complete-case, available-case analysis, mean substitution, regression imputation, and last observation carried forward have been found to produce bias (Cole, 2008). Additionally, imputation methods such as mean imputation, regression imputation, stochastic regression imputation, hot deck imputation, substitution, cold deck imputation and composite methods are limited because they tend to underestimate the sampling variance of estimates (Little & Rubin, 2020). To this end, maximum likelihood (ML) and MI are generally the strategies recommended (Baraldi & Enders, 2010).

There are three mechanisms of missing data, discussed in Enders (2010): Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). Data classified as MCAR is data that is missing as a function of randomness and is not a result of any measured variable (Enders, 2010). In the current study this would suggest the probability of a participant missing an is unrelated to the demographic variables *and* all four CHC factors measured. Missing data by design is considered MCAR (Little & Rubin, 2020). It is possible to have a combination of missing data mechanisms.

There are multiple causes of missing data in the current study. Because the data was collected in phases and had time limits, there is missing data by design. Participants in the ITOS data collection were exposed to different items than those in the ICS. Planned missing data designs "are an efficient way to manage cost, improve data quality, and reduce participant fatigue and practice effects" (Rhemtulla & Little, 2012, p. 425).

However, outside of these deliberate research design decisions there is also data missing for other reasons. Firstly, there is missing data that is likely a function of the variables under investigation themselves (i.e. age, Lexical Knowledge, Induction, Visualisation and Working Memory). Intelligence is intrinsically associated with age; differences in age would likely lead to experiences of varying difficulty with the items, and thus decisions whether to proceed with or cease testing. Similarly, differences in the participants abilities themselves may have also resulted in similar behaviours. Some missing data may also be a result of variables that are not under investigation (e.g. socio-economic status, school exposure, etc.) and could not be measured due to ethical considerations or attempting to reduce the task demands of participants. Establishing the existence of such relationships can be difficult because the variable is not included in the measurement model, and as a result data are MCAR (missingness is unrelated to all variables under investigation), MAR (missingness is related to a measured variable other than the variable missing) *and* MNAR (missingness is related to the variable itself).

When looking at missing data in the entire data set rather than by each individual phase of data collection (Table 4-12), there are inflated proportions of missing data that were addressed at a phase and item set level. As the current study considers each CHC narrow ability as a unidimensional variable, exploration of the missing data patterns in the current chapter was completed on a narrow ability by narrow ability basis. The current subsection will include a consideration of methods to address missing data for the purposes of the IRT, Mokken, CFA and reliability analyses.

Missing data in the current study is predominantly from the ICS Adult sample, which is unexpected in relation to response rates from the ITOS, but not unexpected given the increase in number of items. The missing data is a result of participants either reaching their self-set time limit or dropping out of the study early. This was reflected in higher proportions of incomplete data sets for each CHC factors Item Set B relative to the number incomplete data sets for each CHC factors Item Set A and anchor items. Table 4-13, Figure 4-21, and Figure 4-22 summarise these findings. In a similar vein to the ‘Three-Form Design’ (Enders, 2010), anchor items within Lexical Knowledge were used as the core set of items across all three phases of sampling.

In the past missing data required deep analysis and pattern identification, ML (such as FIML) and MI “are well suited for virtually any missing data pattern” (Enders, 2010, p. 5). MI has been used for data sets that have thousands of participants with hundreds of variables (Schenker et al., 2006) and those that included missingness rates of approximately 66% (He et al., 2010). MI methods make use of a algorithms and are more flexible than FIML. Using MI generates *m* imputed data sets (typically 3 to 5; Figure 4-23), whereas FIML does not actually fill in the data sets (Graham, 2009); *m* data sets result in *m* estimates for each parameter that can be pooled together (Rubin, 1987). Additionally, MI methods allow the use of auxiliary variables in the model which may not require imputation themselves. These specific characteristics of MI suggest it is an appropriate method of addressing missing data in the current set.

Table 4-12. Percentage of Missing Data by CHC Ability

CHC Ability	*n Items*	Mean	SD	Median	Min	Max
Lexical Knowledge	105	48.15	16.84	55.02	17.22	71.21
Induction	78	60.55	20.29	63.29	17.27	89.78
Visualisation	76	39.83	24.98	28.64	13.83	80.15
Working Memory	44	36.91	24.84	33.71	6.17	90.48

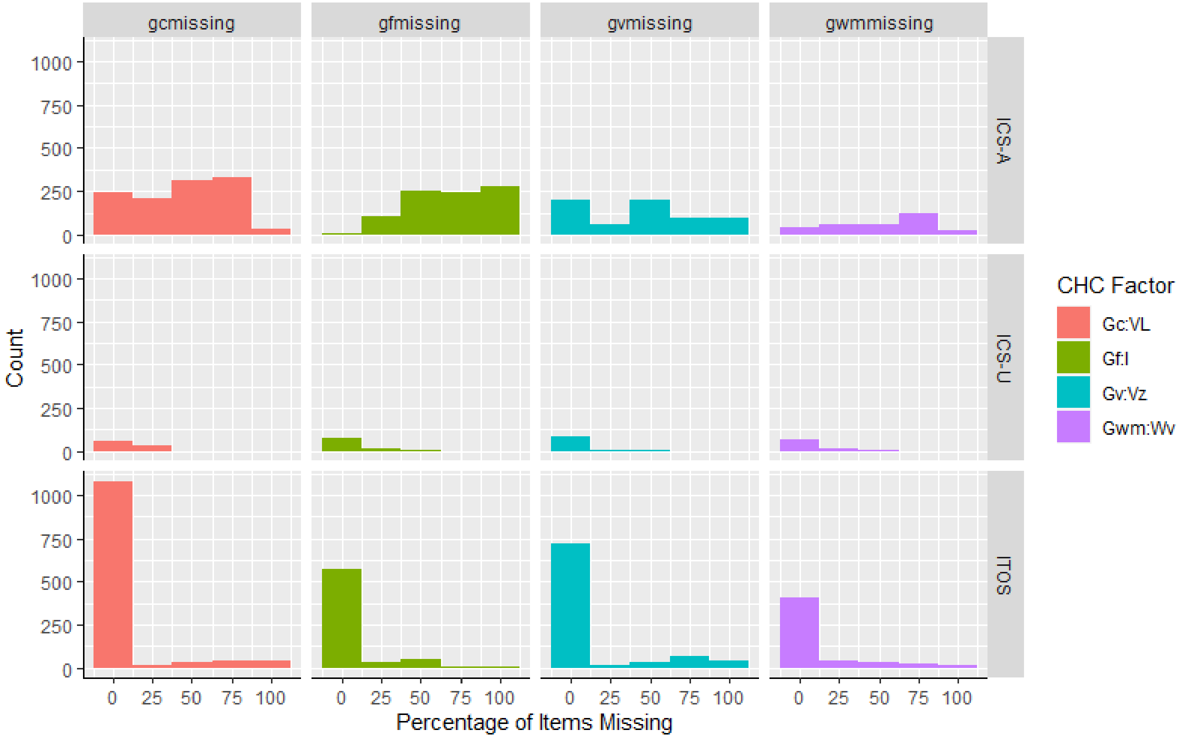
Note. This is the percentage of missing data by participant, for example there was a participant with 90.48% of items unanswered in the Working Memory data set.

Table 4-13. Percentage of Complete Participant Data

Subtest	Item Set A	Anchor Items	Item Set B
Gc:VL	37.94	48.69	11.07
Gf:I	30s: 26.54; 60s: 20.48	30s: 37.97; 60s: 22.35	30s: 7.31; 60s: 4.02
Gv:Vz	28.38	32.26	9.09
Gwm:Wc	20.51	17.10	2.48

Note. This is the percentage of participants that have completed every item within a set of items.

Figure 4-21. Number of Participants with Missing Data (By Phase)



Note. The ICS-A showed the most variation in amount of missing data. This may be related to the recruitment of more young adults or due to changes in the item sets.

Figure 4-22. Percentage of Missing Items (Total Sample)

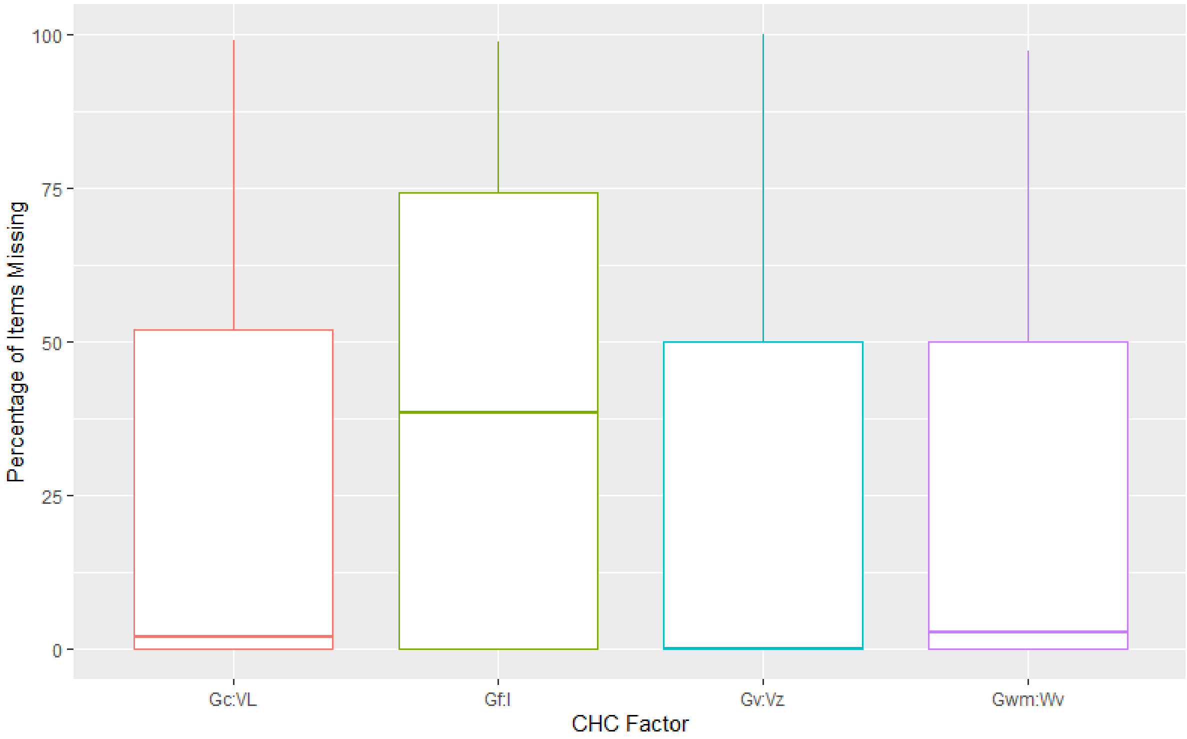
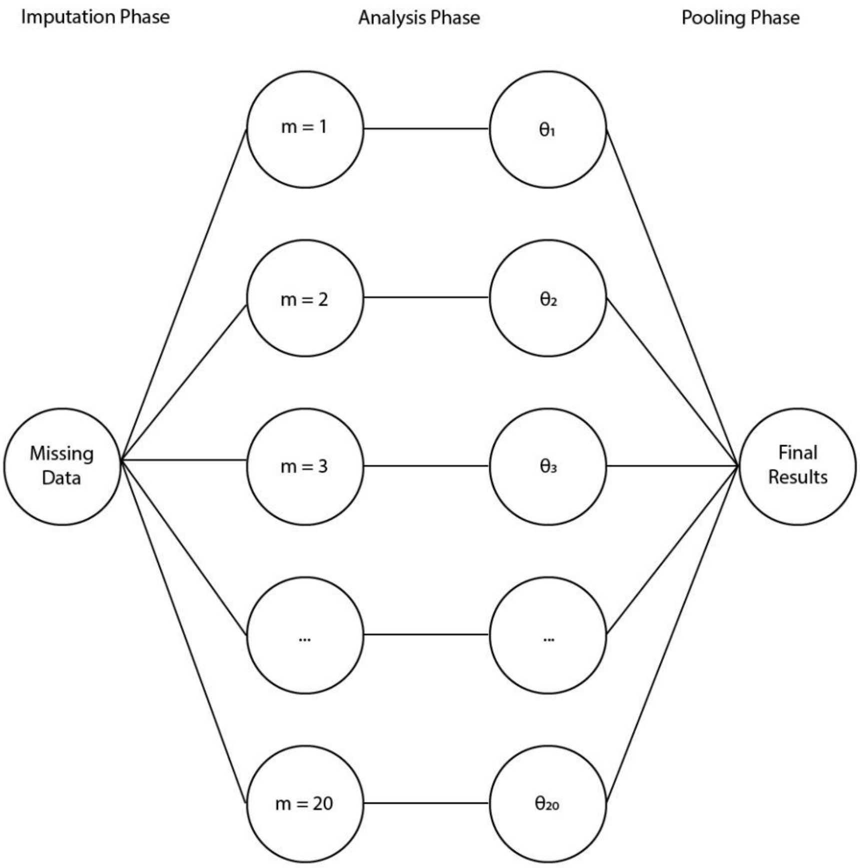


Figure 4-23. MICE Analysis Phases



Multivariate imputation by chained equations (MICE) is a flexible method of MI, including for both continuous and binary data (Azur et al., 2011). This method allows use of different algorithms dependent on each variable being included in the imputation model. Earlier discussion regarding the types of missingness in the data suggested it was likely that missing data was MCAR, MAR *and* MNAR. Given the patterns of missingness relate to the order of item sets, it is reasonable to assume that missing data is a result of planned missingness. Even if the data was not at least 100% MAR, MI has been shown to be robust for non-random missingness (Rässler & Riphahn, 2006).

MICE was completed with all participants without a full set of anchor item data removed. Data was prepared for MICE via several methods to ensure specific items or participants were not having an undue impact on the imputation. Participants who did not have a full set of anchor items (whether correct or incorrect) were removed. Items removed after the ITOS were deleted from the data set. Participants with high amounts of missing data were removed to improve the accurate imputation for other missing data. Research has shown that MICE is effective with even high rates of missingness, so the maximum amount of missingness allowed was set to 33%. Univariate outliers were removed by time and score. Multivariate outliers were removed by analysing the influence of each variable on total score and removing those cases that had a Cook's distance greater than four times the mean. Finally, individuals without a known age were removed as it is believed that age is an important variable in the prediction of performance on individual items.

The *mice* package in *R* allows a variety of imputation methods to create an independent model for each incomplete variable. The Random Forest algorithm was used for the imputations as it is known to produce more efficient imputations and narrower confidence intervals (Shah et al., 2014). The Random Forest algorithm is based on the concept of decision trees; at the most basic level, error is reduced as more "forests" are generated (Breiman, 2001). All items (both score and time taken) in the ICS item sets, as well as age, were included as predictors. The visit sequence (i.e. the order in which the items were imputed) was set to 'monotone' (ordered low to high proportion of missing data). A minimum correlation of .20 was required for these to be used in prediction. When conducting MICE (regardless of the algorithm used) one can consider the influx-outflux of each variable in the model, how important each variable is in the imputation prediction, and how reliant each variable is on the imputation prediction. Items with low outflux and high influx have low predictive power and are highly reliant on the imputation model; these items were removed. The number of imputations was set at five; this meant five complete imputed data sets were generated. The packages used in this study to conduct CFA and IRT analyses include functions to analyse imputed data sets. Where there are errors with MICE, individual data sets were analysed in isolation to determine the cause of errors.

4.3.4.3 Reliability. As per 3.3.4.2, the reliability analysis classified .7 or above as good and was completed using the psych package v.1.9.12 (Ravelle, 2019) for R.

4.3.4.4 Rasch analysis. As per 3.3.4.5 the Rasch analysis focuses on difficulty parameters, holding other parameters (i.e., guessing, discrimination) as stable. Fit statistics for this chapter were the same as those for Chapter 3. Individual items were evaluated with the M2 statistic (Maydeu-Olivares & Joe, 2006), Comparative Fit Index (CFI; Bentler, 1990), the Tucker Lewis Index (TLI;Tucker & Lewis, 1973), the Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1992) and the Standardised

Root Mean Square Residual (SRMSR; Hu & Bentler, 1999). Given the reduction in sample size available for analysis due to the data imputation models used, sample size considerations are more important in this ICS than the ITOS; item calibrations should be stable within half a logit at a sample size of 150, and 250 participants is suggested for high stakes testing (Linacre, 1994). Marginal reliability was also evaluated for each item set: this is an estimate of the reliability of the item set based on the standard error of measurement (SEM) of respondents given variable test lengths (Sireci et al., 1991). Unlike in the ITOS analysis, items were not removed based on stringent application of cutoffs. To retain items, qualitative evaluation and item statistics were used to make decisions about removal. These analyses were carried out using the *mirt v.1.3.1* (Chalmers, 2012) and *WrightMap v.1.2.2* (Torres Irribarri & Freund, 2014) packages.

4.3.4.5 Mokken Analysis. As per 3.3.4.4 the Mokken analysis relied on evaluations of Loevinger's *H* value (Loevinger, 1948) carried out using the *mokken v.2.8.11* package (Van der Ark, 2007, 2012) for R. Items that receive a Loevinger *H* value of below .30 are considered to be inaccurate, between .30 and .40 considered to have low accuracy, between .40 and .50 have moderate accuracy, and values over .50 suggest good ordering (Ligtvoet et al., 2010; Sijtsma & Meijer, 1992).

4.3.4.6 Local Independence. As per 3.3.4.6 Yen's Q3 method of correlated residuals (Yen, 1993) was used to test the local independence of items via *base stats v.3.6.3* (R Core Team, 2020). A cutoff of .20 was set for local dependency and flagged items were evaluated further for consideration of removal.

4.3.4.7 Differential Item Functioning. Some changes were made to the differential item functioning (DIF) analysis for this study compared the ITOS (as per section 3.3.4.8). For the ICS, DIF was analysed for age group, gender, nationality, and device used. The *mirt v.1.3.1* (Chalmers, 2012) and *difR v.5.0* (Magis et al., 2010) packages were utilised. ICCs were analysed for items flagged as displaying DIF, and those that had no obvious, theoretical, or conceptual reason for the DIF were retained.

4.3.4.8 Confirmatory Factor Analysis. CFA was completed only as a general indicator of unidimensionality. Given the wide range of item difficulty and ages in the study, and the deliberate cognitive complexity implemented into some item sets, it is expected the fit statistics were unlikely to always meet the required cutoffs. The goal of this study was to calibrate item sets so they have value in measuring a wide range of ability and make theoretical sense in the context of the CHC taxonomy of abilities for use in a CAT (Chapter 5). It is likely a hierarchical or bifactor model would be more appropriate for this data. The CFAs carried out relied on the same fit statistics as per 3.3.4.3 with a recommended 250 to 500 participants per analysis to meet adequate power (Lewis, 2017). Acceptable fit is determined by evaluating these fit indices in cohesion, aiming for greater than .9 for CFI and TLI, an RMSEA below .01 (excellent), .05 (good), or .08 (mediocre), or an SRMR below .08 (good) (Hooper et al., 2008). The Diagonal Weighted Least Squares (DWLS) estimator was used due to the dichotomous nature of item data. The CFAs were carried out using the *lavaan v0.6-3* (Rosseel, 2012) and *semTools v.0.5-2* (Jorgensen et al., 2019) packages for R.

4.3.4.9 Item Analysis and Removal Sequence. Unlike in the ITOS (section 3.3.4.8) items did not have to repeatedly pass the same checks in an iterative process. While the same analyses were carried out if an item was removed at a certain stage the whole analysis process would not be completed again. Furthermore, less stringent application of cutoff rules was used. This was implemented to take a step back from the conservative approach laid out in the ITOS, allow for consideration of theory-based decisions, and place a focus on IRT. This also assisted in preparing for the simulation (Chapter 5) and validation studies (Chapter 6) by retaining as many items as possible.

4.4 Lexical Knowledge Results

4.4.1 Raw Score Outcomes

The descriptive statistics for time taken and total raw score for the ITOS and ICS samples are included in Table 4-14 and Table 4-15.

The maximum possible score for the ITOS was 55, while the maximum score for the ICS Adult and School Aged samples was 98. There was more varied performance for the ICS Adult and ICS School Aged participants than the ITOS participants (Figure 4-24).

There was a significant difference in the total raw score of participants within each phase: *F*(2,2425) = 29.7, *p* < .001 (Figure 4-25). The total amount of time participants were engaged with the Lexical Knowledge test was significantly different: *F*(2,2425) = 158.4, *p* < .001 (Figure 4-26). Time taken was generally similar for the ITOS and ICS Adult samples but was somewhat higher on average for the ICS School Aged sample, likely due to being supervised.

With the addition of items based on less common words, there was a decrease in performance in latter items (Figure 4-27). As per the ITOS analysis, and quite logically, score improved as a function of more time being taken with the test (Figure 4-28).

There was no significant difference between gender identification groups on total score, *F*(3,2424) = 2.25, *p* = .08 (Figure 4-29), or time taken, *F*(3,2424) = 0.422, *p* = .73 (Figure 4-30).

Table 4-14. Descriptive Statistics for Total Score for the Gc:VL Items

	***n* Items**	***n* Participants**	**Mean**	**SD**	**Min**	**Max**
Item Tryout Study[1]	55	1203	46.95	11.72	1	55
ICS Adult	98	1130	41.65	23.28	1	94
ICS School Aged	98	95	50.05	10.50	7	74
All Phases	105	2428	44.61	18.23	1	94

[1]Descriptive statistics varied from ITOS analysis due to changes in missing data methodology

Table 4-15. Descriptive Statistics for Time Taken with the Gc:VL Items

	n	**Mean**	**SD**	**Min**	**Max**
Item Tryout Study[1]	1203	465.67	159.69	11.02	1569.11
ICS Adult	1130	531.53	261.24	15.03	2065.52
ICS School Aged	95	855.77	118.78	229.16	972.11
All Phases	2428	511.59	225.37	11.02	2065.52

[1]Descriptive statistics varied from ITOS analysis due to changes in missing data methodology

Figure 4-24. Gc:VL Frequency of Participant Raw Score

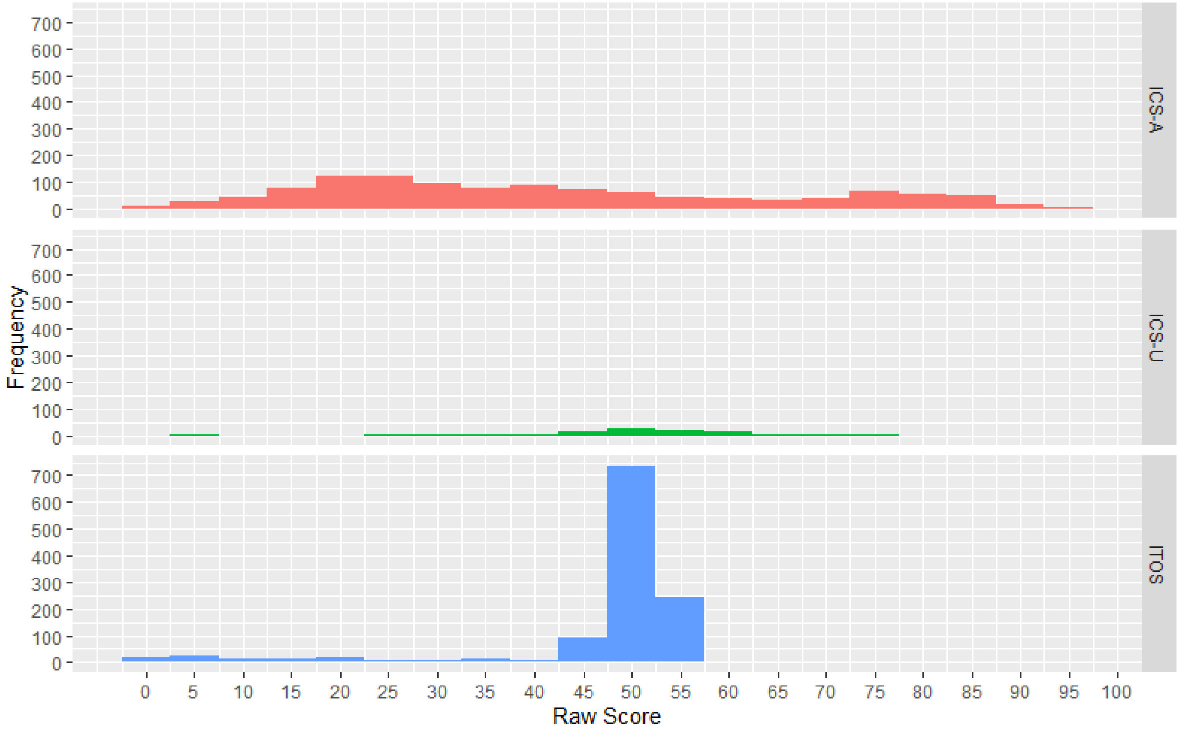


Figure 4-25. Gc:VL Score by Phase

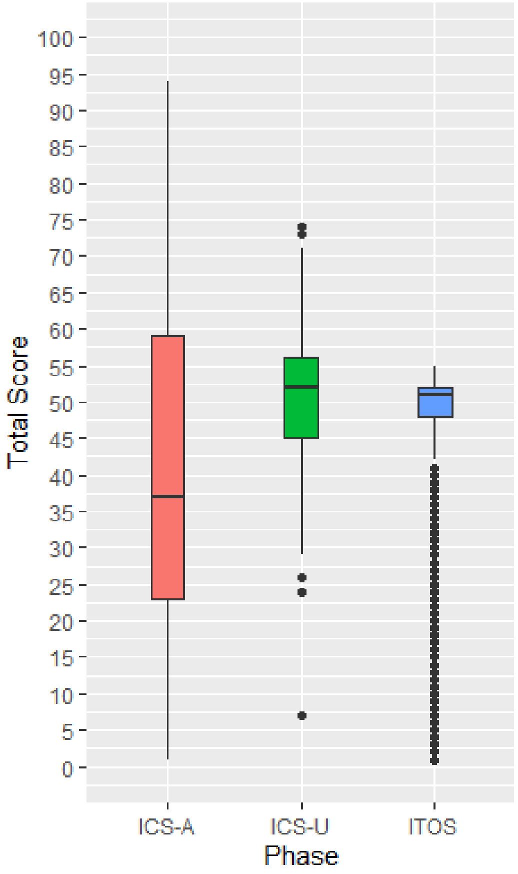


Figure 4-26. Gc:VL Time Taken by Phase

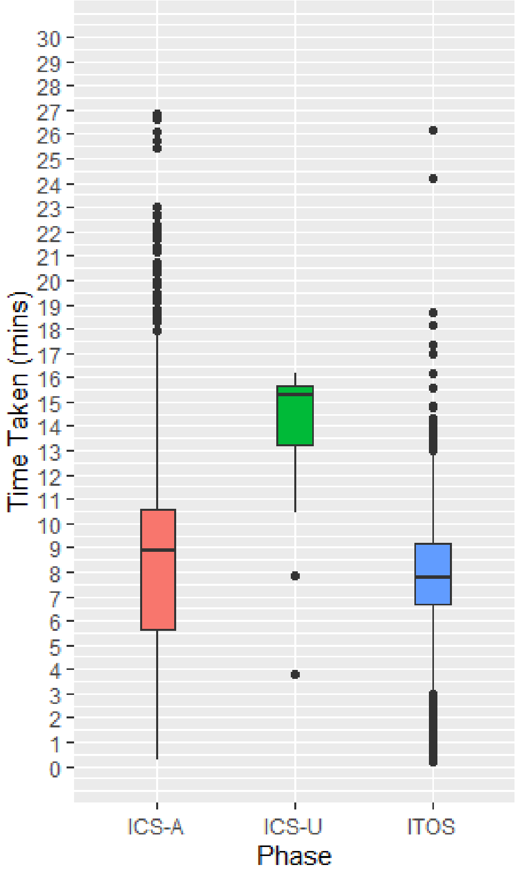


Figure 4-27. Gc:VL Percentage of Items Correct

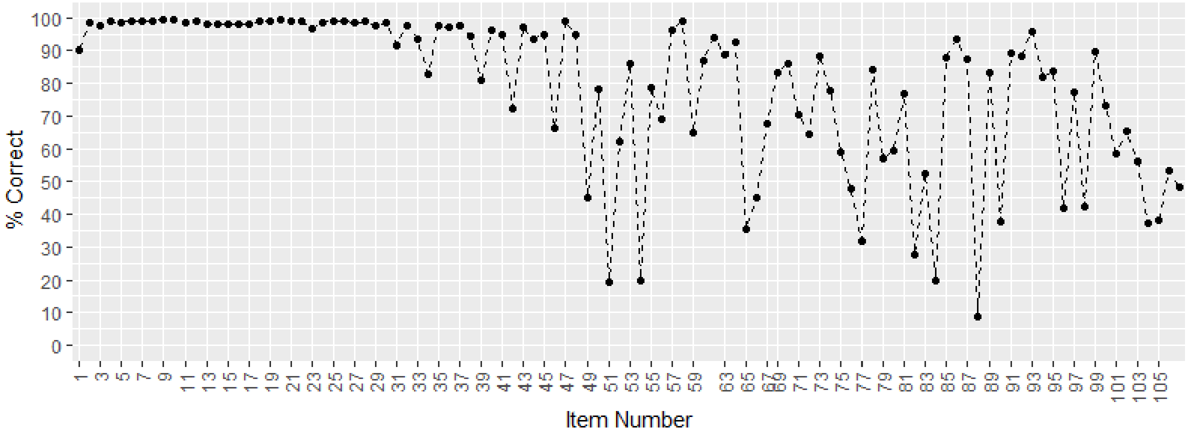


Figure 4-28. Gc:VL Total Score by Time Taken

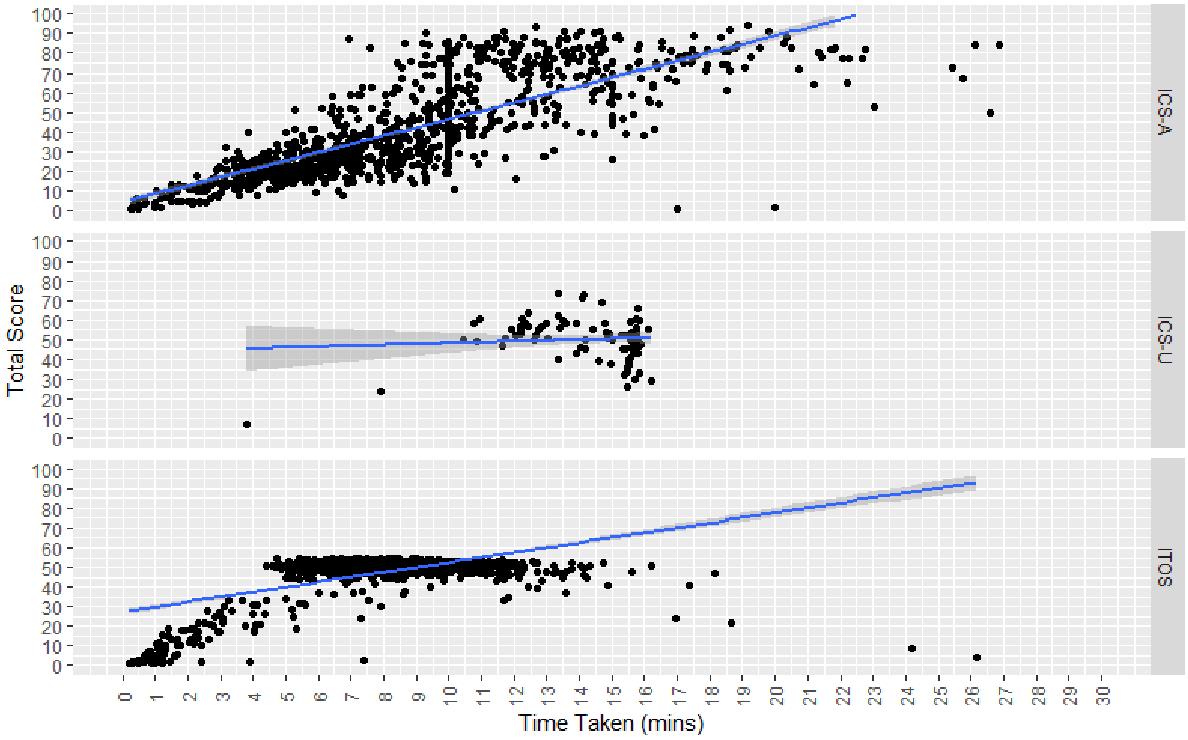


Figure 4-29. Gc:VL Gender by Score

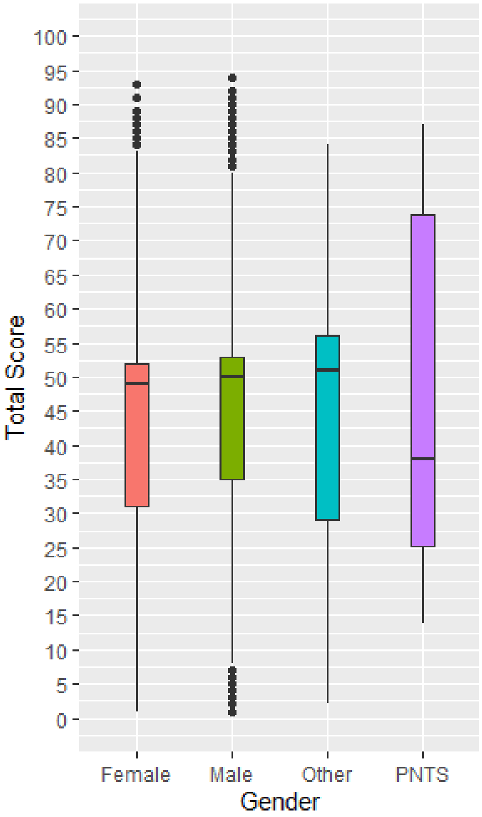
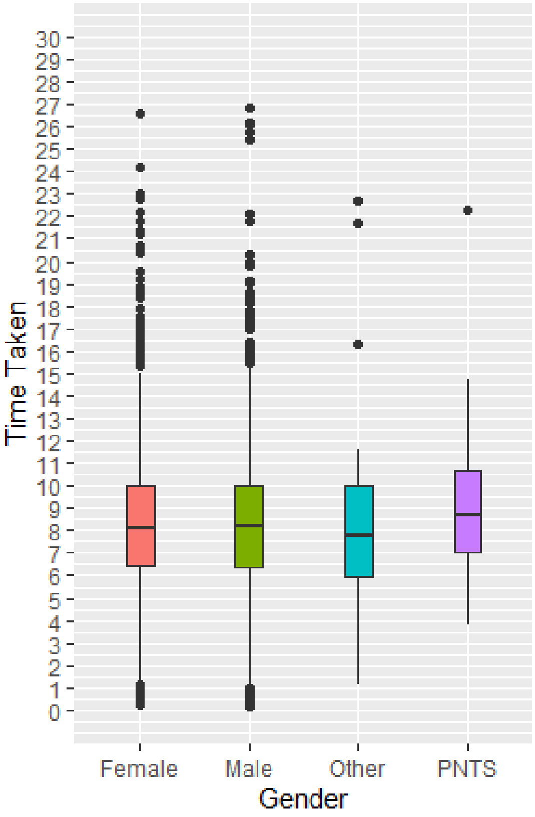


Figure 4-30. Gc:VL Gender by Time Taken



4.4.2 Missing Data

Table 4-16 and Figure 4-31 show the summary statistics of the percentage of missing data within the Lexical Knowledge data set, broken down by each phase of collection. For the ITOS data collection missing data increased linearly. For the ICS Adult data collection, missing data within each variable ranged between 20% and 80% with more significant missing data within Item Set A (items 1-55), administered last for adults. For the ICS School Aged data collection, as with adults, missing data appears to increase in the third set of items (Item Set B; items 67-107).

When considering missing data percentages per participant (Table 4-17), the higher mean and median percentage of missing data by participant for the ICS Adult sample suggests a high number of these participants dropped out of the study, either due to self-imposed time limits or by choice.

Figure 4-32 shows the pattern of missingness across item sets. If an item was missing from an item set, the entire item set was classed as incomplete. Blue cells show the frequency of complete item.

Table 4-16. Percentage of Gc:VL Missing Data by Item

Phase	*n Items*	Mean	SD	Median	Min	Max
ITOS	55	7.40	2.95	8.15	0.25	11.22
ICS Adult	98	43.49	20.39	40.18	13.27	78.58
ICS School Aged	98	10.46	11.61	1.05	0	29.47

Table 4-17. Percentage of Gc:VL Missing Data by Participant

Phase	*n* Participants	Mean	SD	Median	Min	Max
ITOS	1203	7.4	22.42	0	0	98.18
ICS Adult	1130	43.49	27.63	47.96	0	98.98
ICS School Aged	95	10.46	13.64	5.1	0	81.63

Note. Missing data for this table is calculated based on the items available to each participant.

Figure 4-31. Percentage of Gc:VL Missing Data by Phase

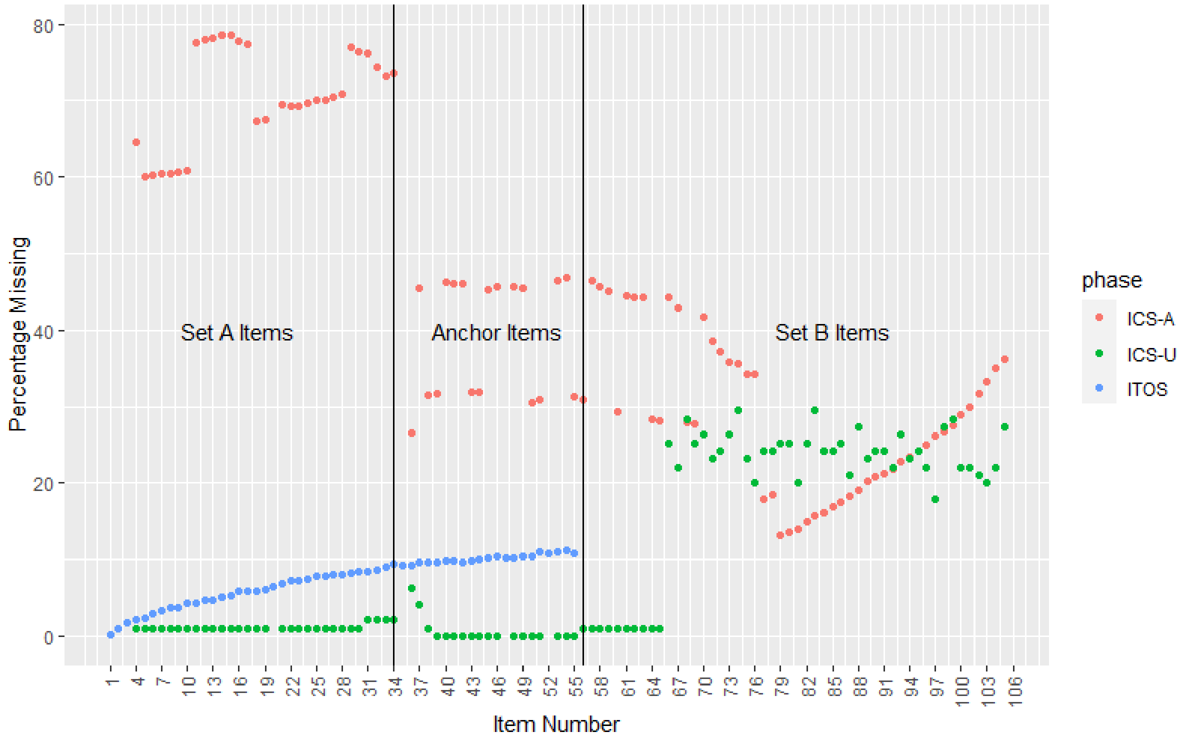
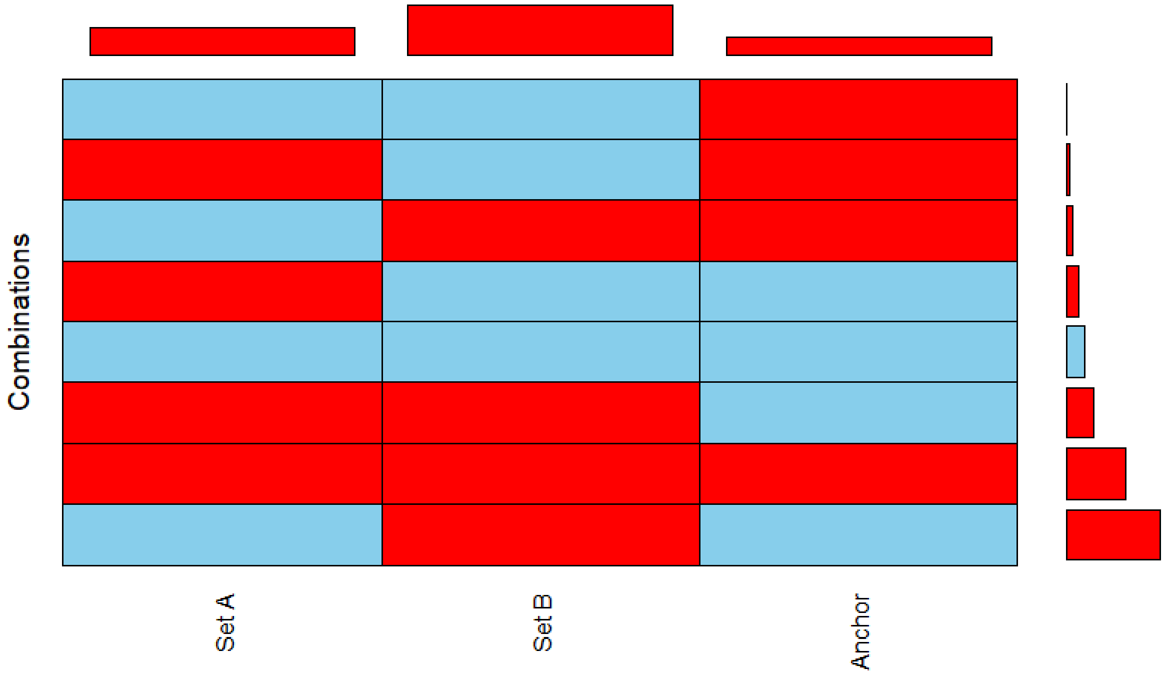


Figure 4-32. Missing Data Pattern for Gc:VL Item Sets



Note. Red cells indicate missing data while blue cells indicate present data. The bars above and beside the cells represent the proportion of participants with that pattern of data.

4.4.3 Data Imputation

While missing data analysis was broken into separate data collection phases and thus missingness calculated on specific item sets administered, the data imputation considered the dataset in entirety. Data was prepared for MICE in stages (Table 4-18). Because item removal resulted in 48 remaining items in the ITOS, participants in the ITOS analysis could have only completed 48 of the 98 items included in the ICS; they automatically had 51.02% of missing data.

This has implications for the interpretation of missing data compared to the previous total missing data analysis (4.4.2). For example, previously a participant was not counted as having missing data if they were in the ITOS sample and had completed every Lexical Knowledge item between items 1 and 55; they had completed every item available to them. In this part of the analysis, for the purposes of imputing data, they were missing data for items 56 through 107 (Figure 4-33). As an additional example, if they only completed items 22 or 24, they would not identify as non-responders, however in this case they would have 100% missing data because the items they did complete were no longer in the data set (Figure 4-34). The pattern of missingness based on item sets after these alterations is represented in Figure 4-35, and suggests after removing of missing data in the steps outlined in Figure 4-34, the most common pattern of data remaining is all three sets of items being complete.

After the data cleaning, items 10 (Hammer) and 12 (Chef) were flagged as constants in the imputation model, and item 51 (Phlegmatic) was flagged as problematic for Random Forest prediction. Items 15 (Laundry), 16 (Wombat), 17 (Hinge), 18 (Winter), 19 (Disappointed), 21 (Flute), 22 (Cauliflower), 23 (Winning), 24 (Barbeque), 25 (Busy), 26 (Shoulder), 27 (Windmill), 28 (Sour), 29 (Dinosaur), 30 (Compass), 31 (Pasties), 32

(Solid), 33 (Oasis) and 34 (Quadrilateral) were found to have poor outflux values and were removed (Figure 4-36 and Figure 4-37). These items have low outflux and thus have low predictive power for the imputation models. They also skew to the right of the graph suggesting higher influx, meaning they are more dependent on the imputation model. Removing these items ensures the imputation models are less biased.

The multivariate data imputation was successfully carried out with 177 individual formulas executed to impute data for the 177 variables remaining. No problems were identified with the imputation, shown in Figure 4-38 and Figure 4-39.

Figure 4-40 and Figure 4-41 show comparisons between the data before and after imputation.

Table 4-18. ICS Gc:VL Missing Data Cleaning

Phase	***n Items***	**Mean**	**SD**	**Median**	**Min**	**Max**
Initial Data	105	48.15	16.84	55.02	17.22	71.21
Full Anchor Items	105	40.59	28.21	38.60	0	78.10
Removed ITOS Items[1]	98	40.72	29.21	61.58	0	78.10
Low Response Items[2]	94	39.19	28.83	21.69	0	74.87
High Missingness	94	10.43	12.64	4.04	0	36.85
Missing Age	94	10.42	12.65	4.26	0	32.98
Time Outliers	94	10.50	12.44	4.55	0	36.87
Score Outliers	94	10.48	12.53	4.43	0	37.24
Multivariate Outliers	94	10.17	11.98	4.57	0	32.98

[1]Items 1, 2, 3, 20, 35, 47, 61 and 68; [2]Items 104, 105, 106, 107

Figure 4-33. Gc:VL Missing Data for Data Imputation Preparation

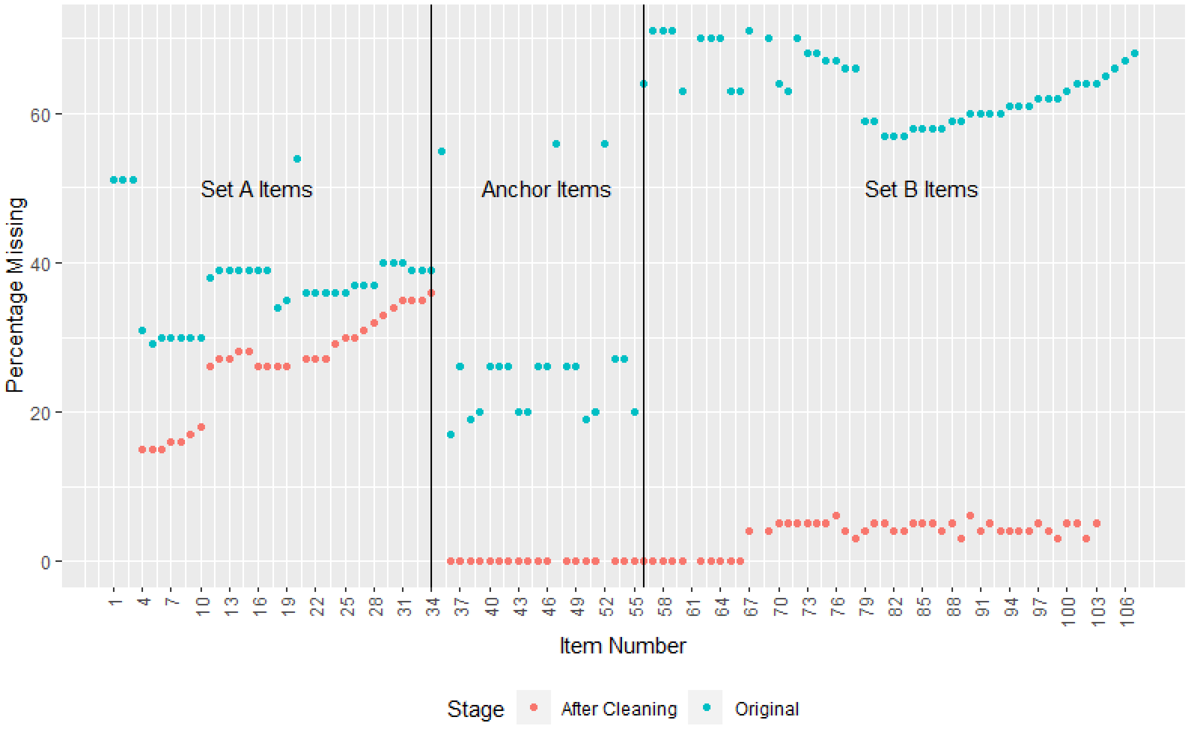


Figure 4-34. Percentage of Missing Gc:VL Items by Participant

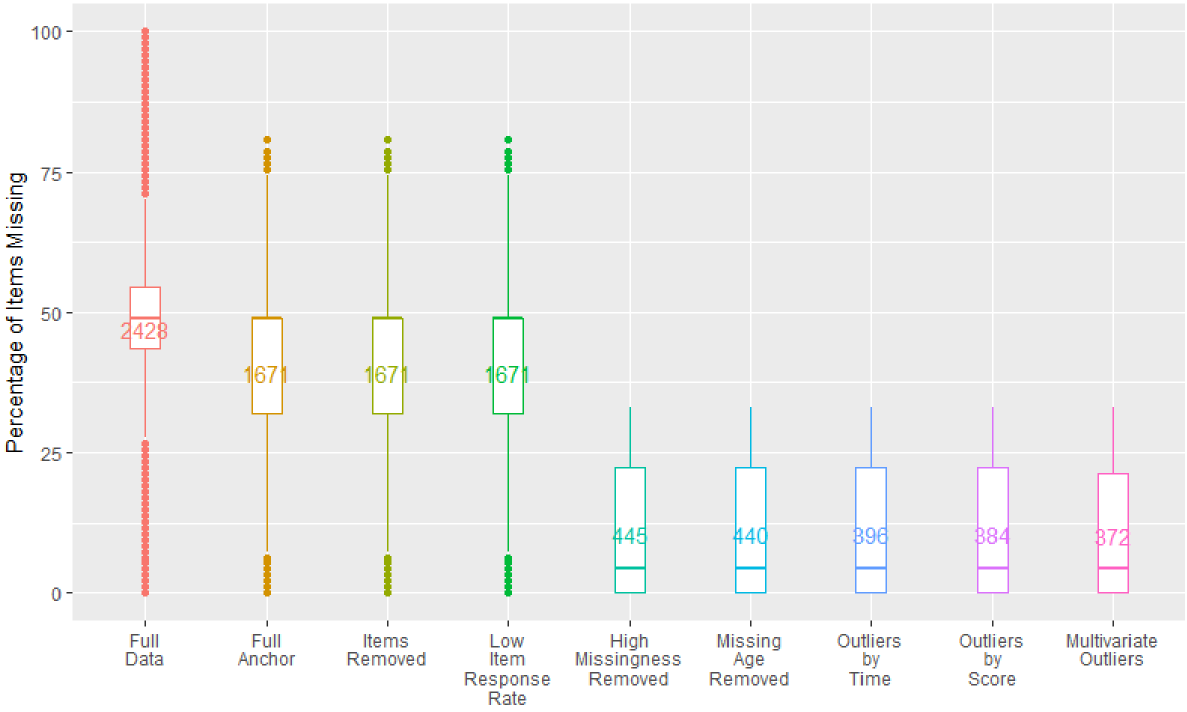
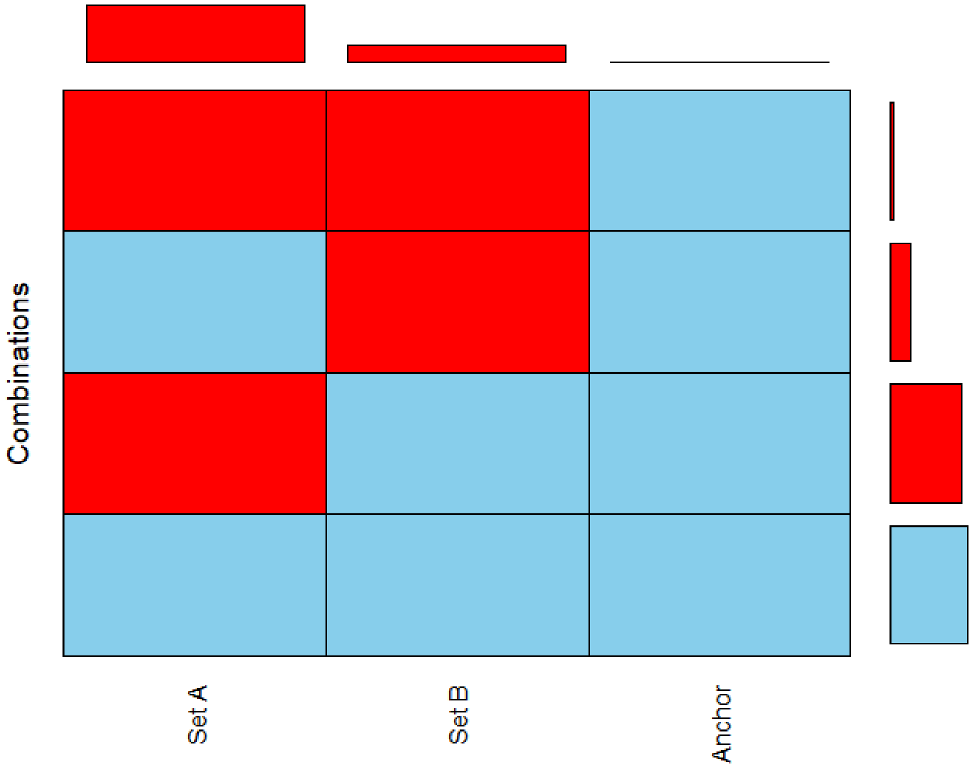


Figure 4-35. Missing Data Pattern Gc:VL with Full Anchor Items



Note. Red cells indicate missing data while blue cells indicate present data. The bars above and beside the cells represent the proportion of participants with that pattern of data.

Figure 4-36. Outflux-Influx Prior to Item Removal

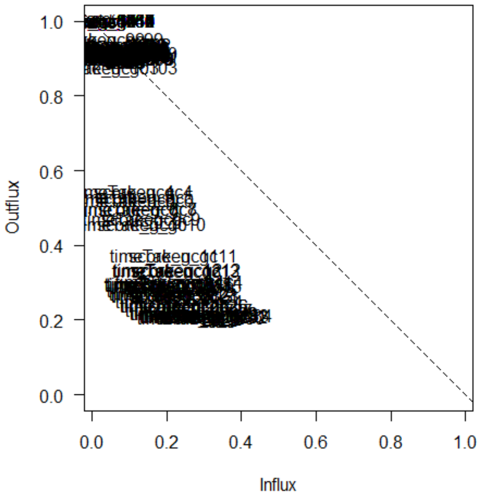
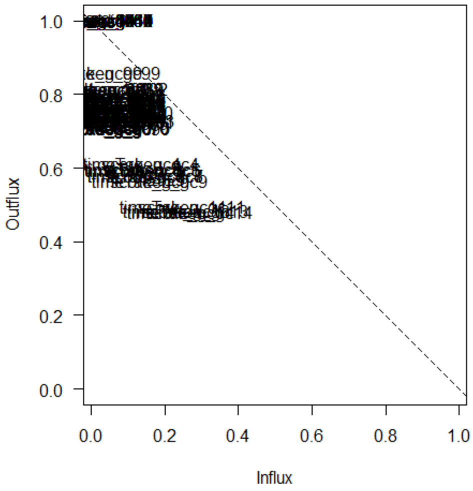


Figure 4-37. Outflux-Influx After Item Removal



Note. Items with low outflux have poor predictive power for the imputation model and items with high influx have high reliance on the imputation model. Ideally you have high outflux and low influx.

Figure 4-38. Gc:VL Score by Imputation

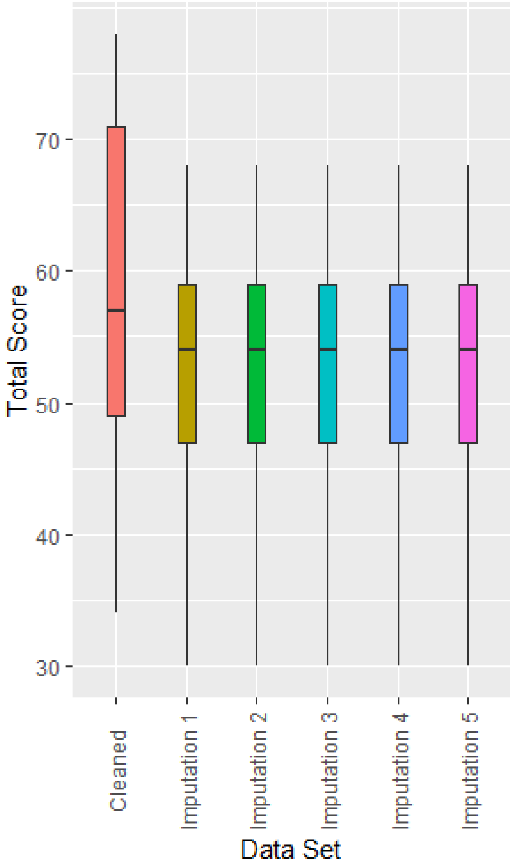


Figure 4-39. Gc:VL Total Time by Imputation

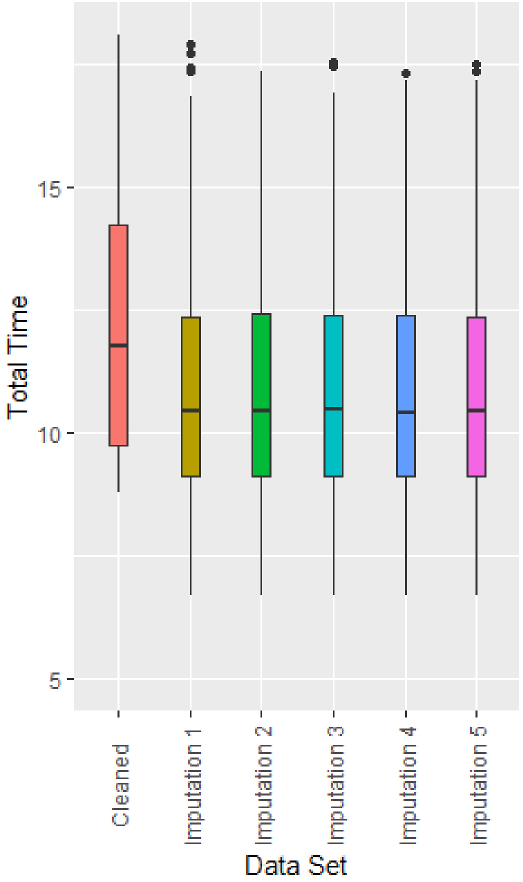


Figure 4-40. Gc:VL ICS Percentage of Items Correct

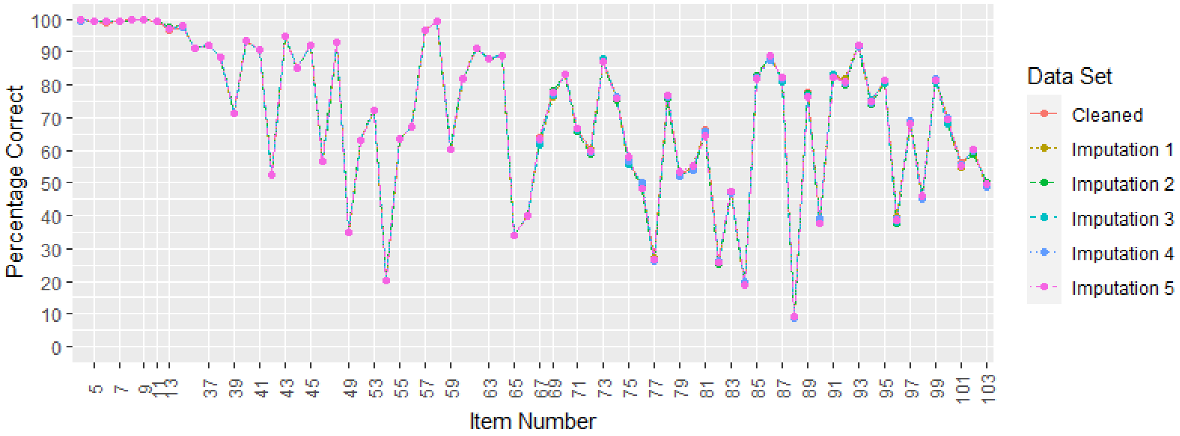
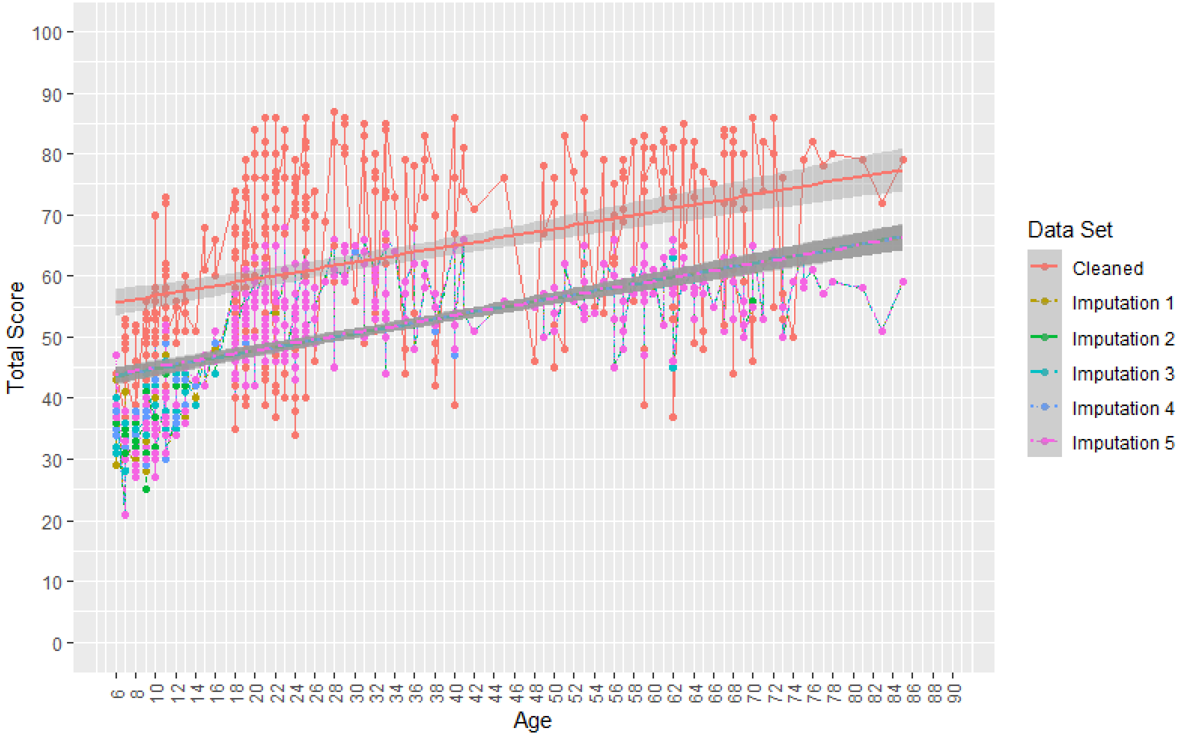


Figure 4-41. Gc:VL Total Score by Age



4.4.4 Reliability

Cronbach's alpha for the Lexical Knowledge 72-item set was .90. This suggests the items have good internal consistency. This reliability is lower than in the ITOS, likely a result of the items measuring a wider range of difficulties. Reliability also stayed relatively high, likely due to a high number of items.

4.4.5 Rasch Analysis

Rasch analysis was conducted on all five imputed data sets. For the 72-item set there was close to acceptable fit across all fit indices. Item fit details are in Appendix I.

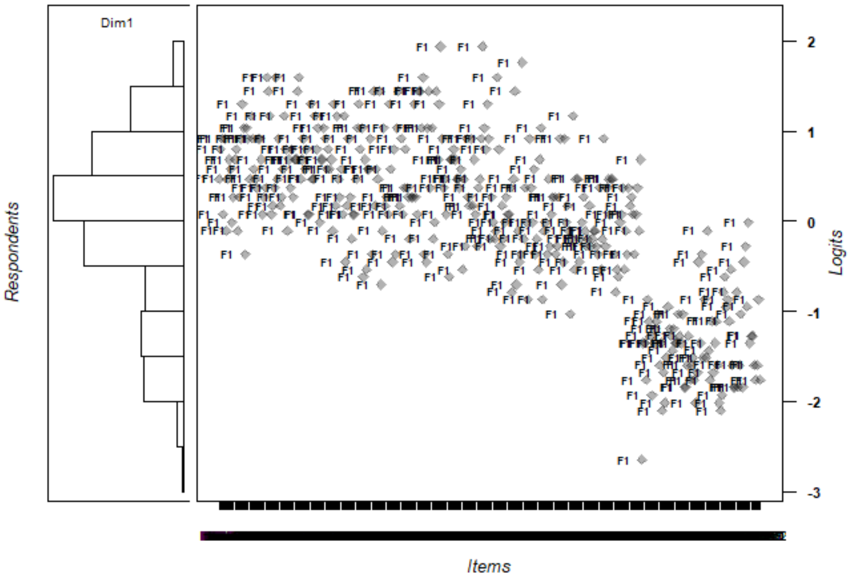
Table 4-19. Rasch Scale Fit Statistics for Gc:VL 72-Item Set

Data	M2	DF	RMSEA	SRMR	TLI	CFI	Marginal Rxx
Imp 1	5431.72***	2555	.06	.11	0.88	0.88	.90
Imp 2	5356.62***	2555	.05	.11	0.88	0.88	.90
Imp 3	5419.41***	2555	.05	.11	0.88	0.88	.90
Imp 4	5600.77***	2555	.06	.11	0.87	0.87	.90
Imp 5	5527.72***	2555	.06	.11	0.87	0.87	.90

Note. *p < .05; **p < .01; ***p < .001

Person fit statistics were calculated (Figure 4-42) with 42 participants found to have Zh values above +2 and below -2. Further analysis identified these were 38 participants aged between 6 and 18, and four participants aged above 50. These participants were likely flagged as 'poorly fitting' as they performed inconsistently on an item-by-item basis compared to other participants. This study included people of very low, average, and very high ability and thus there is expected to be significant variation in performance which may be flagged by overly sensitive statistical tests. Univariate and multivariate outliers were already addressed in the data imputation stage.

Figure 4-42. Gc:VL 72 Item Set Wright Map



For the 72-item set, 43 items were flagged as having poor item fit. These items were analysed from both a qualitative and psychometric perspective.

Across all five imputations, items 4 (Gate), 5 (Sad), 6 (Pineapple), 7 (Koala), 8 (Basketball), 9 (City), 11 (Washing), and 58 (Splinter) returned 'NA' or 'NaN' fit statistics. Item analysis showed very minimal variation in performance on these items across the age groups, with almost all participants obtaining a correct answer (i.e. data sparseness).

Such items are deemed very easy but do have some clinical use for screening children with very low trait levels of Lexical Knowledge. In a CAT, not all items need to be administered to an individual, and item exposure rules can ensure people are not exposed to overly repetitive items. Item 7 (Koala) is culturally loaded, and both items 9 (City) and 11 (Washing) had poor-quality pictures and so were removed. Almost no participants obtained an incorrect answer for items 4 (Gate), 5 (Sad), and 8 (Basketball). Items 6 (Pineapple) and 58 (Splinter) showed the most variation in performance for people aged 6-17 and were retained as they are believed to be clinically useful.

After removal of the above six items, a Rasch analysis was conducted again and 20 items remained problematic. Appendix J shows the item performance by age groups. For items 38 (Fortnight), 40 (Spanner), 56 (Incensed), 62 (Toast), 99 (Stipulate) and 100 (Maître D), only younger participants appeared to have difficulty, suggesting they are good items for discriminating ability in children and adolescents and were retained. Items 96 (Virescent) and 102 (Dystopian) had variable performance and the stimuli are not ambiguous if you know the definition of the word, and thus were retained. Items 54 and 55 were previously flagged for monitoring during new development following the ITOS and were removed. Items 42 (Transparent), 75 (Lateral), 79 (Verbose), 82 (Emblazon), 83 (Luddite), 84 (Scintillate), 88 (Astute), 90 (Bogart), and 98 (Perennial) would be useful items clinically, however the response options were ambiguous even with knowledge of the definition. This combined with low quality images resulted in their removal. Item 65 (Digraph) appeared to have very unpredictable performance, likely to be a result of requiring very specialised knowledge (i.e. linguistics). It was deemed an inappropriate measure of Lexical Knowledge and was removed. These calibrations resulted in a 54-item set.

Table 4-20. Rasch Scale Fit Statistics Gc:VL 54-Item Set

Data	M2	DF	RMSEA	SRMR	TLI	CFI	Marginal Rxx
Imp 1	2642.44***	1430	.05	.11	0.94	0.94	.86
Imp 2	2563.56***	1430	.05	.11	0.95	0.95	.86
Imp 3	2636.93***	1430	.05	.11	0.94	0.94	.86
Imp 4	2703.46***	1430	.05	.11	0.94	0.94	.86
Imp 5	2626.54***	1430	.05	.11	0.94	0.94	.86

Note. *p < .05; **p < .01; ***p < .001

4.4.6 Mokken Analysis

Mokken analysis was completed with the 54-item set. Thirty items were found to have Loevinger's *H* values below .30, with many from Item Set B (i.e. more unlikely to have been completed by school aged participants and predicted via MICE). Some items that did not scale well were also flagged above in the Rasch Analysis. Item 14 (Bridge), 48 (Artichoke), 63 (Witness), 71 (Proficient), 76 (Assuage), and 80 (Lycanthropy) were flagged by the Mokken analysis and evaluation of their response options revealed some ambiguity that was not determined by the expert panel. These items were removed.

Table 4-21. Mokken Analysis for Gc:VL 54 Item Set

Data Set	Loevinger's *H*	Standard Error
54-items		
Imputation 1	.30	.01
Imputation 2	.30	.01
Imputation 3	.30	.01
Imputation 4	.30	.01
Imputation 5	.30	.01
48 items		
Imputation 1	.33	.01
Imputation 2	.33	.01
Imputation 3	.33	.01
Imputation 4	.33	.01
Imputation 5	.33	.01

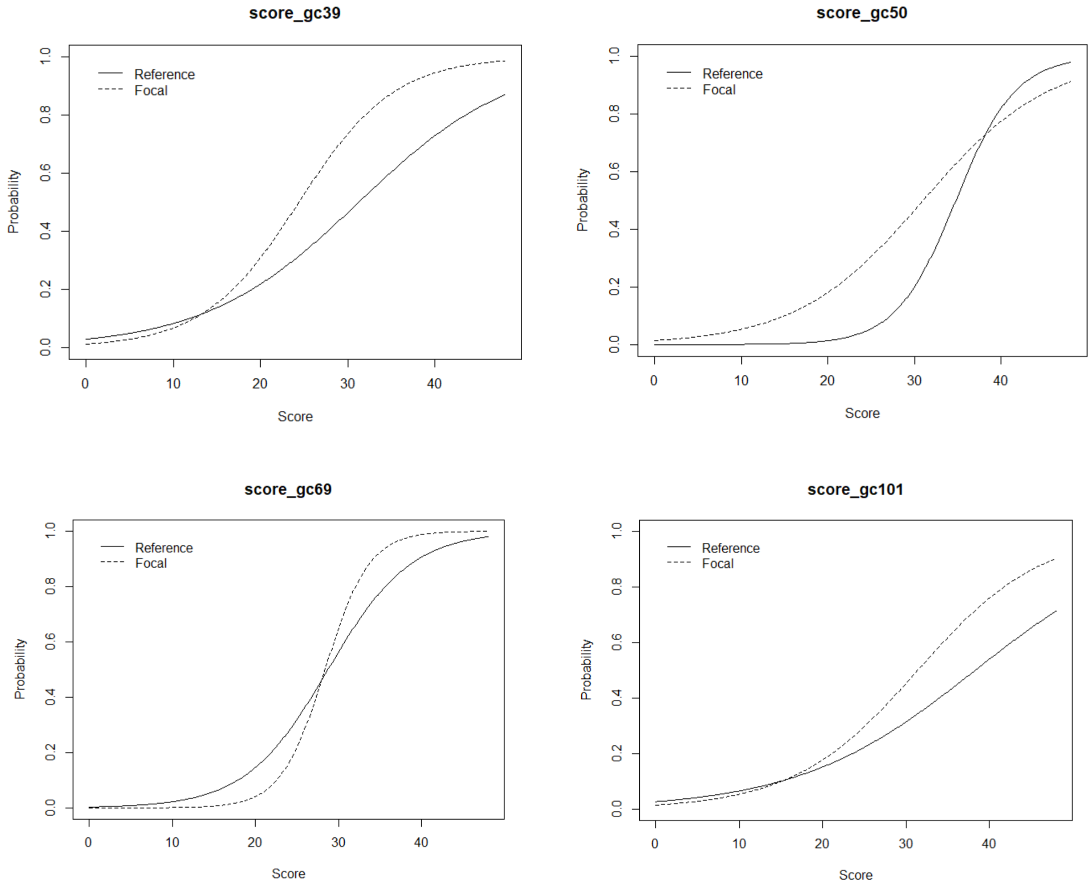
4.4.7 Local Independence

Local dependency was found for the following items: 6, 7, 9, 11, 36, 37, 38, 41, 43, 44, 53, 55, 57, 62, 64, and 81. Each item's content did not reveal any potential learning a participant could gain to assist in correctly answering a later item. Some items were flagged by earlier analyses due to sparseness of data and retained due to clinical utility. Item 37 (Duet) had significantly higher local dependence (.47) and was removed.

4.4.8 Differential Item Functioning

No DIF was found for nationality or device usage. Items 39, 50, 69, and 101 were found to have DIF for gender. No theoretical reason was identified for the discrepancy in performance and thus items were retained.

Figure 4-43. Gc:VL DIF by Gender



Note. The focal group for gender was Male ("m") and the reference group was Female ("f")

4.4.9 Confirmatory Factor Analysis

CFA was completed with the Lexical Knowledge 47-item set. The CFA converged for imputations 1, 2, 3, and 4 and an acceptable fit was found. A higher SRMR in the context of low RSMEA suggests there are some residuals that correlate highly, whereas the RSMEA is lower due to the higher degrees of freedom (DF). Only items 58 (Splinter), 96 (Virescent) and 100 (Maître D) were found to have loadings below .30.

Table 4-22. CFA for Gc:VL 47-Item Set

Imputation	χ^2	DF	RMSEA	SRMR	TLI	CFI
Imp 1	1485.21***	1034	.03	.11	0.98	0.98
Imp 2	1450.54***	1034	.03	.11	0.98	0.98
Imp 3	1427.32***	1034	.03	.11	0.98	0.98
Imp 4	1466.15***	1034	.03	.11	0.98	0.98

Note. *p < .05; **p < .01; ***p < .001

4.4.10 Rasch Item Parameters, ICC and Test Information

Rasch fit statistics (Table 4-23) for the 47-items reflected that of the CFA. The final parameters calculated after item calibration used rules from Rubin (1987) to pool parameter and standard error estimates across the five imputed data sets. These are in Appendix K. The items range from -6.41 θ (item 8; Basketball) to 2.67 θ (item 88; Astute) as shown in Figure 4-44. The highest point of information, and thus the point of highest reliability, is at approximately -2 θ as shown in Figure 4-45 and Figure 4-46, respectively.

Table 4-23. Rasch Scale Fit Statistics for Gc:VL 47-Item Set

Data	M2	DF	RMSEA	SRMR	TLI	CFI	Marginal Rxx
Imp 1	2016.92***	1080	.05	.11	0.95	0.95	.84
Imp 2	1939.36***	1080	.05	.11	0.95	0.95	.84
Imp 3	1965.96***	1080	.05	.11	0.95	0.95	.84
Imp 4	1998.05***	1080	.05	.11	0.95	0.95	.84
Imp 5	1972.87***	1080	.05	.11	0.95	0.95	.84

Note. *p < .05; **p < .01; ***p < .001

Figure 4-44. Gc:VL ICS Rasch ICCs

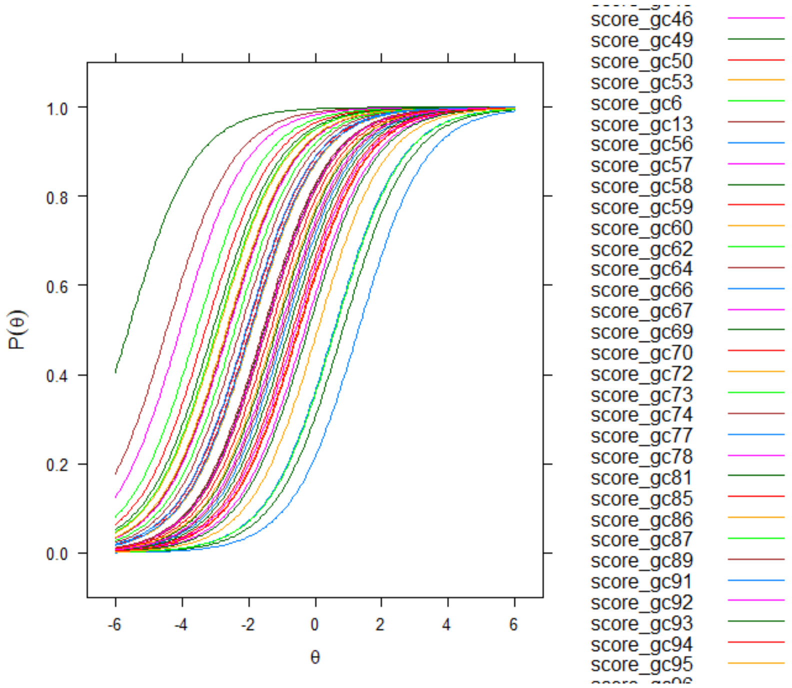


Figure 4-45. Gc:VL ICS Rasch TIC

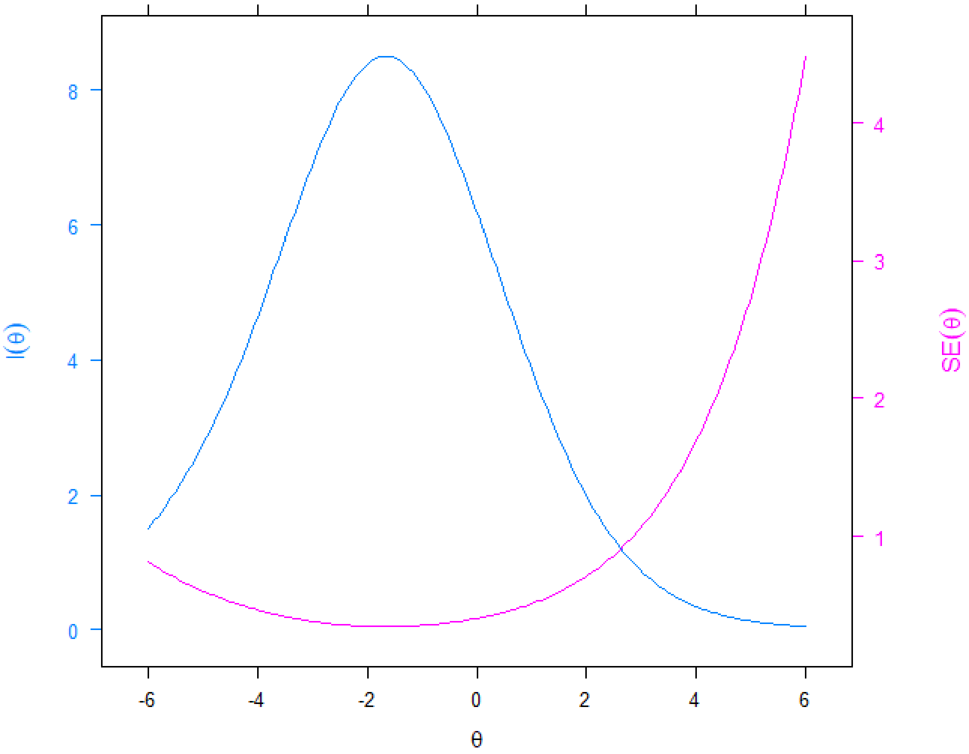
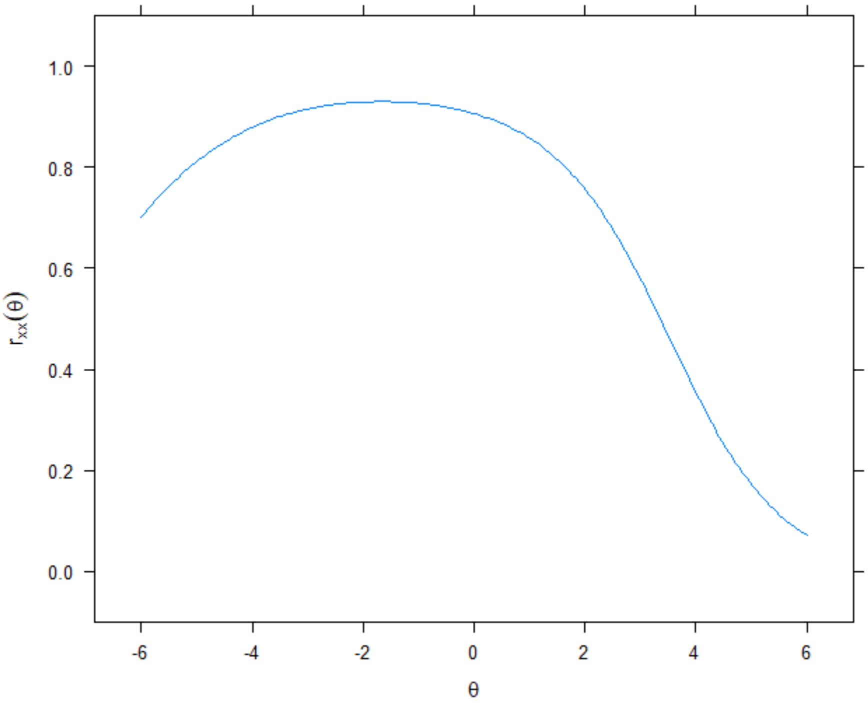


Figure 4-46. Gc:VL ICS Rasch Test Reliability Curve



4.5 Induction Results

4.5.1 Raw Data

The descriptive statistics for time taken and total raw score for the ITOS and ICS samples are included in Table 4-24 and Table 4-25.

The frequencies of the total raw scores for participants are in Figure 4-47. The maximum possible score for the ITOS sample was 33, whereas the maximum total score for the ICS Adult and ICS School Aged samples was 78. Despite more items being available for the ICS Adult participants, more participants scored lower, though this does not account for time spent on the test.

There was a significant difference in the raw score of participants within each phase: *F*(2,1641) = 115.6, *p* < .001 (Figure 4-48). The total time taken for the Induction test was significantly different: *F*(2,1641) = 76.77, *p* < .001 (Figure 4-49).

Performance across the Induction items remained varied, including within the new items (Figure 4-50). Many participants dropped out of the Induction test early in the ICS Adult sample (Figure 4-51), which also shows a somewhat linear relationship between time taken and higher scores.

There was a significant difference between gender identification groups on total score, $F(3,1640) = 5.01$, $p < .01$ (Figure 4-52). Post hoc comparisons using the Tukey HSD test indicated the mean score for the Prefer Not To Say gender identification group was statistically lower than the Male gender identification group, $p = .02$. There was no significant difference between other gender identification groups for total score.

There was also a difference between gender identification groups on total time taken, $F(3,1640) = 3.58$, $p = .01$ (Figure 4-53). Post hoc comparisons using the Tukey HSD test indicated that mean score for the Other gender identification group was statistically lower than the Male gender identification group, $p = .02$. There was no significant difference between other gender identification groups for time taken.

Table 4-24. Descriptive Statistics for Raw Score with the Gf:I Items

	***n* Items**	***n* Participants**	**Mean**	**SD**	**Min**	**Max**
ITOS[1]	33	670	19.95	4.77	1	31
ICS Adult	78	877	14.16	10.83	1	46
ICS School Aged	78	97	23.85	7.95	1	42
All Phases	78	1644	17.09	9.28	1	46

[1]Descriptive statistics varied from the ITOS analysis due changes in missing data methodology

Table 4-25. Descriptive Statistics for Time Taken with the Gf:I Items

	n	**Mean**	**SD**	**Min**	**Max**
ITOS[1]	670	699.02	228.01	6.24	1346.69
ICS Adult	877	518.22	435.33	14.88	2559.11
ICS School Aged	97	876.91	299.38	13.60	1305.18
All Phases	1644	613.07	373.32	6.24	2559.11

[1]Descriptive statistics varied from the ITOS analysis due changes in missing data methodology

Figure 4-47. Gf:I Frequency of Participant Raw Score

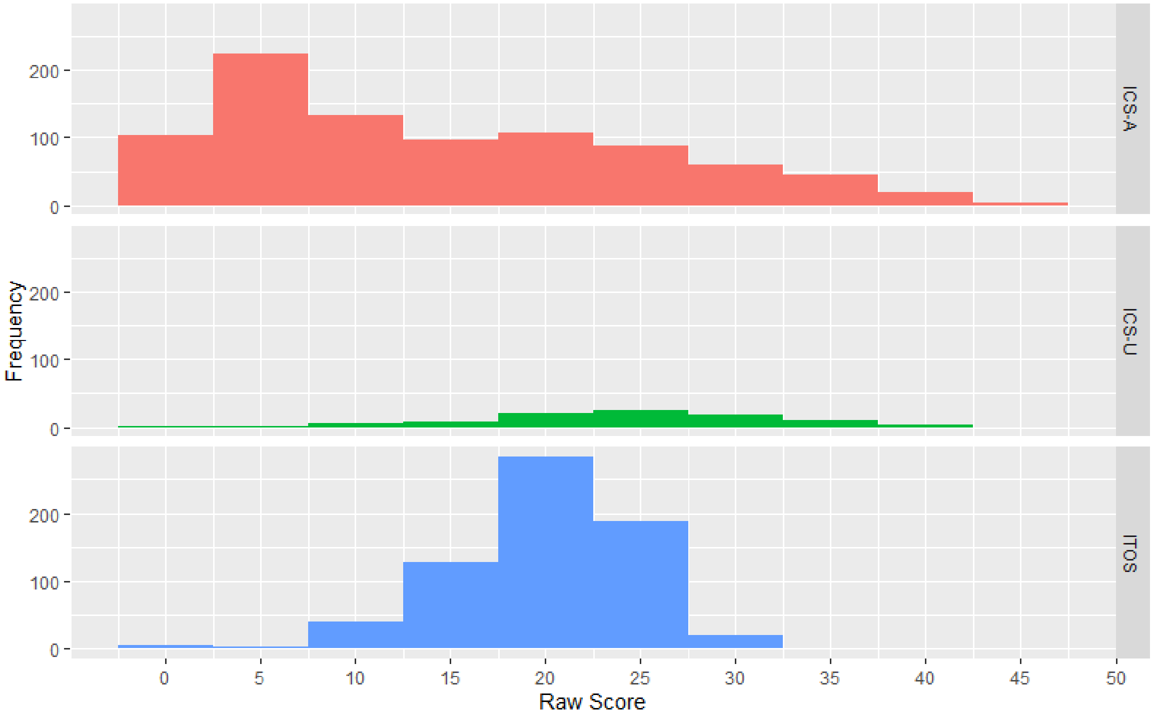


Figure 4-48. Gf:I Total Score by Phase

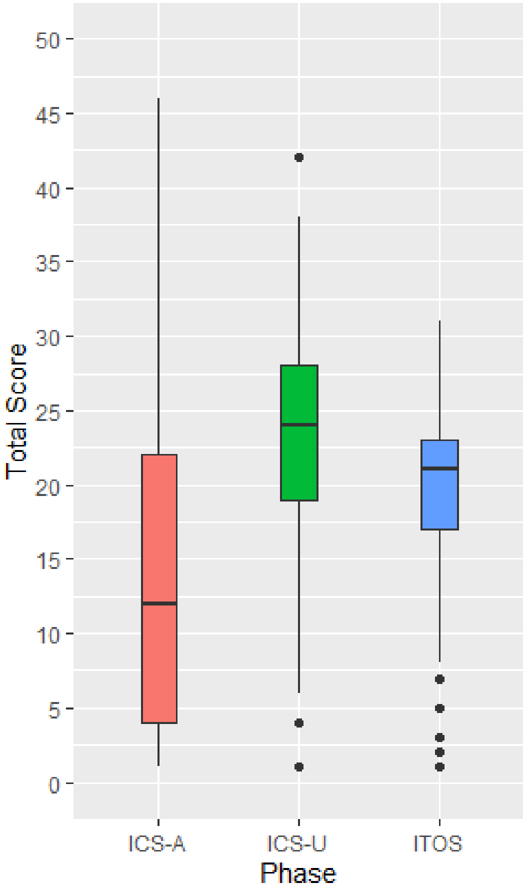


Figure 4-49. Gf:I Time Taken by Phase

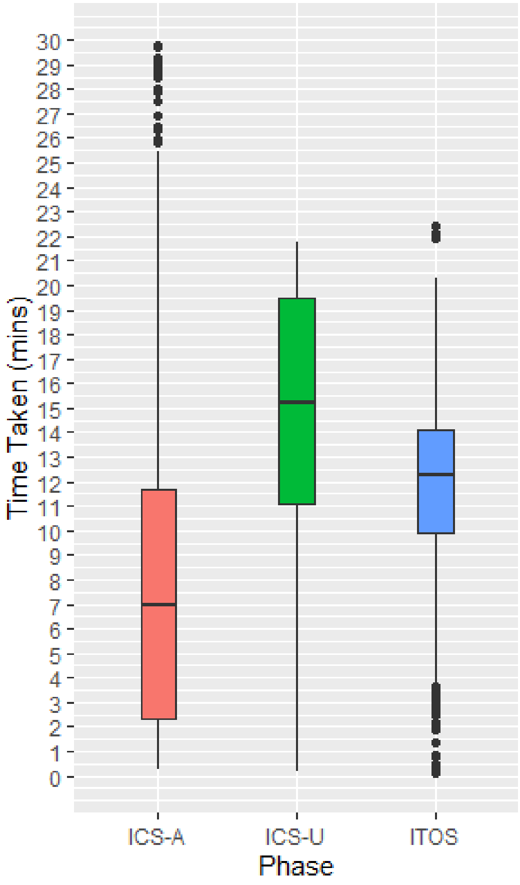


Figure 4-50. Gf:I Percentage of Items Correct

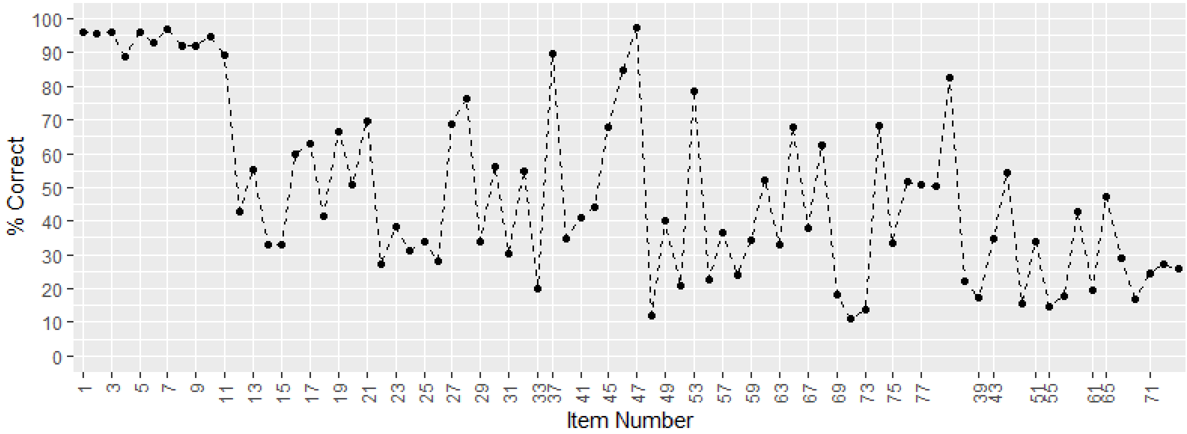


Figure 4-51. Gf:I Total Score by Time Taken

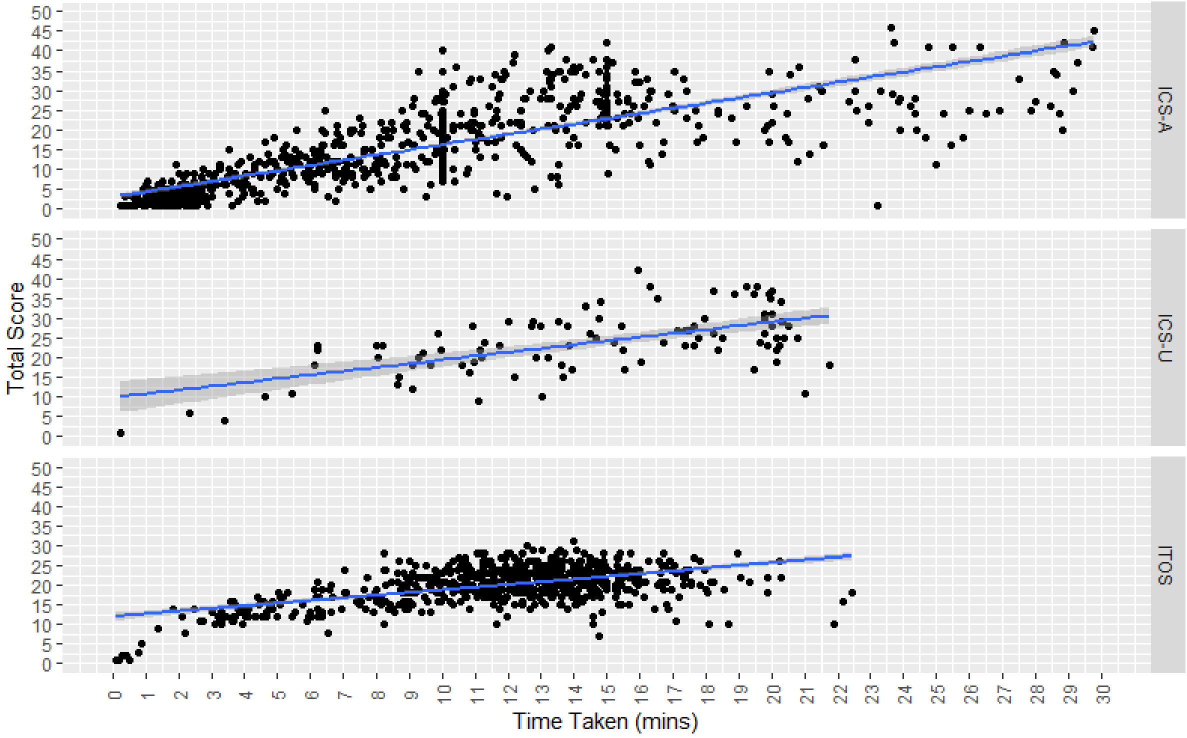


Figure 4-52. Gf:I Gender by Total Score

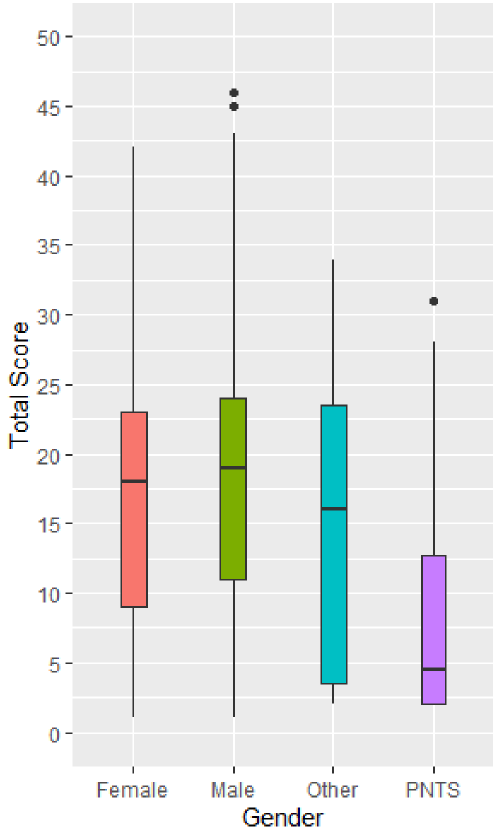
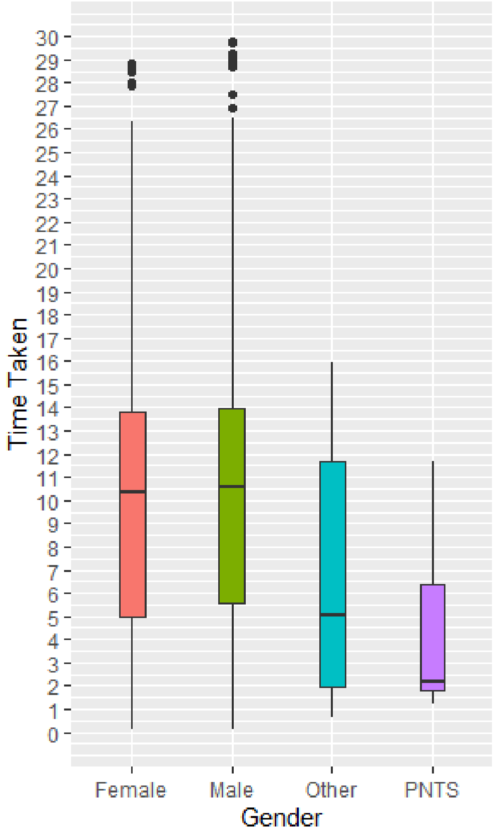


Figure 4-53. Gf:I Gender by Time Taken



4.5.2 Missing Data

For the ICS data collection, Induction items were split into 30 second and 60 second groups, as well as broken down into their item sets as per 4.3.2.6. The 30 second items were administered prior to the 60 second items. While no items were removed from the ITOS, there are some items of concern that were noted for monitoring.

Table 4-26 and Figure 4-54 show summary statistics of the percentage of missing data within the Induction data set, broken down by each phase of data collection. Due to the disordered administration of items (by number) based on different time limits between item sets, there is variation in performance across items in the ICS Adult phase.

When considering the missing data percentages per participant in Table 4-27, the higher mean and median percentage of missing data per participant for the ICS Adult sample suggests a high number of participants in this sample dropped out of the study, likely due to either self-imposed time limits or by choice. Figure 4-55 shows the pattern of missingness across the item sets. For each participant, an item set was classed as incomplete if an item was missing from that set.

Table 4-26. Percentage of Gf:I Missing Data by Item

Phase	*n Items*	Mean	SD	Median	Min	Max
ITOS	33	6.76	5.43	4.93	0.75	17.01
ICS Adult	78	66.24	21.95	74.29	30.78	93.16
ICS School Aged	78	9.11	7.95	7.22	1.03	28.87

Table 4-27. Percentage of Gf:I Missing Data by Participant

Phase	*n* **Participants**	Mean	SD	Median	Min	Max
ITOS	670	6.76	17.3	0	0	96.97
ICS Adult	877	66.24	24.62	71.79	0	98.72
ICS School Aged	97	9.11	18.71	0	0	98.72

Note. Missing data for this table is calculated based on the items available to each participant.

Figure 4-54. Percentage of Gf:I Missing Data by Phase

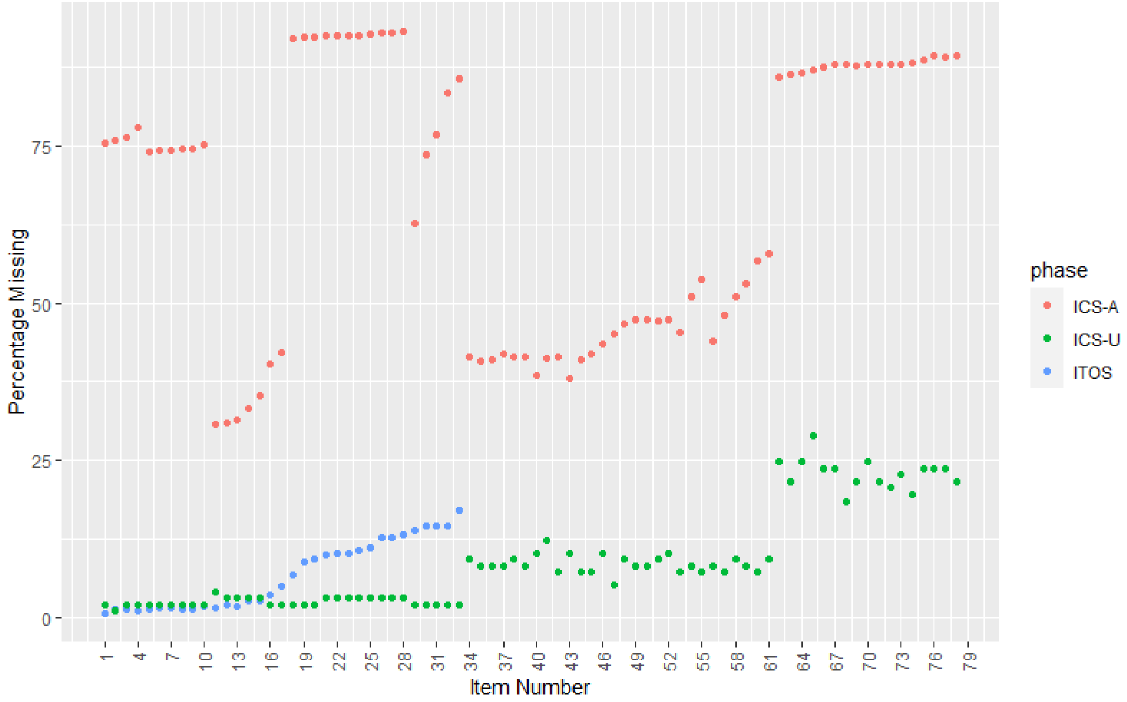
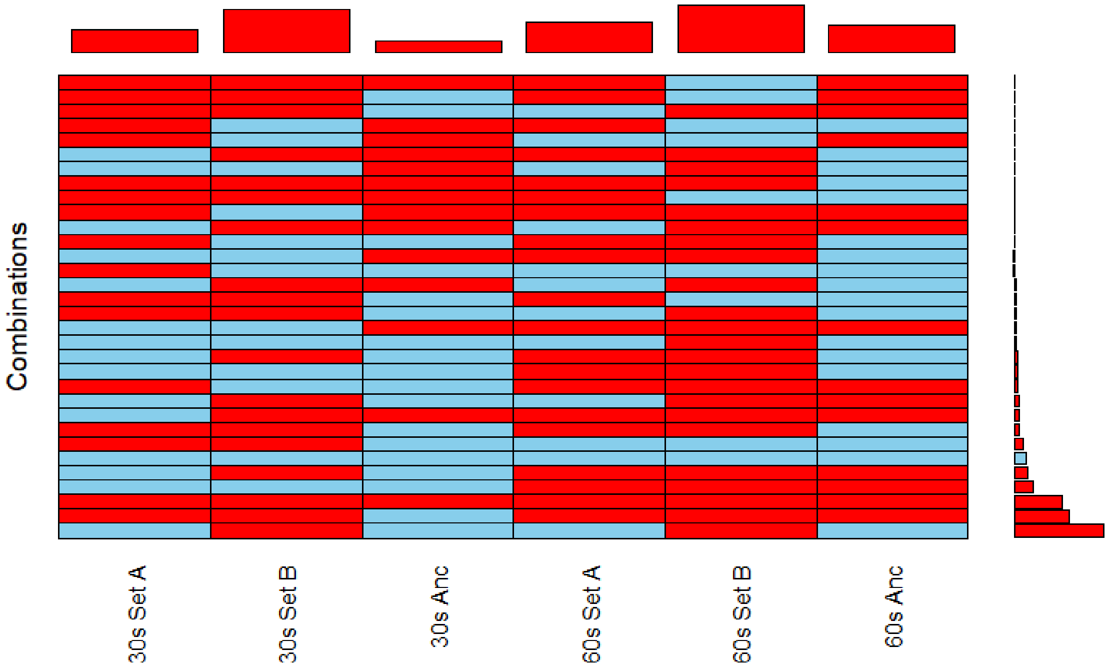


Figure 4-55. Missing Data Pattern for Gf:I Sets



Note. Red cells indicate missing data while blue cells indicate present data. The bars above and beside the cells represent the proportion of participants with that pattern of data.

4.5.3 Data Imputation

As per the Lexical Knowledge analysis the data imputation will consider the dataset in entirety. Data was prepared for MICE in stages as detailed in the methodology section and demonstrated in Table 4-28.

This has implications for the interpretation of missing data compared to the previous section. For example, in the missing data analysis section (4.5.2), a participant was not counted as having missing data if they were in the ITOS sample and had completed every Induction item between items 1 and 33.In this part of the analysis however, they were considered to be missing data for items 34 through 78 (Figure 4-56). Some items were removed prior to data imputation cleaning due to concerns about the reliability of the items being raised by the expert panel. The most significant drops in participants were caused by removal of those that either did not have full anchor items or had high rates of missingness (Figure 4-57). The pattern of missingness based on item sets after these alterations is presented in Figure 4-58, and suggests the most common pattern of data remaining is all six items sets being complete.

After data cleaning, items were analysed to determine if they would be flagged as problematic for MICE. No items were flagged as problematic for the Random Forest algorithm. Items 1-10 were flagged as having an outflux below .30 (Figure 4-59 and Figure 4-60) and were removed.

MICE was successfully carried out with 81 individual formulas executed for 81 variables. No problems were identified with the imputation. Figure 4-61, Figure 4-62, Figure 4-63 and Figure 4-64 show comparisons between the data before and after imputation.

Table 4-28. ICS Gf:I Missing Data Cleaning

Phase	***n Items***	**Mean**	**SD**	**Median**	**Min**	**Max**
Initial Missing Data	78	60.55	20.29	63.29	17.27	89.78
Full Anchor Items	78	46.84	35.27	73.19	0	84.32
Removed Items[1]	72	50.44	34.31	73.32	0	84.32
Low Response Rate[2]	47	35.13	33.53	12.33	0	74.53
High Missingness	47	10.65	13.61	2.29	0	37.14
Missing Age	47	10.61	13.60	2.33	0	37.21
Time Outliers	47	10.67	13.27	2.60	0	35.71
Score Outliers	47	10.65	13.35	2.67	0	36.00
Multivariate Outliers	47	10.661	13.73	2.82	0	37.32

[1]Items 22, 24, 25, 30, 31, 33, [2]Items 34, 38, 39, 43, 44, 50, 51, 55, 56, ,60, 61, 65, 66, 69, 70-80

Figure 4-56. Gf:I Missing Data for Data Imputation Preparation

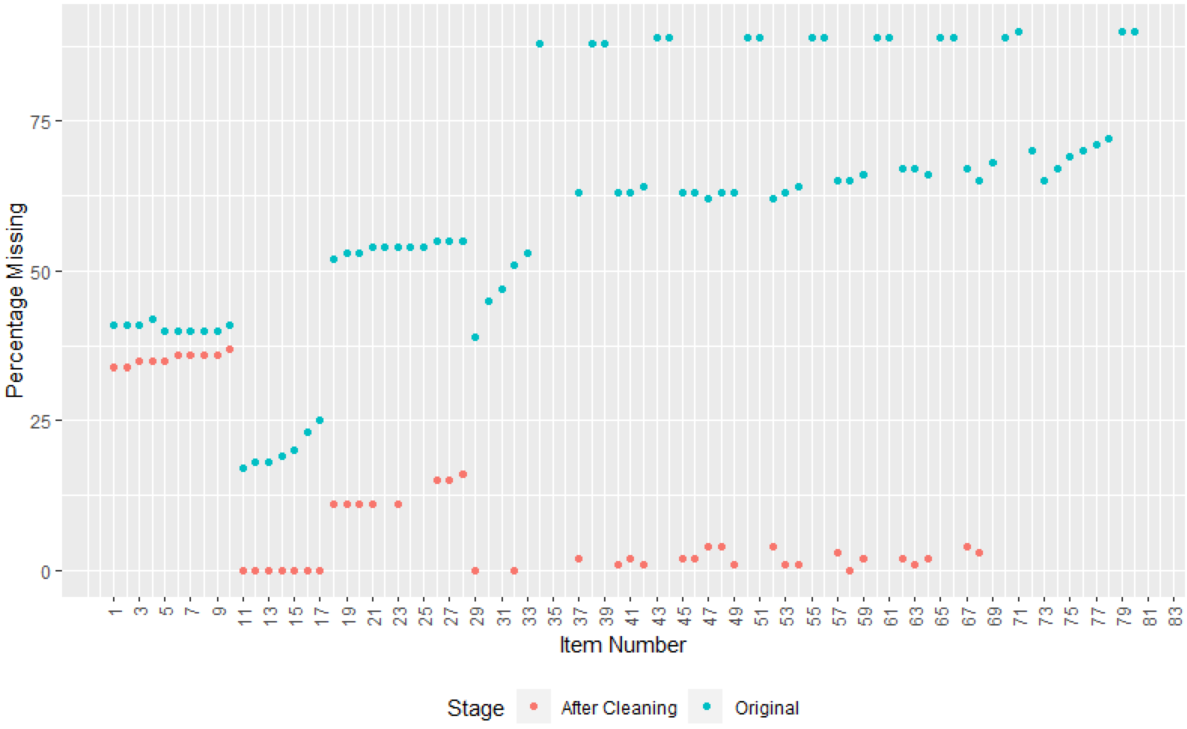


Figure 4-57. Percentage of Missing Gf:I Items by Participant

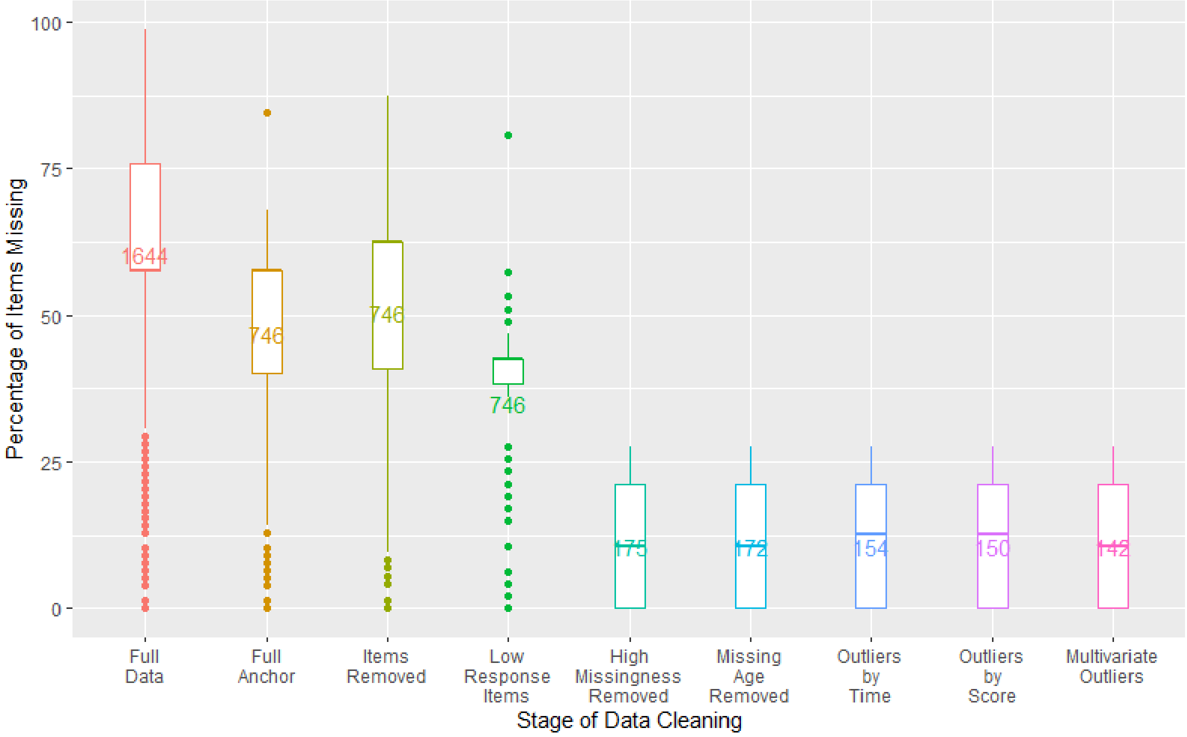
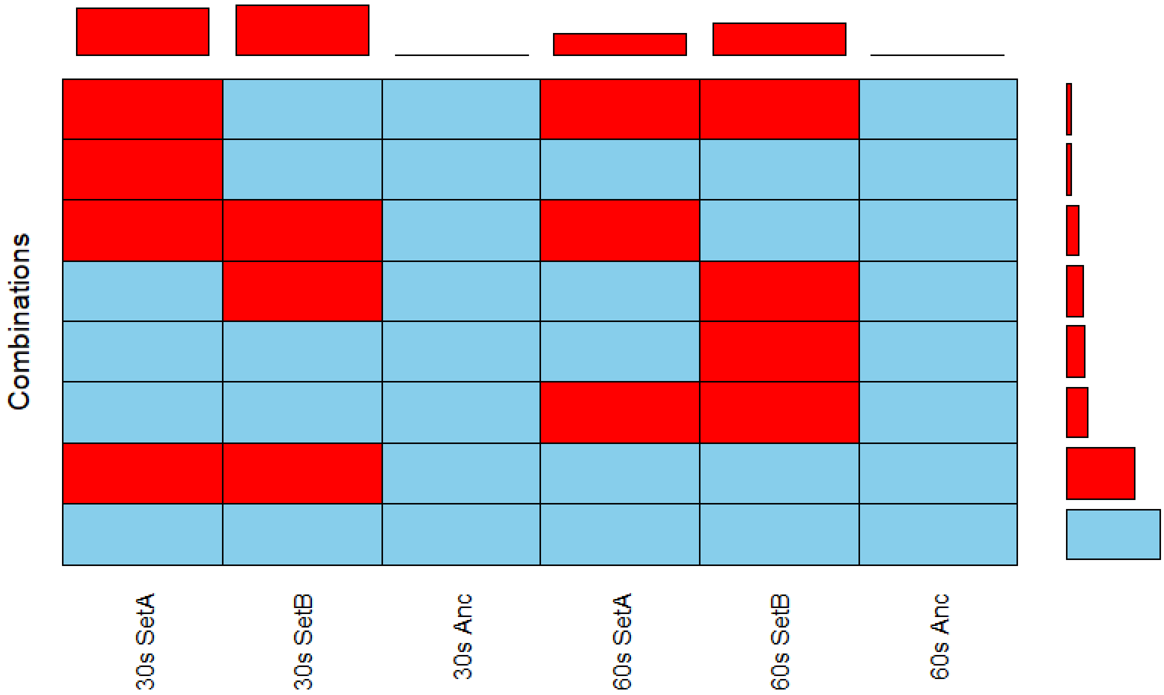


Figure 4-58. Missing Data Pattern for Gf:I with Full Anchor Items



Note. Red cells indicate missing data while blue cells indicate present data. The bars above and beside the cells represent the proportion of participants with that pattern of data.

Figure 4-59. Outflux-Influx Before Removal

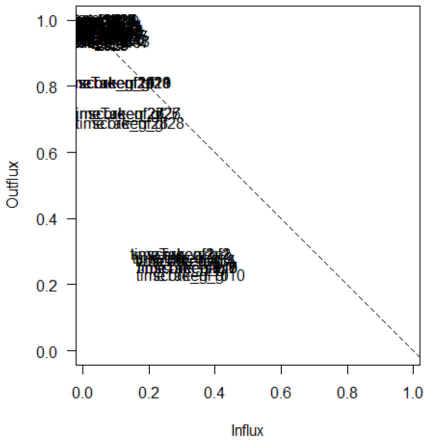
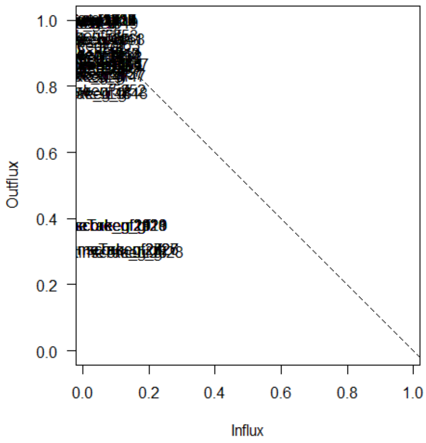


Figure 4-60. Outflux-Influx After Removal



Note. Items with low outflux have poor predictive power for the imputation model and items with high influx have high reliance on the imputation model. Ideally you have high outflux and low influx.

Figure 4-61. Gf:I Score by Imputation

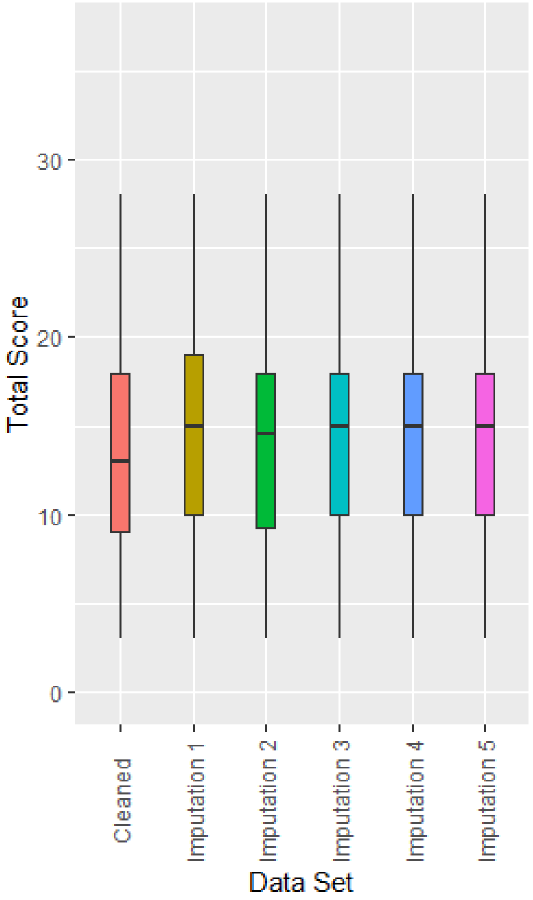


Figure 4-62. Gf:I Total Time by Imputation

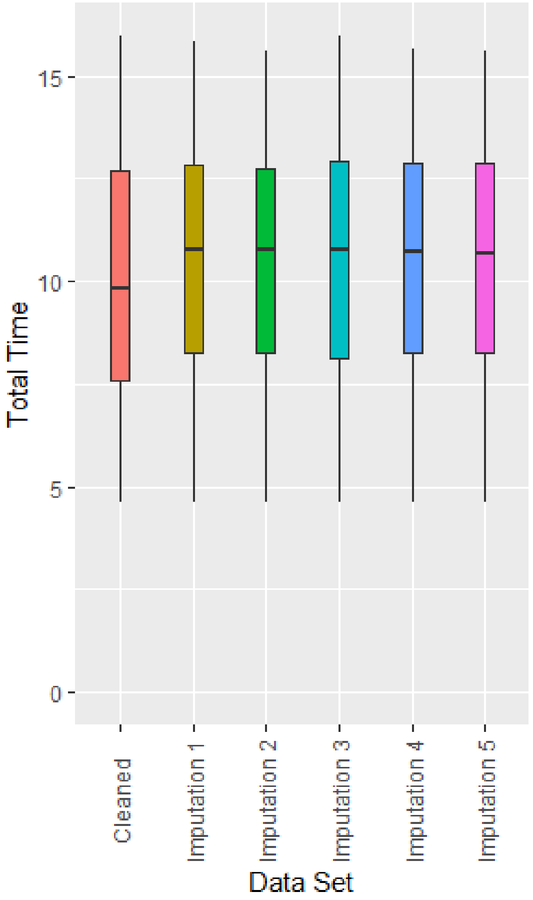


Figure 4-63. Gf:I ICS Percentage of Items Correct

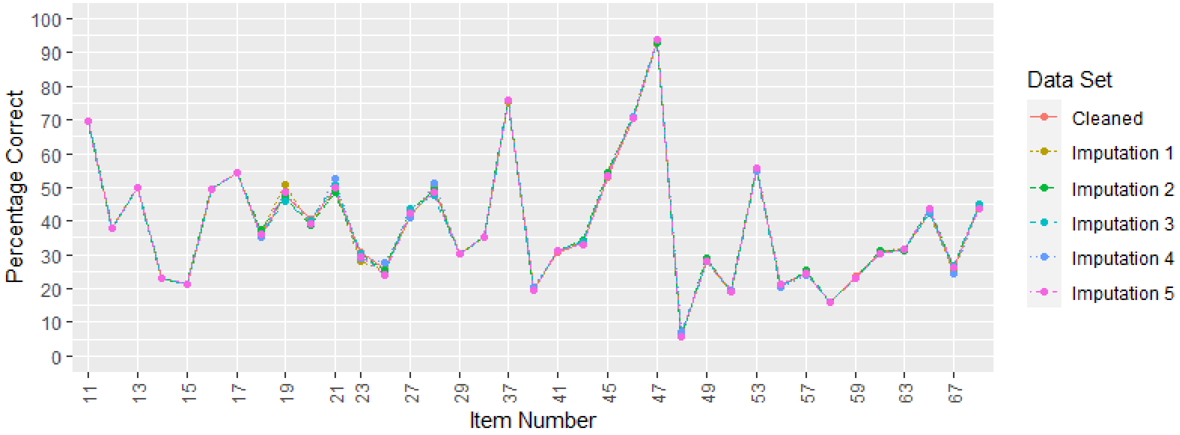
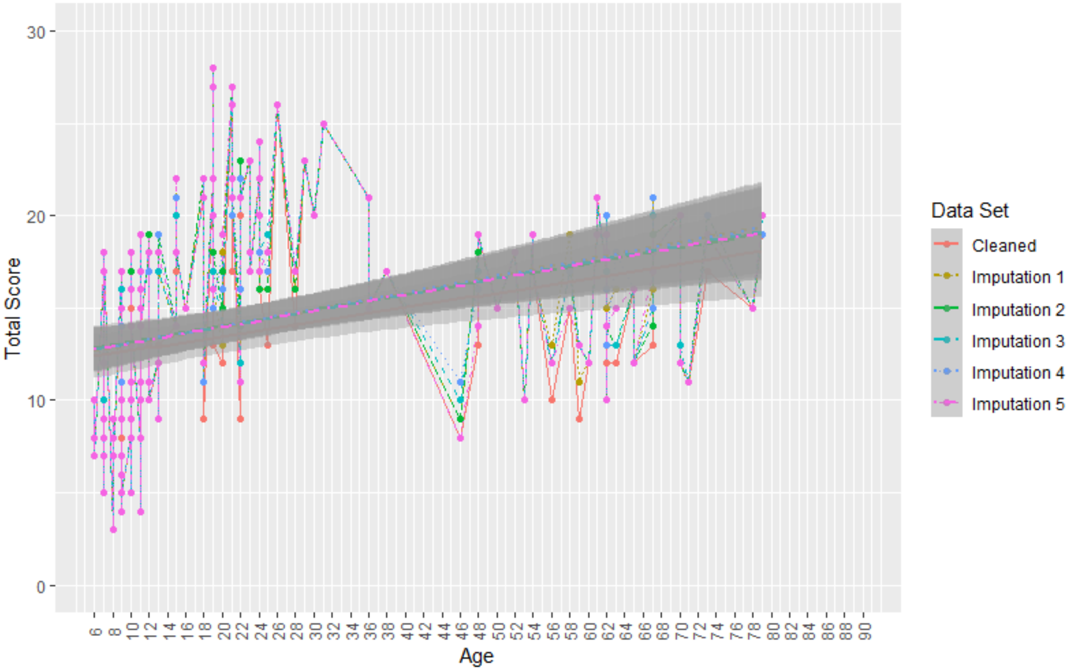


Figure 4-64. Gf:I Total Score by Age



4.5.4 Reliability

Cronbach's alpha for the Induction items in the ICS was calculated across each of the imputed data sets and found to have a mean of .78 (good). This was lower than the Cronbach's alpha in the ITOS, reflecting an increased diversity in difficulty.

4.5.5 Rasch Analysis

Rasch analysis was conducted on all five imputed data sets. For the 37-item set there was close to acceptable fit across all fit indices. Item fit details are in Table 4-29. None of the items flagged for monitoring from the ITOS (items 22, 24, 25, 30, 31 and 33) were in the imputed data sets due to being removed in preparation for the MICE.

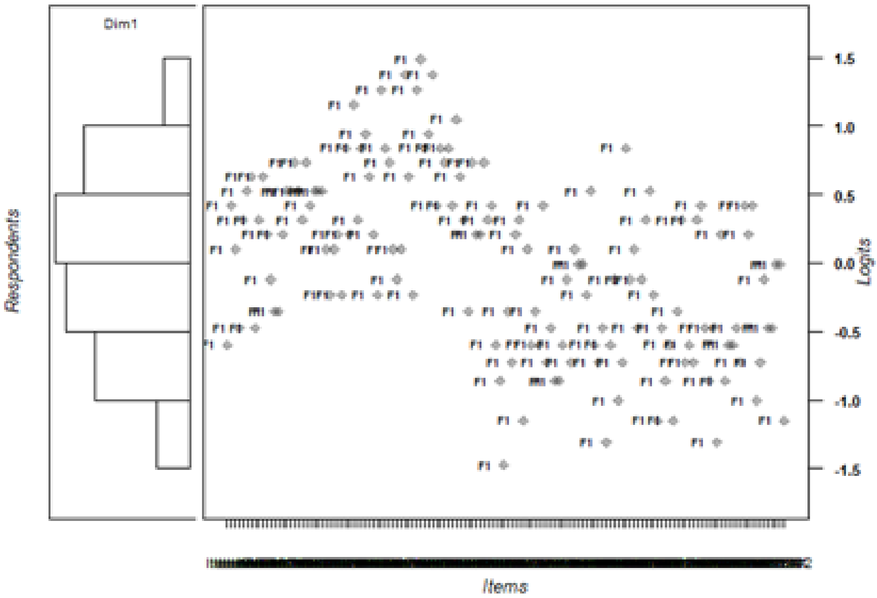
Table 4-29. Rasch Scale Fit Statistics for Gf:I 37-Item Set

Data	M2	DF	RMSEA	SRMR	TLI	CFI	Marginal Rxx
Imp 1	876.64***	665	.05	.11	0.82	0.82	.86
Imp 2	858.56***	665	.05	.11	0.83	0.83	.86
Imp 3	857.95***	665	.05	.10	0.82	0.82	.86
Imp 4	863.78***	665	.05	.11	0.82	0.82	.86
Imp 5	836.12***	665	.04	.10	0.85	0.85	.86

Note. *p < .05; **p < .01; ***p < .001

Person fit statistics were calculated (Figure 4-65), and six participants found to have Zh values outside ±2. Consistent with the Lexical Knowledge items, four of these participants were School Aged. The range of performance was more centralised for Induction (-1.5 to 1.5) compared to Lexical Knowledge (-3.0 to 3.0). This was likely due to the removal of participants after identifying missing data during MICE preparation.

Figure 4-65. Gf:I 37-Item Set Wright Map



For the 37-item set, items flagged as having poor fit was dependent on the imputation data set. Items 17 and 52 were flagged in all five Rasch models, while item 20 was flagged in imputations 1, 2 and 5, item 37 in imputations 3 and 4, and item 49 in imputation 4. These items were further analysed.

For item 17 (Figure 4-66), the pattern progresses based on adding two shapes and changing the colour *and* shape. The answer is nine triangles, however the response options provided two options with nine triangles, one in an ordered fashion (incorrect) and a disordered fashion (correct). This was deemed ambiguous and removed.

For item 52 (Figure 4-67) the final pattern in the bottom right cell is unpredictable as the pattern of change from the top left to either the bottom left or top right is the same; there is not enough information available for the participant to reliably answer. The other response options are clearly incorrect so participants may obtain the answer by a process of deduction. As this was a test of induction, the item was removed.

For item 20 (Figure 4-68) the answer relies on horizontal flipping. As the matrix is only 2x2 rather than 3x2 where the pattern is displayed more consistently, it may have caused some uncertainty. Unlike item 52, where it is not possible to detect the rule without referring to the response options, the rule for item 20 can be determined without consulting potential answers. This appeared to be most challenging for 6- to 17-year-olds in comparison to 18- to 29-year-olds, suggesting it was a good item to retain.

For item 37 (Figure 4-70) there were only four response options (a teal hexagon, a yellow star, a green square and a green hexagon) with a fairly obvious rule. Only 1 out of 57 adults obtained an incorrect answer here, likely due to other factors outside of Induction ability or very low ability. Some children obtained the incorrect answer, likely due to the size of the matrix and the distractor green square. This item was flagged in only two of the imputations, so was retained.

For item 49 (Figure 4-69) the shape rotates 90 degrees to the right, however due to the order of shapes in the bottom row, and there being three response options with rotated purple triangles on a yellow background, there is some opportunity for participants who are not paying close attention to make a mistake. Looking at item performance across age groups, there appeared to be inconsistency and thus this item was removed.

After removing these three items, Rasch was conducted again and these calibrations resulted in a 34-item set (Table 4-30).

Figure 4-66. Gf:I Item 17

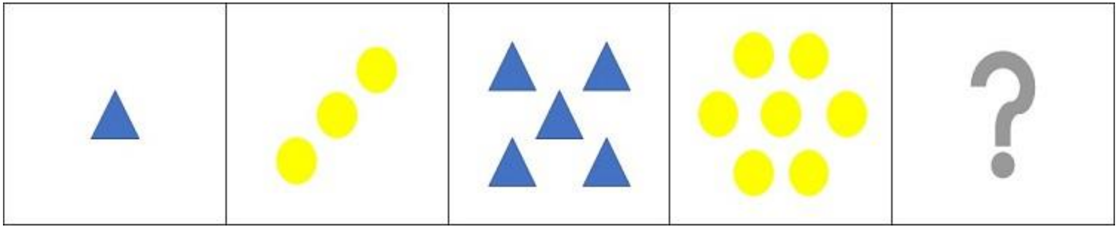


Figure 4-67. Gf:I Item 52

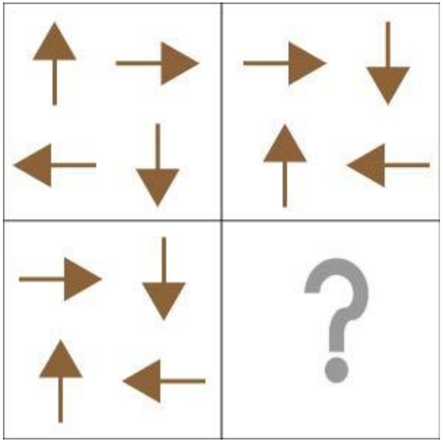


Figure 4-68. Gf:I Item 20

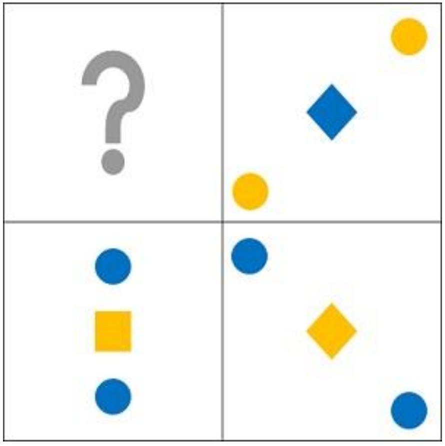


Figure 4-69. Gf:I Item 49

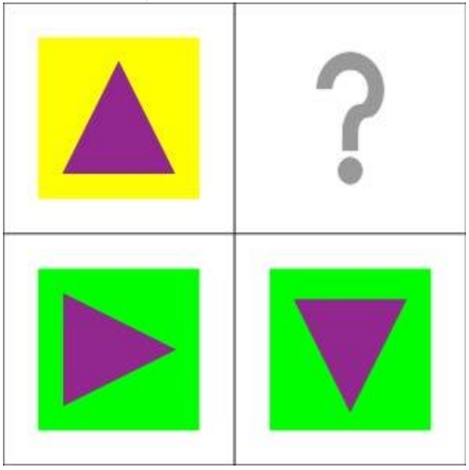


Figure 4-70. Gf:I Item 37

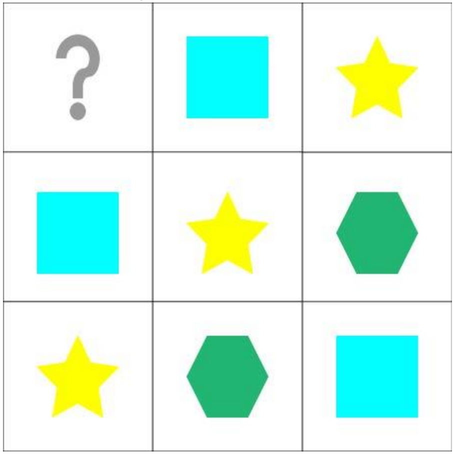


Table 4-30. Rasch Scale Fit Statistics Gf:I 34-Item Set

Data	M2	DF	RMSEA	SRMR	TLI	CFI	Marginal Rxx
Imp 1	711.60***	560	.04	.10	0.86	0.86	.85
Imp 2	692.23***	560	.04	.10	0.88	0.88	.85
Imp 3	691.37***	560	.04	.10	0.87	0.87	.85
Imp 4	693.29***	560	.04	.10	0.88	0.88	.85
Imp 5	672.80***	560	.04	.10	0.89	0.90	.85

Note. **p* < .05; ***p* < .01; ****p* < .001

4.5.6 Mokken Analysis

Mokken analysis was completed with the 34-item set. Thirty items were found to have Loevinger's *H* values below .30; likely due to substantial variation in item complexity. There was no discernible pattern across items characteristics: matrix size, number of response options, straight lines or curves, rotations, number of rules, number of colours, and use of sequential progression. Items 13, 15, 18, 20, 29, 40, 41, 42, 48, 54, 58, and 63 displayed low Loevinger's *H* values. Item 20, already discussed above, was not removed however all others were. This produced noticeable improvements in fit.

Table 4-31. Mokken Analysis for Gf:I 34 Item Set

Data Set	**Loevinger's *H***	**Standard Error**
34 items		
Imputation 1	.15	.02
Imputation 2	.15	.02
Imputation 3	.15	.02
Imputation 4	.15	.02
Imputation 5	.15	.02
23 items		
Imputation 1	.22	.02
Imputation 2	.22	.02
Imputation 3	.21	.02
Imputation 4	.22	.02
Imputation 5	.22	.02

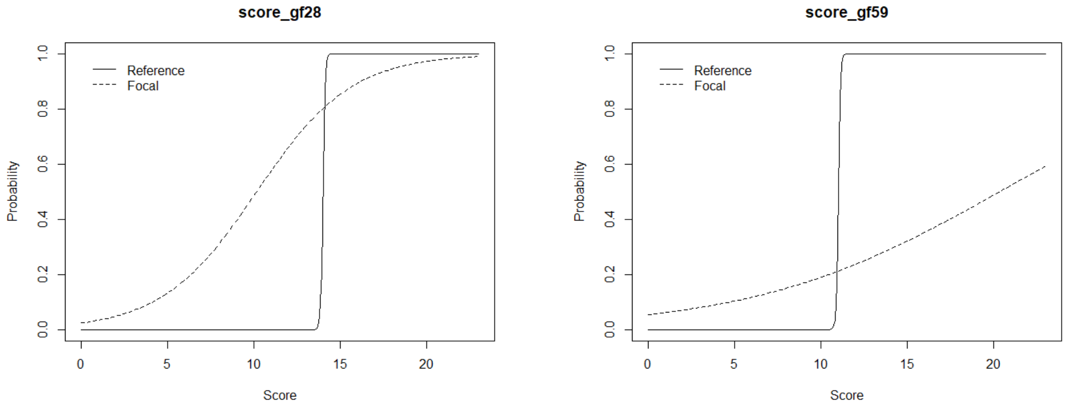
4.5.7 Local Independence

Local dependency was flagged for items 11, 29, 32, 37,42, 45 and 53 with only minor inflation of residuals above the .20 cutoff. There are no obvious learnings in these items that could be gained from one item that would assist in answering a latter question correctly. However, there are some similarities in terms of task style. For example, items 29 and 42 (Yen's Q of .23) both require mental shape rotation, while items 32 and 45 (Yen's Q of .22) both require moving colours within a shape. There does not appear to be enough evidence that these items are not theoretically unidimensional, and while there may be some cognitive complexity introduced to the tasks that produce residual correlations, this does not interfere with the underlying measurement of Induction ability and thus these items were retained.

4.5.8 Differential Item Functioning

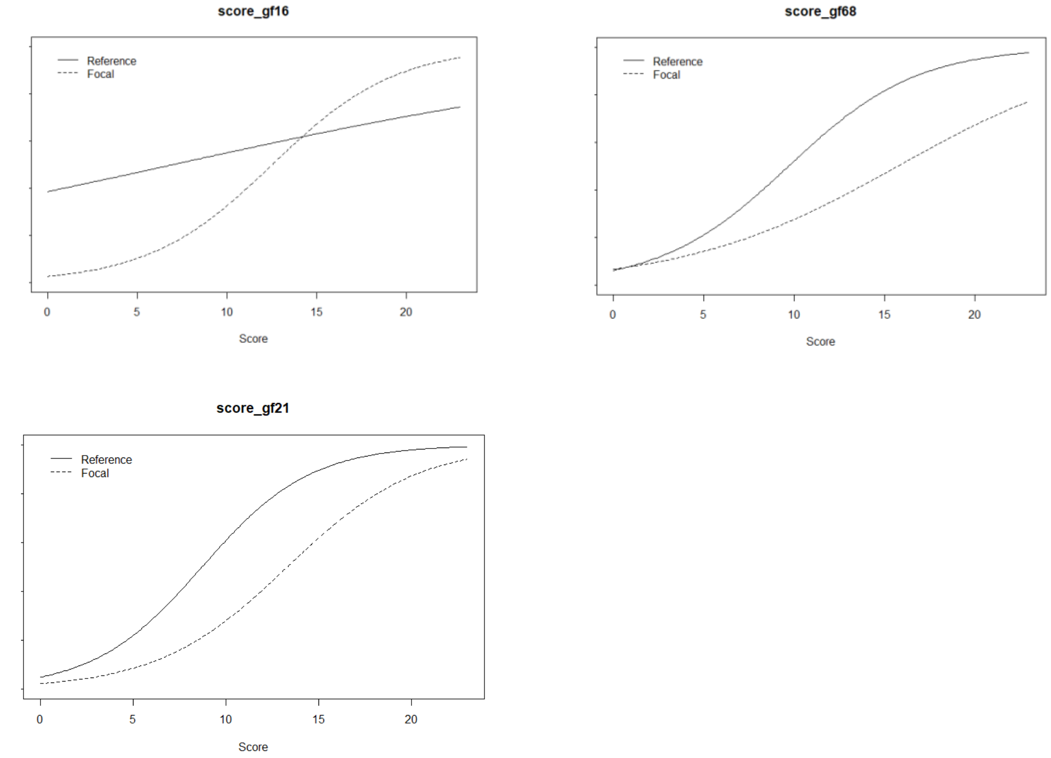
Items 68 and 16 were identified as having DIF by device, gender, and nationality, while items 21, 28, and 59 demonstrated inconsistent DIF by nationality and device. Evaluation of the performance of different groups showed that DIF results should be interpreted with caution in the current data set due to the low number of participants who identified as non-Australian or as another gender. For example, participants from Australia had an approximate 50% probability (68 of 134) of obtaining a correct answer for item 28, whereas non-Australians had an approximate 10% probability (1 of 8). This was demonstrated by significant changes in probability based on groups in the figures below. Further data is required for the Induction items to further establish true DIF.

Figure 4-71. Gf:I DIF by Device



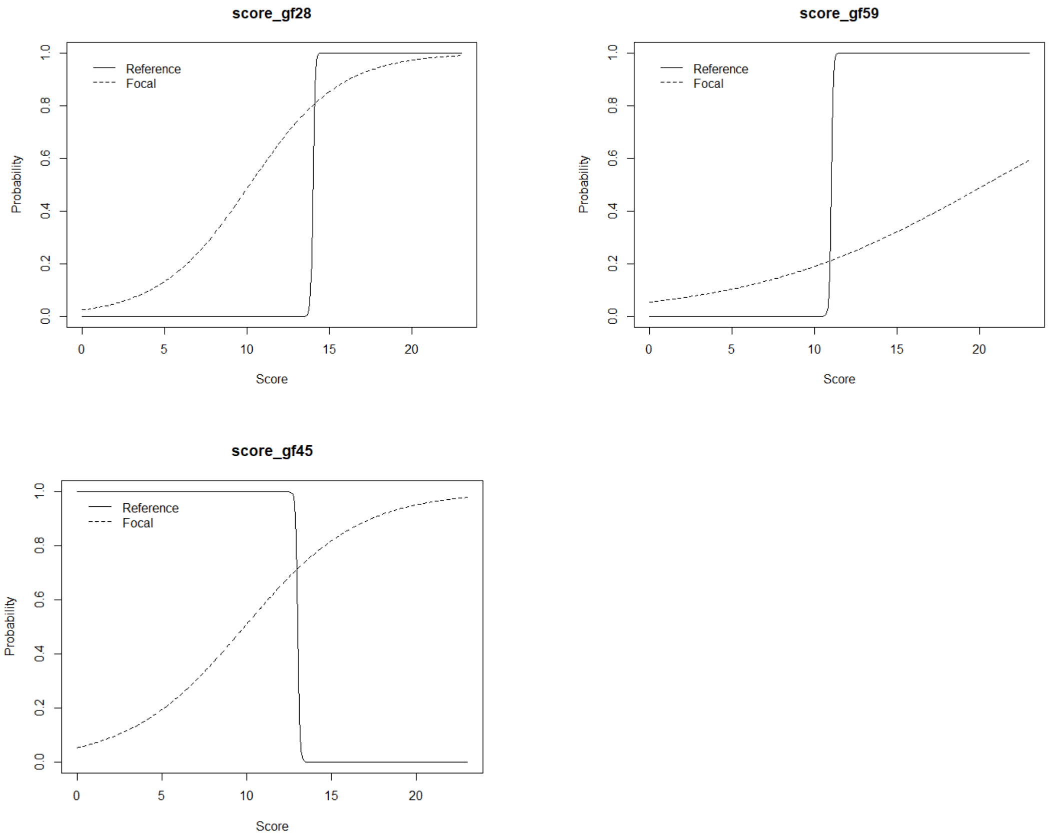
Note. Focal device was iPhone

Figure 4-72. Gf:I DIF by Gender



Note. Focal group was Male (“m”) and reference group was Female (“f”)

Figure 4-73. Gf:I DIF by Nationality



Note: Focal group was Australian and reference group was non-Australian

4.5.9 Confirmatory Factor Analysis

CFA was completed with the Induction 23-item set. The CFA converged only for imputations 1, 2, 4 and 5, with the fit indices suggesting the model had overfit due to the low sample size relative to the number of items. Only items 16, 20, 26 and 59 had factor loadings below .30; items 16, 20 and 59 had been flagged earlier in these results.

Table 4-32. CFA for Induction 23-Item Set

Imputation	χ^2	DF	RMSEA	SRMR	TLI	CFI
Imp 1	205.10	230	.00	.12	1.022	1.000
Imp 2	199.89	230	.00	.12	1.026	1.000
Imp 4	192.79	230	.00	.12	1.033	1.000
Imp 5	184.40	230	.00	.12	1.042	1.000

Note. *$p < .05$; **$p < .01$; ***$p < .001$

4.5.10 Rasch Item Parameters, ICC and Test Information

Rasch fit statistics for the 23-items reflected that of the CFA for Induction, and the Rasch modelling for Lexical Knowledge (i.e. low RMSEA, TLI, and CFI, but elevated SRMR). The final parameters were calculated after item calibration using the rules from Rubin (1987) to pool parameter and standard error estimates across the five imputed data sets (Appendix M). The items range from -3.06 θ (Item 47) to 1.42 θ (item 59) as shown in Figure 4-74. The highest point of information, and thus the point of highest reliability, is at approximately 1 θ as shown in Figure 4-75 and Figure 4-76 respectively.

Table 4-33. Rasch Scale Fit Statistics for Gf:I 23-Item Set

Data	M2	DF	RMSEA	SRMR	TLI	CFI	Marginal Rxx
Imp 1	326.78	252	.05	.10	0.92	0.92	.80
Imp 2	321.26	252	.04	.10	0.92	0.92	.80
Imp 3	316.41	252	.04	.10	0.92	0.92	.80
Imp 4	326.63	252	.05	.10	0.92	0.92	.80
Imp 5	310.74	252	.04	.10	0.93	0.93	.80

Note. *$p < .05$; **$p < .01$; ***$p < .001$

Figure 4-74. Gf:I ICS Rasch ICCs

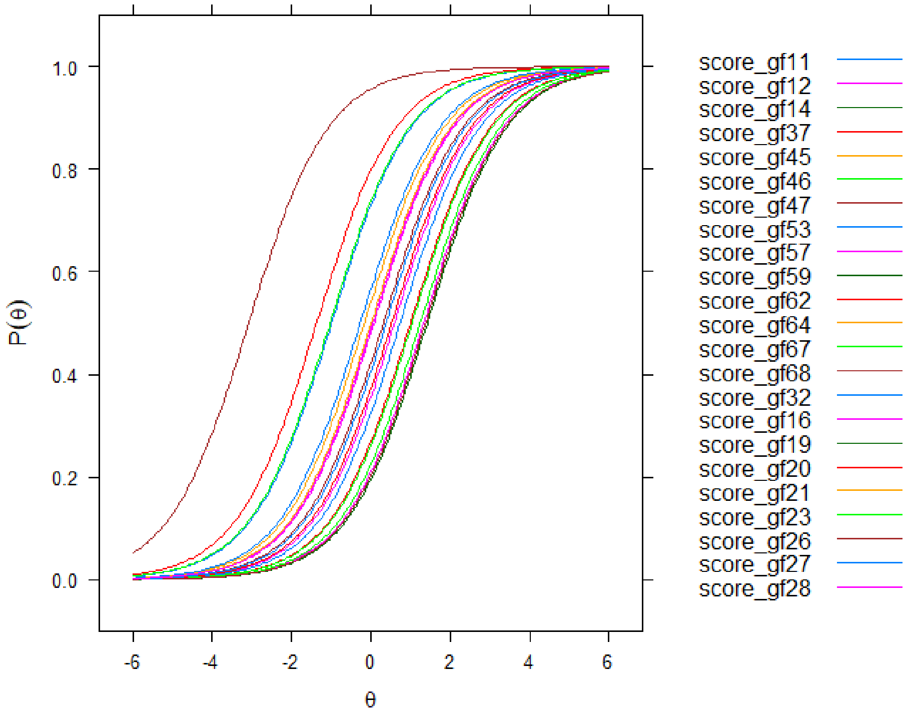


Figure 4-75. Gf:I ICS Rasch TIC

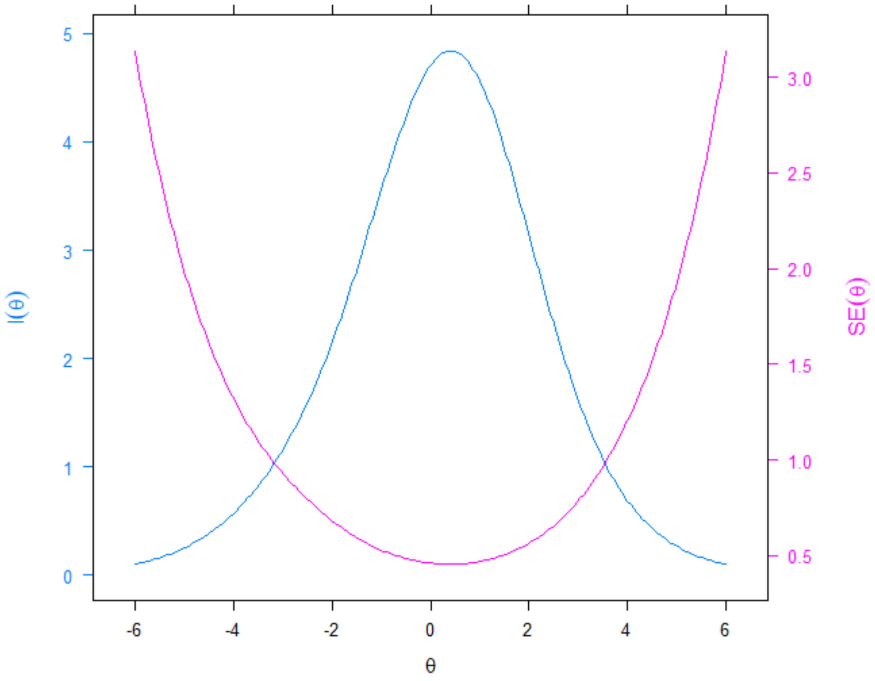
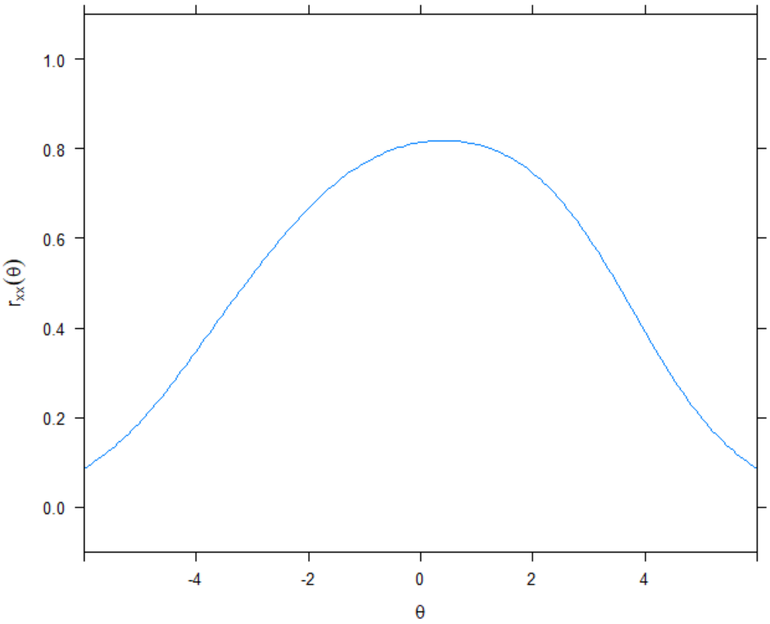


Figure 4-76. Gf:I ICS Rasch Test Reliability Curve



4.6 Visualisation Results

4.6.1 Raw Data

The descriptive statistics for time taken and total raw score for the ITOS and ICS samples are included in Table 4-34 and Table 4-35.

The frequencies of the total raw score for participants are in Figure 4-77. The maximum possible score for the ITOS sample was 52, while the maximum total score for both ICS Adult and ICS School Aged samples was 72. A majority of the ITOS participants performed within the higher ranges, however the Visualisation test performance for the ICS Adult participants was far more varied.

There was a significant difference in the total raw score of participants within each phase: *F*(2,1624) = 13.88, *p* < .001 (Figure 4-78). The total time taken for the Visualisation test showed significant difference: *F*(2,1624) = 13.46, *p* < .011 (Figure 4-79)

Performance across the Visualisation items remained varied, including within the new items (Figure 4-80).

There was no significant difference between the gender identification groups (Female, Male, Other, Prefer Not To Say) on total score, *F*(3,1623) = 1.99, *p* = .11 (Figure 4-82), or on time taken, *F*(3,1623) = 2.44, *p* = .06 (Figure 4-83).

Table 4-34. Descriptive Statistics for Raw Score with the Gv:Vz Items

	***n* Items**	***n* Participants**	**Mean**	**SD**	**Min**	**Max**
ITOS[1]	52	876	34.85	12.16	1	52
ICS Adult	72	648	32.27	18.73	1	67
ICS School Aged	72	103	39.97	8.05	1	57
All Phases	72	1627	34.15	15.07	1	67

[1]Descriptive statistics varied from the ITOS analysis due changes in missing data methodology

Table 4-35. Descriptive Statistics for Time Taken with the Gv:Vz Items

	n	**Mean**	**SD**	**Min**	**Max**
ITOS[1]	876	592.76	243.75	6.57	1424.16
ICS Adult	648	625.34	322.70	6.72	1676.52
ICS School Aged	103	737.89	216.43	4.36	1131.65
All Phases	1627	614.93	278.59	4.36	1676.52

[1]Descriptive statistics varied from the ITOS analysis due changes in missing data methodology

Figure 4-77. Gv:Vz Frequency of Participant Raw Score

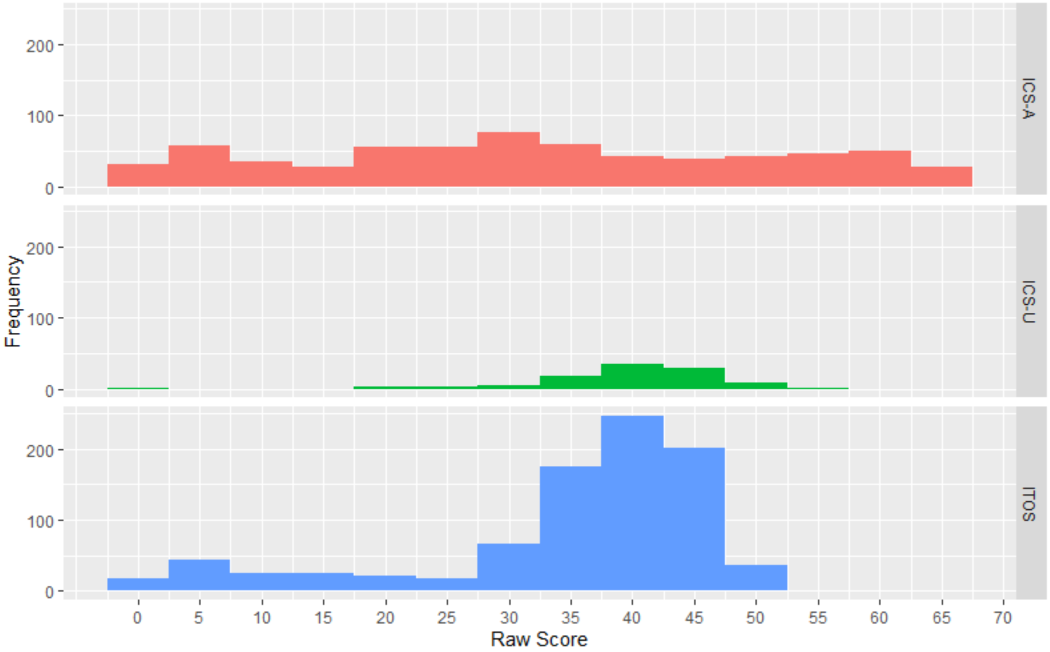


Figure 4-78. Gv:Vz Total Score by Phase

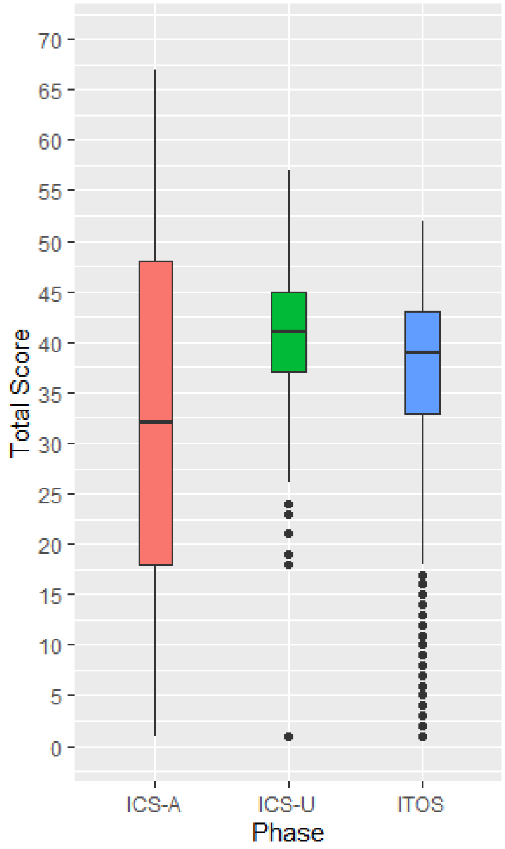


Figure 4-79. Gv:Vz Time Taken by Phase

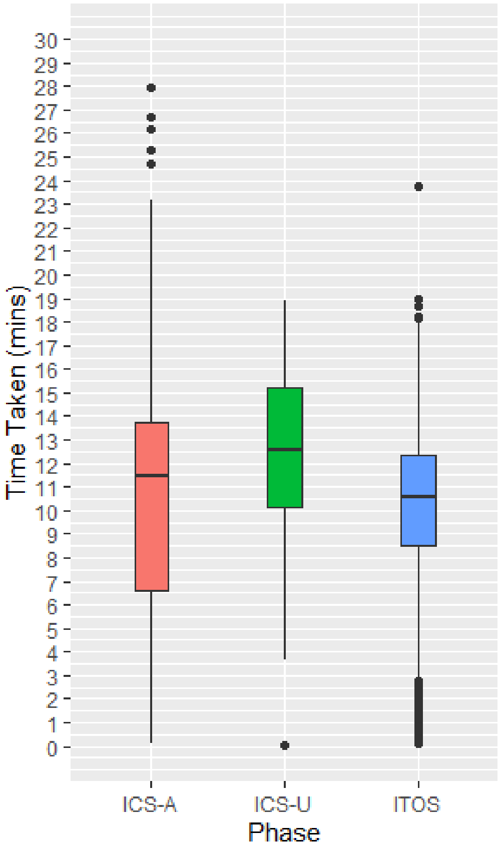


Figure 4-80. Visualisation Percentage of Items Correct

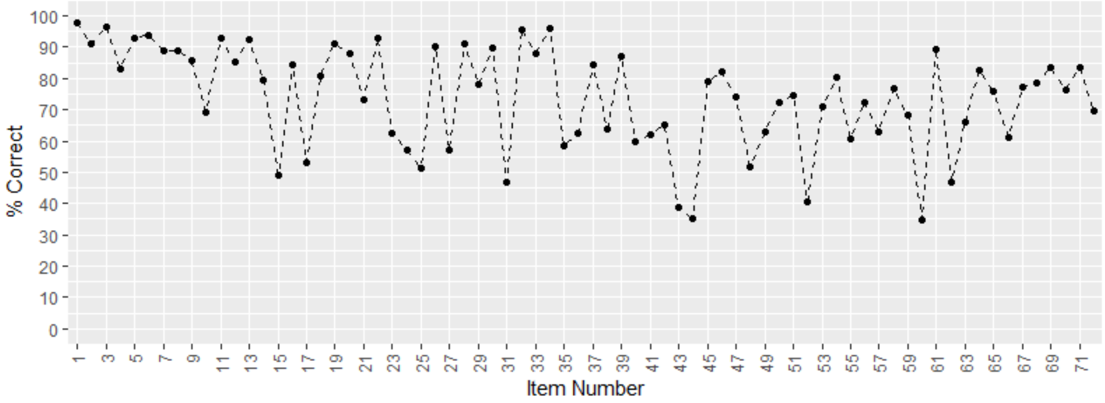


Figure 4-81. Gv:Vz Total Score by Time Taken

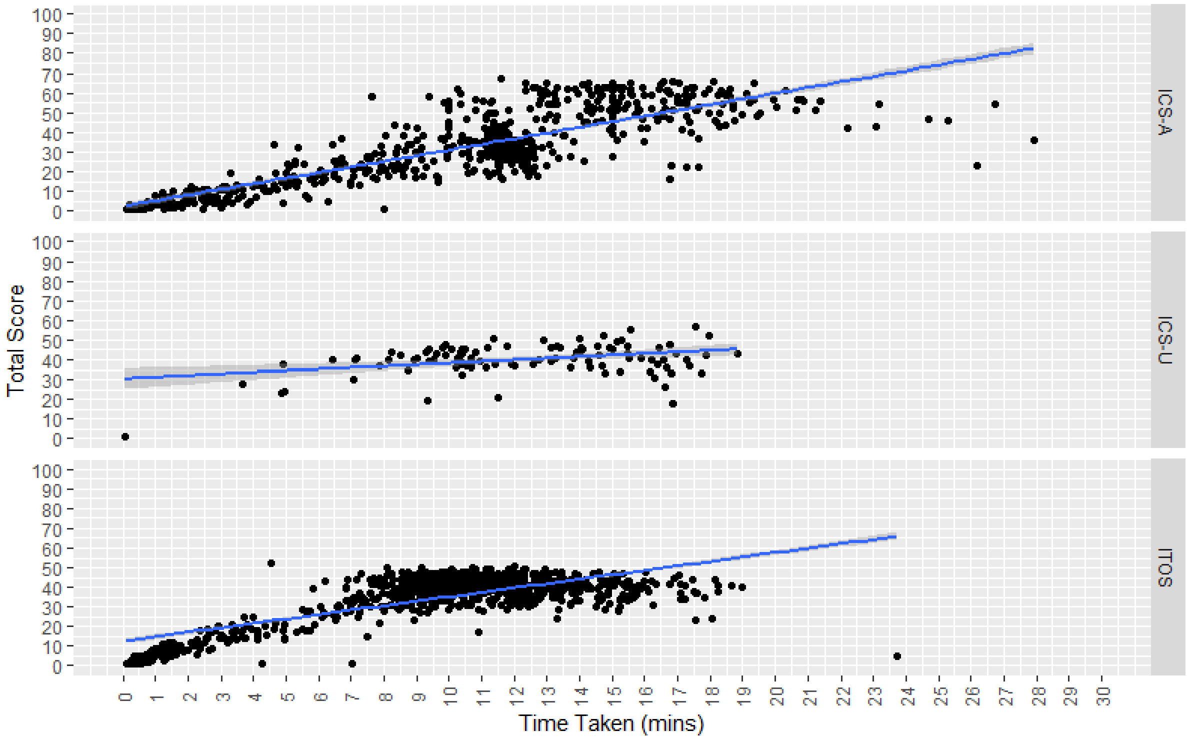


Figure 4-82. Gv:Vz Gender by Total Score

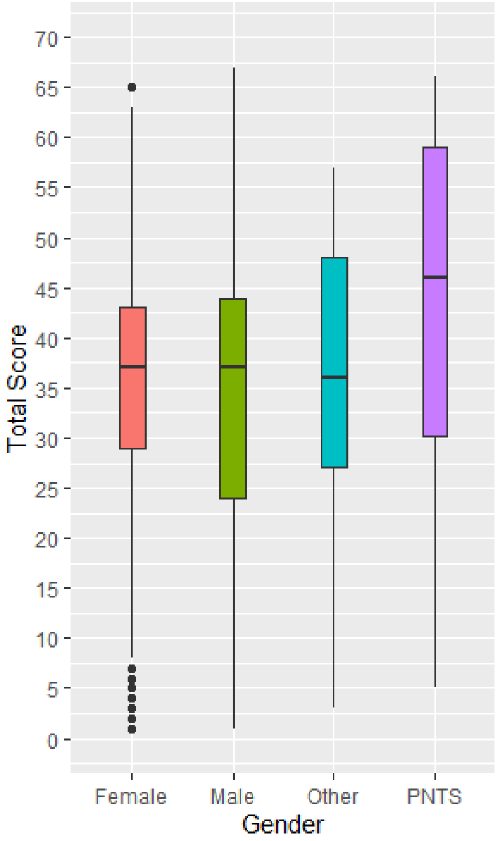
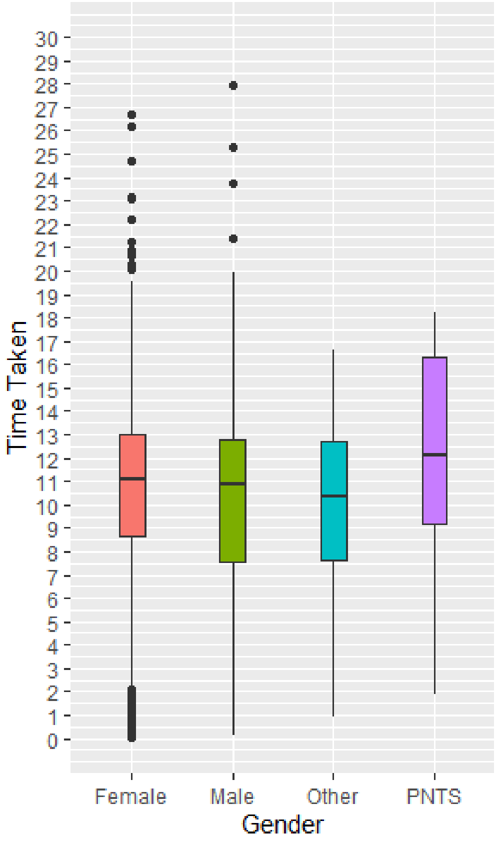


Figure 4-83. Gv:Vz Gender by Time Taken



4.6.2 Missing Data

For the ICS data collection items were administered in the order of anchor Items (21-52), Item Set B (23-72) and Item Set A (1-21) for adults, and anchor Items (21-52), Item Set A (1-21) and Item Set B (23-72) for school aged participants.

Table 4-36 and Figure 4-84 show summary statistics of the percentage of missing data within the Visualisation data set broken down by each phase of data collection. Consistent with the order of administration, missing data increases (i.e. highest missing data percentage was noted in Set A for ICS Adult and Set B for ICS School Aged).

When considering missing data percentages (Table 4-37), the higher mean and median percentage of missing data by participant for the ICS Adult sample suggests a high number of participants in this sample dropped out of the study.

Figure 4-85 shows the pattern of missingness across item sets. For each participant if an item was missing from a set, that item set was classed as incomplete. Thus, the blue cells show the frequency of those item sets that were deemed complete.

Table 4-36. Percentage of Gv:Vz Missing Data by Item

Phase	*n Items*	Mean	SD	Median	Min	Max
ITOS	52	13.24	5.44	15.30	0.23	18.84
ICS Adult	72	43.25	19.38	40.82	16.05	70.37
ICS School Aged	72	6.42	4.3	3.88	0.97	17.48

Table 4-37. Percentage of Gv:Vz Missing Data by Participant

Phase	*n* Participants	Mean	SD	Median	Min	Max
ITOS	876	13.24	28.44	0	0	98.08
ICS Adult	648	43.25	34.56	45.83	0	100[1]
ICS School Aged	103	6.42	15.83	0	0	98.61

Note. Missing data for this table is calculated based on the items available to each participant. [1]39 participants achieved correct scores on the training items but then did not proceed to the test items.

Figure 4-84. Percentage of Gv:Vz Missing Data by Phase

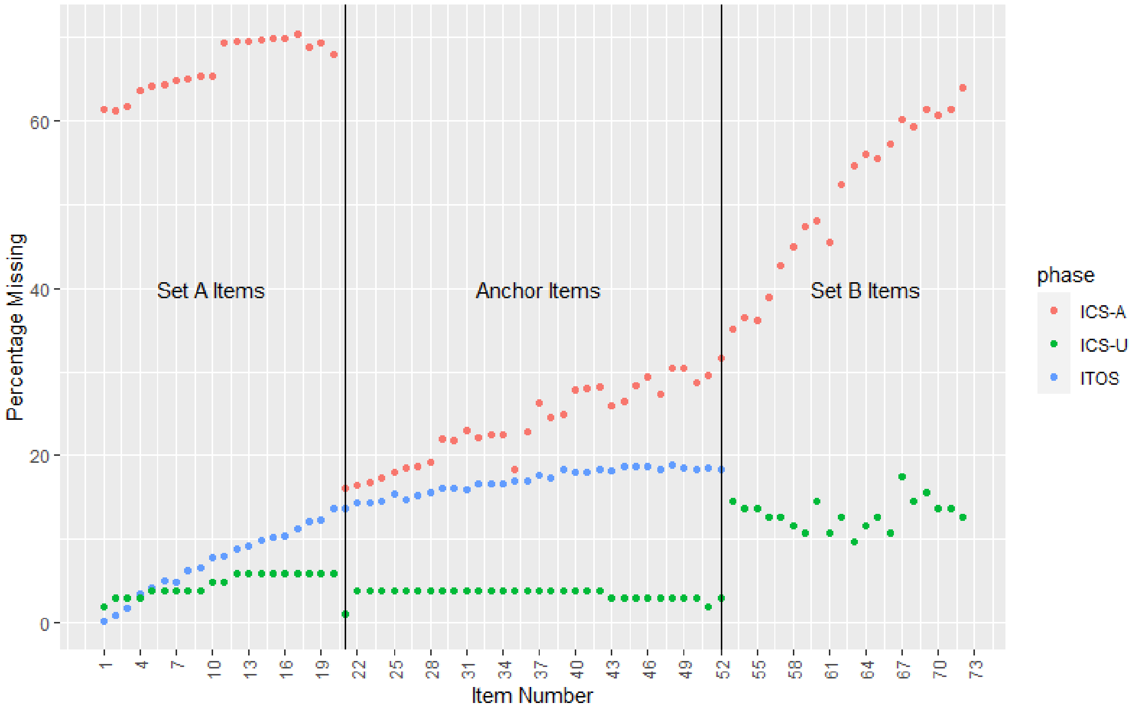
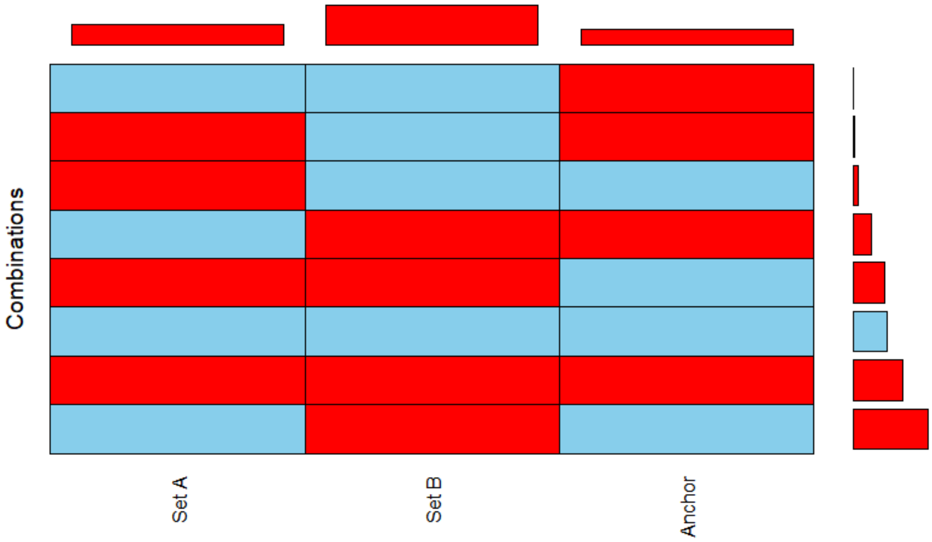


Figure 4-85. Missing Data Pattern for Gv:Vz Sets



Note. Red cells indicate missing data while blue cells indicate present data. The bars above and beside the cells represent the proportion of participants with that pattern of data.

4.6.3 Data Imputation

Like previous narrow abilities analysed, the data imputation will consider the dataset in entirety. Data was prepared for MICE in stages as detailed in the methodology and demonstrated in Table 4-38. This had some implications for the interpretation of missing data compared to the previous section (4.6.2). For example, in the missing data analysis section, a participant was not counted as having missing data if they were in the ITOS sample and completed every Visualisation item between items 1 and 52. In this part of the analysis however, they were considered as missing data for items 53 through 72 (Figure 4-86). Unlike for the Lexical Knowledge item set, no items were removed for the Visualisation items after the ITOS analysis, and thus the most significant drop in participants was caused by removing participants who did not have a full set of anchor Items (Figure 4-87). The pattern of missingness based on item sets after these alterations is represented in Figure 4-88.

After the data cleaning above, no items were flagged as problematic for the imputation model. Items 53 to 72 were identified as having an outflux below .30 suggesting imputation for these items may be difficult. As shown in Figure 4-89 a number of these items had a low outflux value meaning their observed data did not connect well to the missing data on other variables. Variables in the top left have more predictive power, whereas variables that come closer to the bottom right are overly reliant on the imputation model. Items noted above were removed prior to MICE.

MICE was successfully carried out with 111 individual formulas executed to impute data for the remaining 111 variables. No problems were identified with the imputation. Figure 4-91, Figure 4-92, Figure 4-93 and Figure 4-94 show comparisons between the data before and after imputation.

Table 4-38. ICS Gv:Vz Missing Data Cleaning

Phase	*n Items*	Mean	SD	Median	Min	Max
Initial Missing Data	76	39.83	24.98	28.64	13.83	80.15
Full Anchor Items	76	26.68	30.04	18.37	0	82.81
Removed ITOS Items[1]	76	26.68	30.04	18.37	0	82.81
Low Response Items[2]	72	23.58	27.70	17.83	0	26.43
High Missingness	72	19.34	29.08	2.53	0	67.38
Missing Age	72	19.37	29.11	2.54	0	67.46
Time Outliers	72	20.06	30.07	2.71	0	69.68
Score Outliers	72	20.17	30.33	2.54	0	70.05
Multivariate Outliers	72	20.58	32.14	0.79	0	72.35

Figure 4-86. Gv:Vz Missing Data for Data Imputation Preparation

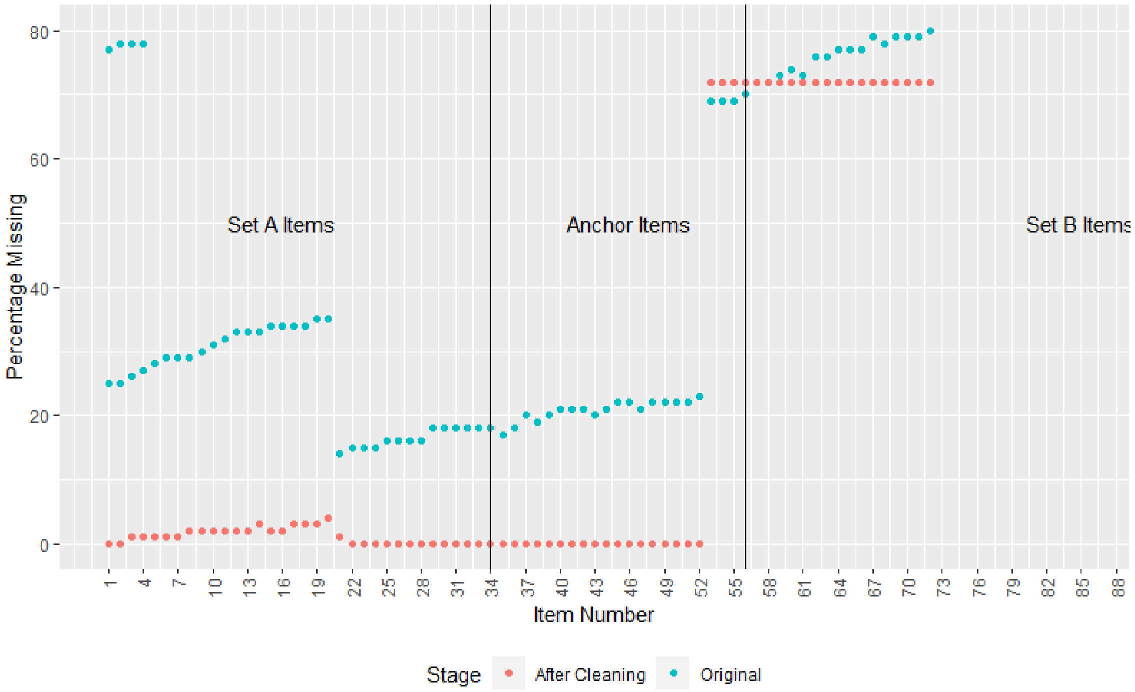
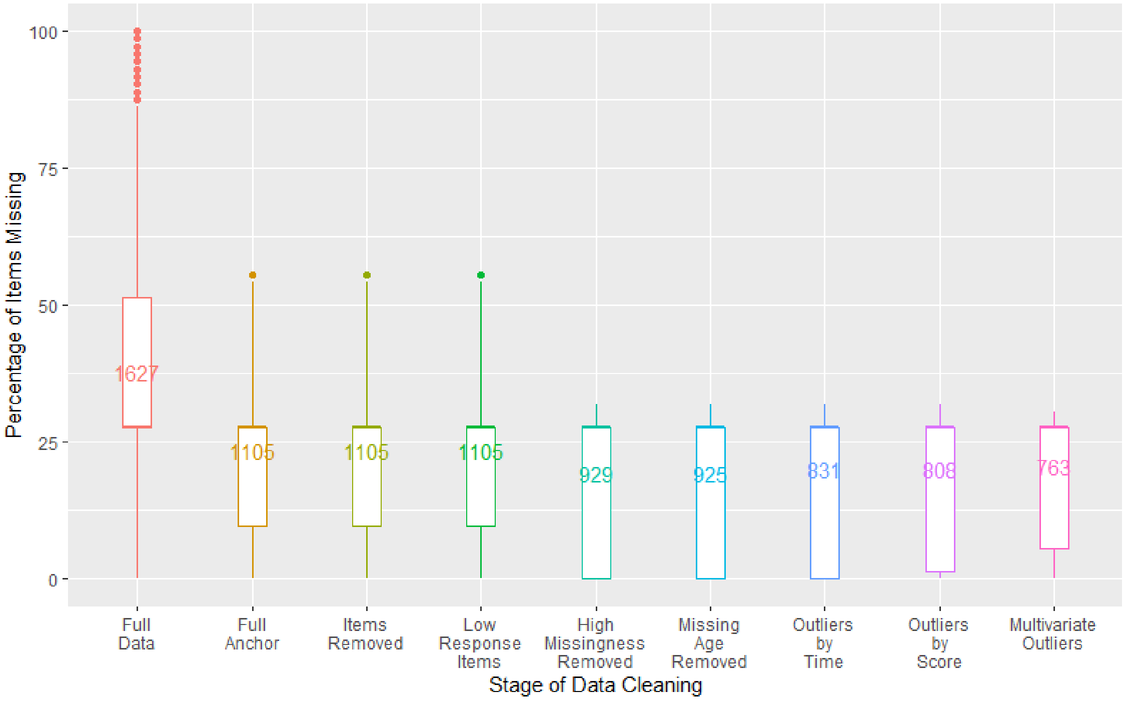
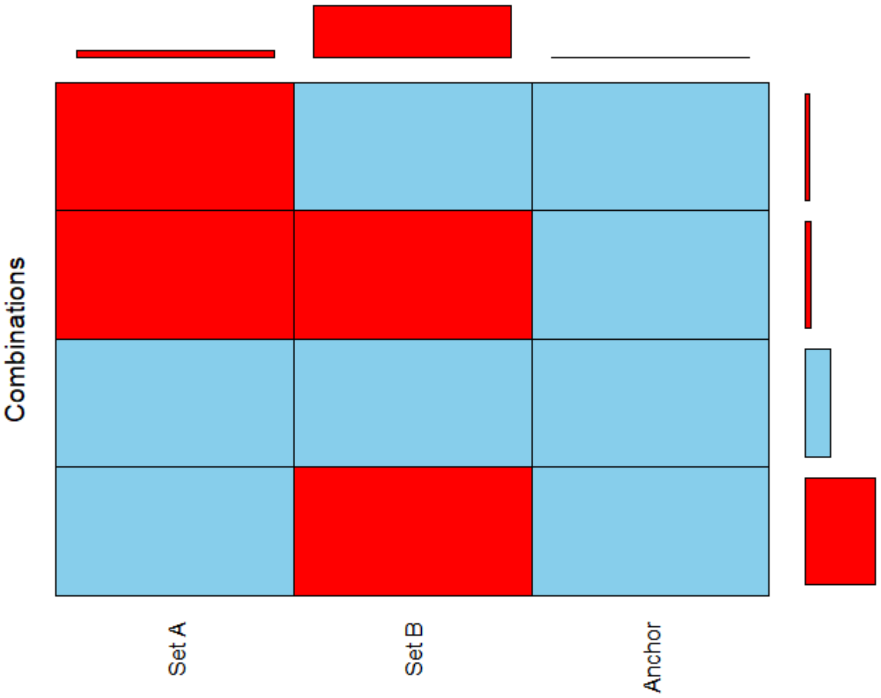


Figure 4-87. Percentage of Missing Gv:Vz Items by Participant



Note. No items were removed from the Visualisation items after ITOS, but step is retained for consistency

Figure 4-88. Missing Data Pattern for Gv:Vz with Full Anchor Items



Note. Red cells indicate missing data while blue cells indicate present data. The bars above and beside the cells represent the proportion of participants with that pattern of data.

Figure 4-89. Outflux-Influx pattern before removal

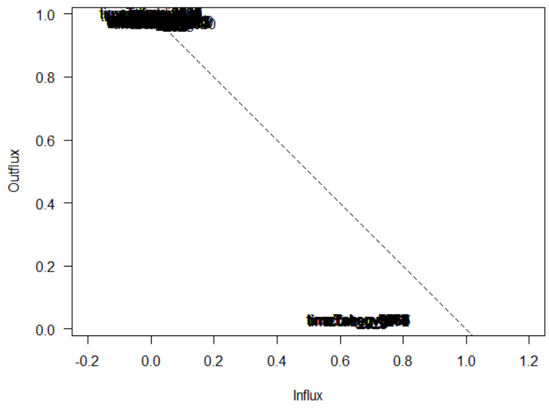
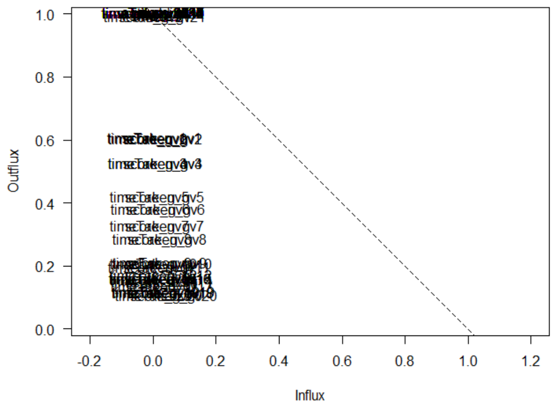


Figure 4-90. Outflux-Influx pattern after removal



Note. Items with low outflux have poor predictive power for the imputation model and items with high influx have high reliance on the imputation model. Ideally you have high outflux and low influx.

Figure 4-91. Gv:Vz Score by Imputation

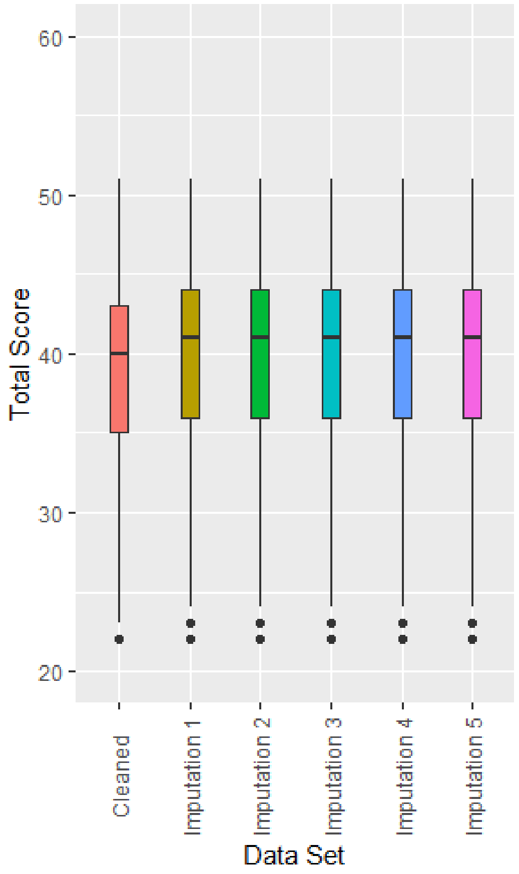


Figure 4-92. Gv:Vz Total Time by Imputation

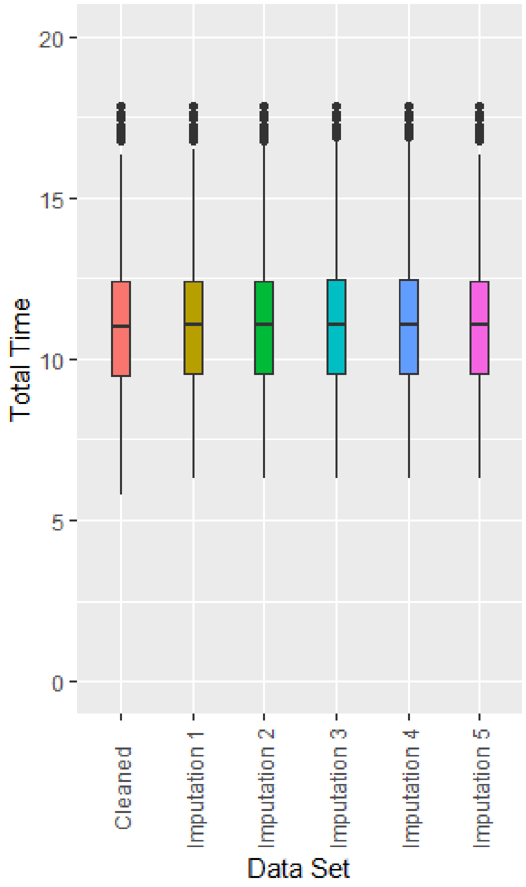


Figure 4-93. Gv:Vz ICS Percentage of Items Correct

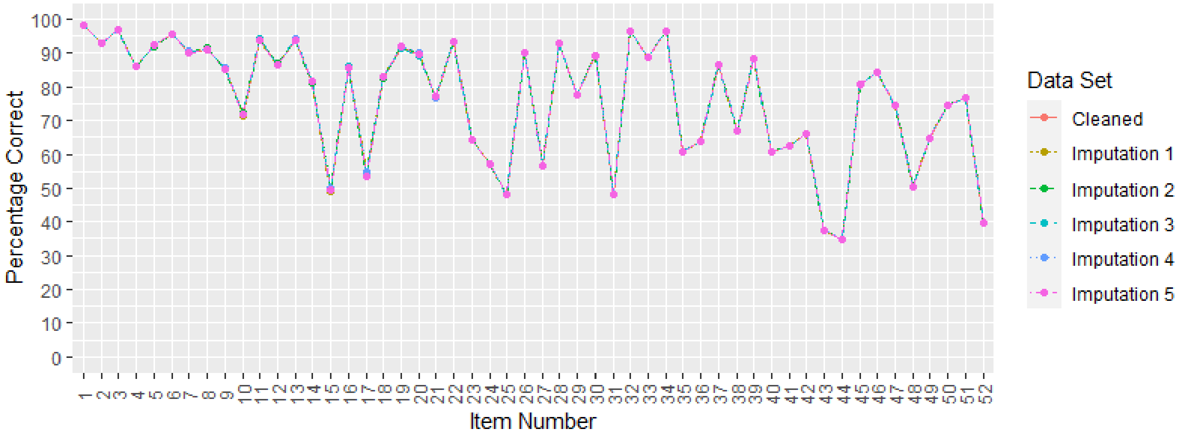
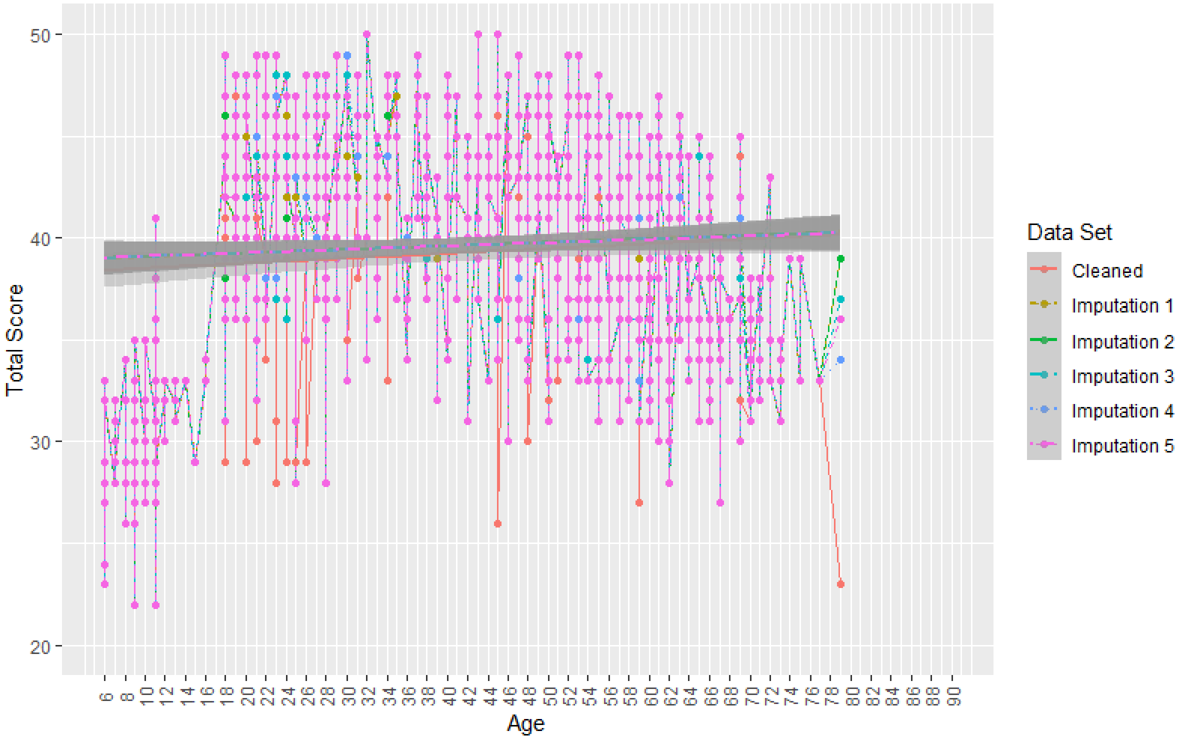


Figure 4-94. Gv:Vz Total Score by Age



4.6.4 Reliability

Cronbach's alpha for the Visualisation items in the ICS was calculated across each imputed data set and was found to have a mean Cronbach's alpha of .76. This was lower than in the ITOS, likely reflecting an increased diversity in item difficulty. Regardless, this level of reliability can be classified as good.

4.6.5 Rasch Analysis

Rasch analysis was conducted on all five imputed data sets. For the 52-item set there was poor fit with the Rasch model. Item fit details are in Appendix N.

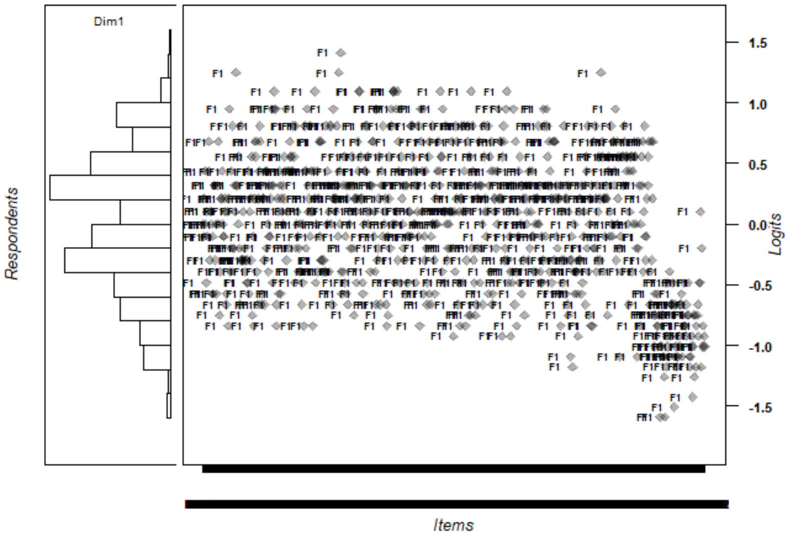
Person fit statistics were calculated, and 43 participants were found to have Zh values above +2 and below -2. Analysis of these participants identified these were 32 participants aged 6-18, nine participants above age 50, and two participants aged between 20 and 30. As with the Lexical Knowledge section, this may be a result of the wide age range tested in this study, as demonstrated in Figure 4-95.

Table 4-39. Rasch Scale Fit Statistics for Gv:Vz 52-Item Set

Data	M2	DF	RMSEA	SRMR	TLI	CFI	Marginal Rxx
Imp 1	5082.36	1325	.06	.09	0.60	0.61	.86
Imp 2	5084.35	1325	.06	.09	0.61	0.61	.86
Imp 3	5097.05	1325	.06	.09	0.61	0.61	.86
Imp 4	5094.20	1325	.06	.09	0.61	0.61	.86
Imp 5	5074.63	1325	.06	.09	0.61	0.61	.86

Note. **p* < .05; ***p* < .01; ****p* < .001

Figure 4-95. Gv:Vz 52-item Wright Map



For the 52-item set, 20 items were flagged for poor item fit. There was no clear pattern as to why these items didn't fit well, as they differed in number of shapes, internal cues, similarity of shapes, and whether they were old or new items.

Items 23, 44, and 52 showed the highest misfit and evaluation showed they were all low performing items and without a border. Evaluation of the item stimuli does show if determining whether the shape parts are rotated together to make the complete shape is somewhat ambiguous and for many participants it was likely no more than a guess given a dichotomous option of 'same' or 'different'. These items were removed.

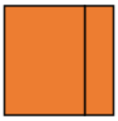
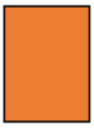
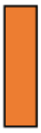
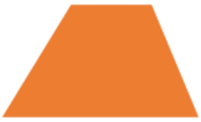
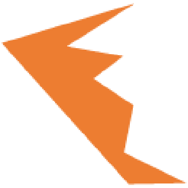
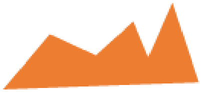
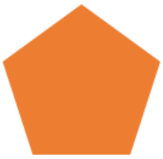
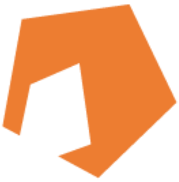
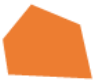
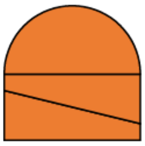
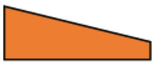
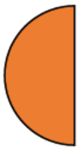
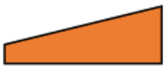
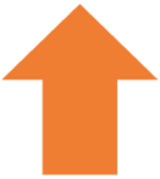
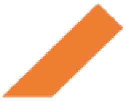
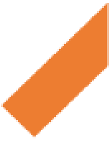
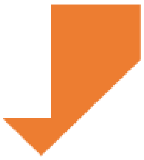
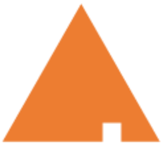
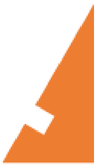
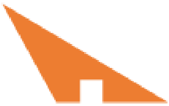
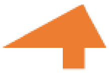
Items 3, 5 and 12, while quite easy, were generally only scored incorrectly by school aged participants. A few adults also obtained incorrect answers, suggesting the items' low difficulty at face value led to low attention and subsequently inconsistent scores. This can only be addressed with further proctored. These items were retained.

Items 10 and 13 demonstrated an inconsistent number of incorrect responses by adults despite their qualitatively evaluated ease. These items fit poorly with the Rasch model and were removed from this analysis but should be included in future studies.

Item 17 appears ambiguous due to the length of lines not being particularly obvious, so it is reasonable to infer many participants were effectively guessing; this item was removed. Such lines are more obvious in item 19, reinforced by the fact that generally only children obtained an incorrect answer here. This item was retained as it does appear to be measuring implementation of the rules and mental.

A high number of adults answered item 25 incorrectly, possibly due to difficulty of rotating the shape to fit in the un-bordered missing section. This item was retained. Item 27 displayed similar results, with a high number of adults answering incorrectly by not applying the rules of the task appropriately (i.e. no flipping) and was also retained.

Table 4-40. Gv:Vz Items

#	Complete Shape	Shape Parts		
1				
23				
25				
27				
44				
52				

Items 35 (3 response options, internal cues), 36 (3 response options, internal cues), 42 (3 response options, no internal cues), 43 (3 response options, no internal cues), 47 (3 response options, no internal cues), 48 (3 response options, no internal cues), 50 (3 response options, no internal cues) and 51 (3 response options, no internal cues) were also s misfitting and are more difficult items. Some of these were retained to ensure enough items at the difficult end of the spectrum. After evaluating each item and the item fits, items 35, 43, 48, and 50 appeared most ambiguous (causing guessing), had poor fit statistics and were removed. These calibrations resulted in a 42-item set.

Table 4-41. Rasch Scale Fit Statistics for Gv:Vz 42-Item Set

Data	M2	DF	RMSEA	SRMR	TLI	CFI	Marginal Rxx
Imp 1	2527.16	860	.05	.08	0.77	0.77	.81
Imp 2	2527.77	860	.05	.08	0.77	0.77	.81
Imp 3	2537.16	860	.05	.08	0.77	0.77	.81
Imp 4	2539.26	860	.05	.08	0.77	0.77	.81
Imp 5	2529.25	860	.05	.08	0.77	0.77	.81

Note. *$p < .05$; **$p < .01$; ***$p < .001$

4.6.6 Mokken Analysis

Mokken analysis was completed with the 42-item set. Unsurprisingly given the noted difficulty increase in some items in the Rasch analysis, all 42 items were identified as having a Loevinger's *H* value below .30. This suggests poor ordering in the items in terms of probability of obtaining correct answers; this may be caused by people with low Visualisation guessing answers on the test, or by other constructs not intended to be measured by the items. Twelve items were identified as having particularly poor Loevinger's *H* values and no qualitative reason was established as to why; for example, item 1 had a Loevinger's *H* of .018 despite quite clearly joining and matching the shape to the left. Items in this set that were not evaluated earlier were removed, resulting in a 31-item scale with poor Mokken scale fit.

Table 4-42. Mokken Analysis for Gv:Vz 42-Item Set

Data Set	Loevinger's *H*	Standard Error
42-items		
Imputation 1	.14	.01
Imputation 2	.14	.01
Imputation 3	.14	.01
Imputation 4	.14	.01
Imputation 5	.14	.01
31 items		
Imputation 1	.18	.01
Imputation 2	.19	.01
Imputation 3	.19	.01
Imputation 4	.18	.01
Imputation 5	.19	.01

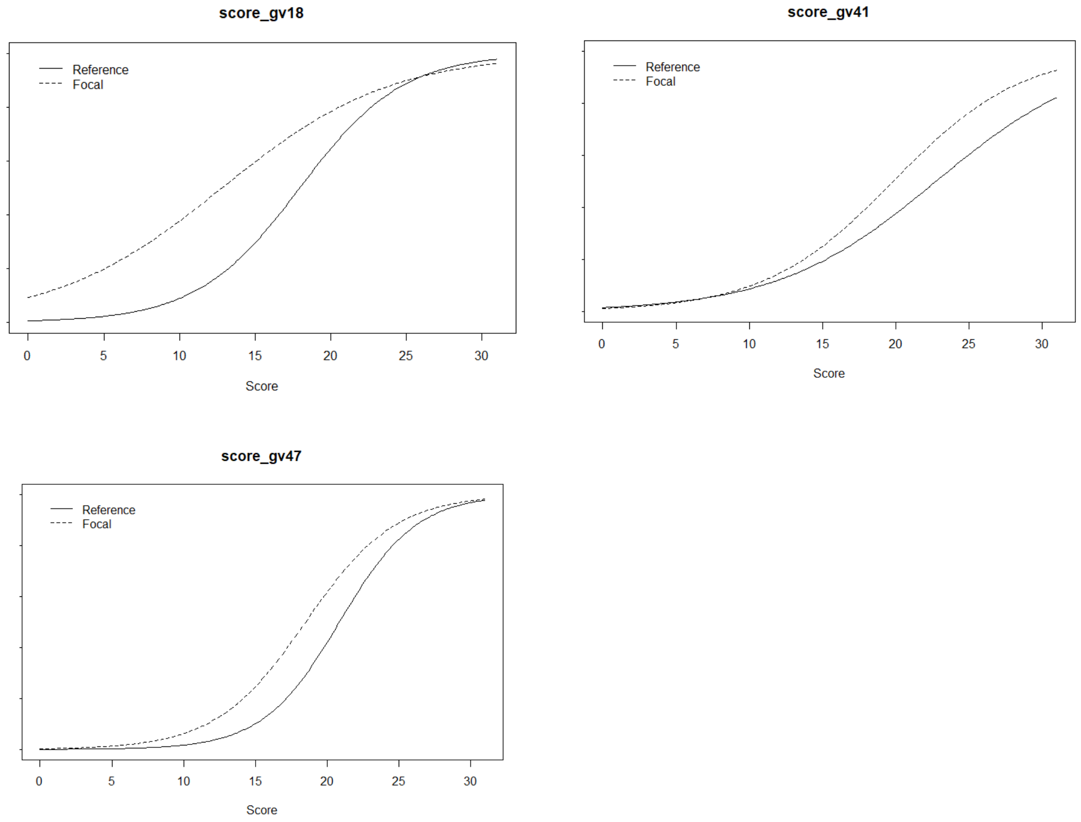
4.6.7 Local Independence

In order to be able to rotate three shapes in one's mind, one must also be able to rotate two shapes. As such, it was unsurprising that items 4, 5, 6, 8, 19, 27, 36, 47, and 50 were all flagged for local dependency. There was no clear relationship between these items outside the reliance on shape similarity, shape rotation, and internal cues to vary the difficulty. No strong locally dependent items were identified and thus all items were retained.

4.6.8 Differential Item Functioning

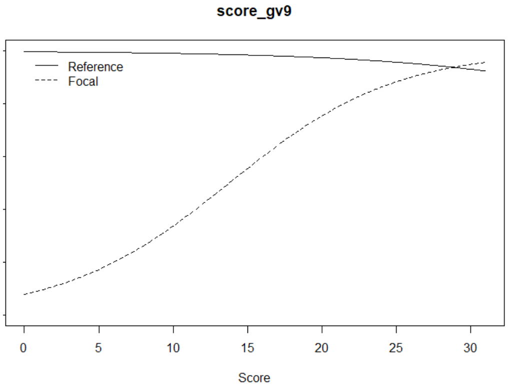
DIF was identified by gender for items 18, 41 and 47. No qualitative reason was identified and thus the items were retained. Item 9 possessed DIF for non-Australians, but this was likely a result of only 6 (out of 108) non-Australian participants answering incorrectly. Thus, this item was not removed. No DIF was identified for device.

Figure 4-96. Gv:Vz DIF by Gender



Note. Focal group was Male (“m”) and reference group was Female (“f”)

Figure 4-97. Gv:Vz DIF by Nationality



Note. Focal group was Australian and reference group was Non-Australian.

4.6.8.1 Confirmatory Factor Analysis

CFA was completed with the Visualisation 31-item set. The CFA converged only for imputations 1, 3 and 5 and a generally acceptable fit was found. As with the other item sets a higher SRMR was identified suggesting correlations of residuals. Only item 25 had a negative loading under .30 and was removed.

Table 4-43. CFA for Gv:Vz 31-Item Set

Imputation	χ^2	DF	RMSEA	SRMR	TLI	CFI
Imp 1	855.08***	434	.04	.11	.931	.936
Imp 3	857.58***	434	.04	.11	.932	.936
Imp 5	850.83***	434	.04	.10	.933	.937

Note. *p < .05; **p < .01; ***p < .001

4.6.9 Rasch Item Parameters, ICC, and Test Information

Rasch fit statistics (Table 4-44) for the 30 items demonstrated improved Rasch fit. RMSEA and SRMR were acceptable and TLI and CFI close to acceptable. The final parameters were calculated after item calibration using the rules from Rubin (1987) to pool parameter and standard error estimates across the five imputed data sets (Appendix O). The items range from -3.89 θ (item 3) to .09 θ (item 31) as shown in Figure 4-98. The highest point of information, and thus the point of highest reliability is at approximately -2 θ as shown in Figure 4-99 and Figure 4-100, respectively.

Table 4-44. Rasch Scale Fit Statistics for Gv:Vz 30-Item Set

Data	M2	DF	RMSEA	SRMR	TLI	CFI	Marginal Rxx
Imp 1	1183.71***	434	.05	.07	0.88	0.88	.76
Imp 2	1188.24***	434	.05	.07	0.88	0.88	.76
Imp 3	1194.63***	434	.05	.07	0.87	0.88	.76
Imp 4	1196.73***	434	.05	.07	0.87	0.87	.76
Imp 5	1185.43***	434	.05	.07	0.88	0.88	.76

Note. *p < .05; **p < .01; ***p < .001

Figure 4-98. Gv:Vz ICS Rasch ICCs

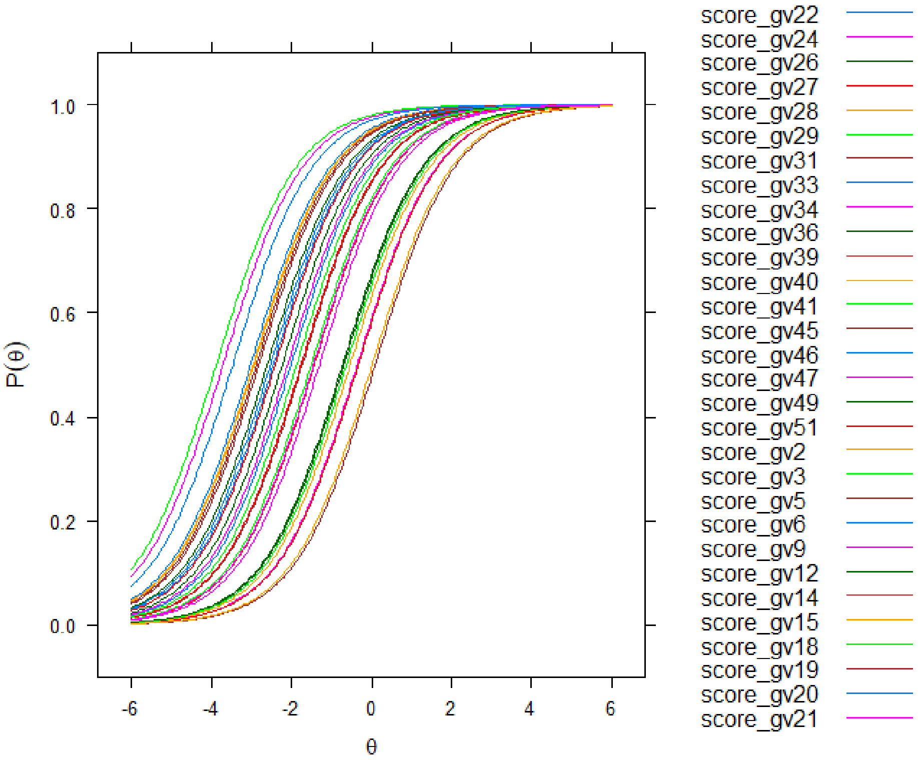


Figure 4-99. Gv:Vz ICS Rasch TIC

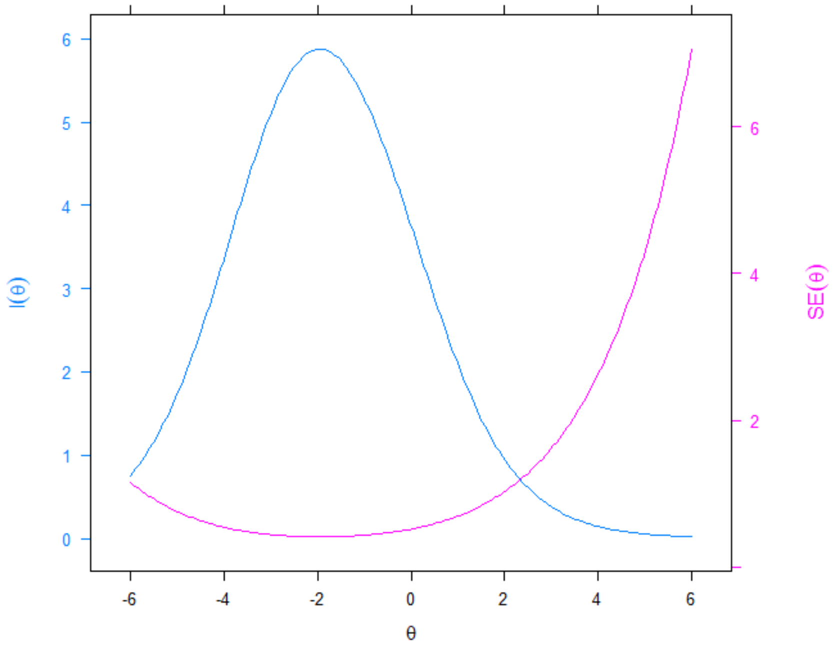
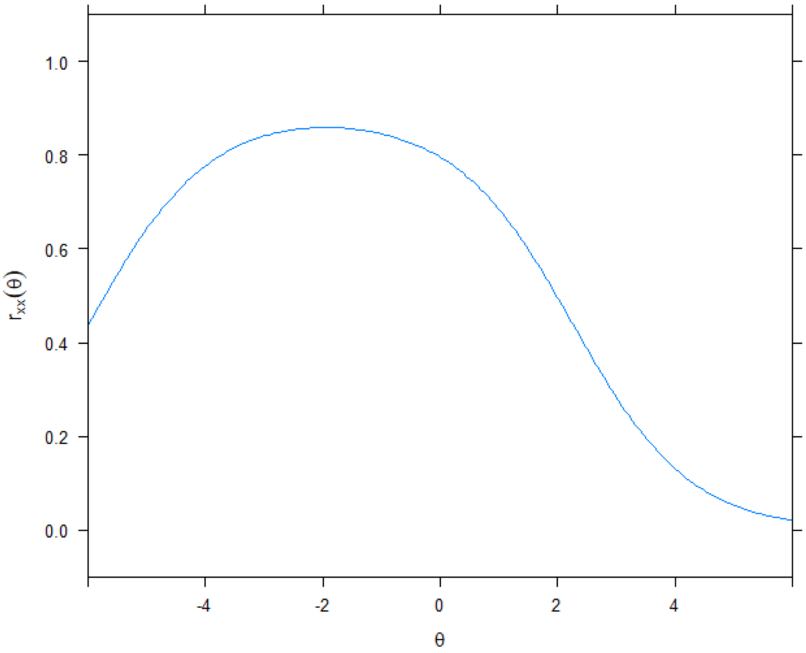


Figure 4-100. Gv:Vz ICS Rasch Test Reliability Curve



4.7 Working Memory Results

4.7.1 Raw Data

The descriptive statistics for time taken and total raw score for the ITOS and ICS samples are included in Table 4-45 and Table 4-46.

The frequencies of the total raw scores for participants are in Figure 4-101. The maximum possible score for the ITOS sample was 38, while the maximum total score for the ICS Adult and ICS School Aged samples was 34.

There was a significant difference in the raw score of participants within each phase: $F(2,921) = 508.9$, $p < .001$ (Figure 4-102). The total time taken for the Working Memory test was significantly different: $F(2,921) = 153.8$, $p < .001$ (Figure 4-103).

Performance across Working Memory showed a relatively linear drop as more pieces of information were required (Figure 4-104). Many participants dropped out of the Working Memory test quite early in the ICS Adult sample (Figure 4-105), which also shows a somewhat linear relationship between time taken and higher scores.

There was no significant difference between gender identification groups on total score, *F*(3,920) = 1.41, *p* = .24, or time taken, *F*(3,920) = 1.44, *p* = .23. This is illustrated in Figure 4-106 and Figure 4-107.

Table 4-45. Descriptive Statistics for Raw Score with the Gwm:Wc Items

	n **Items**	*n* **Participants**	**Mean**	**SD**	**Min**	**Max**
ITOS[1]	38	521	24.60	7.37	1	38
ICS Adult	34	315	11.19	5.59	1	27
ICS School Aged	34	88	9.33	4.05	2	21
All Phases	44	924	18.58	9.49	1	38

[1]Descriptive statistics varied from the ITOS analysis due changes in missing data methodology

Table 4-46. Descriptive Statistics for Time Taken with the Gwm:Wc Items

	n	**Mean**	**SD**	**Min**	**Max**
ITOS[1]	521	703.13	241.50	12.33	1480.00
ICS Adult	315	408.43	276.26	15.18	1146.01
ICS School Aged	88	755.9	188.94	182.13	910.96
All Phases	924	607.64	288.02	12.33	1480.00

[1]Descriptive statistics varied from the ITOS analysis due changes in missing data methodology

Figure 4-101. Gwm:Wc Frequency of Participant Raw Score

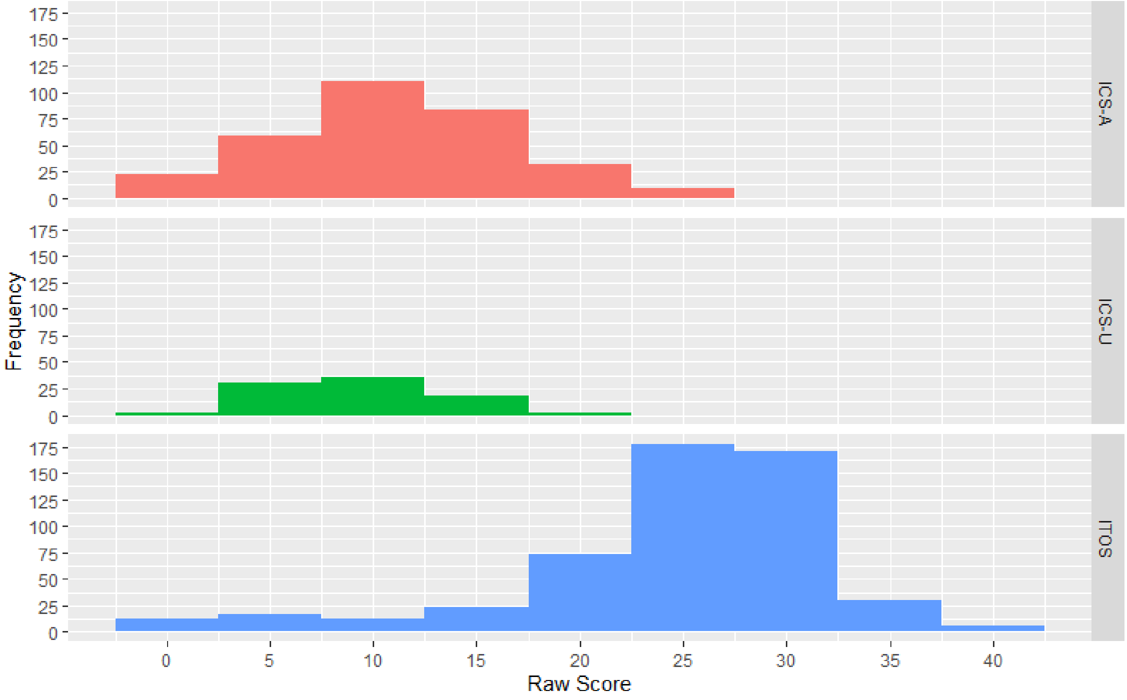


Figure 4-102. Gwm:Wc Score by Phase

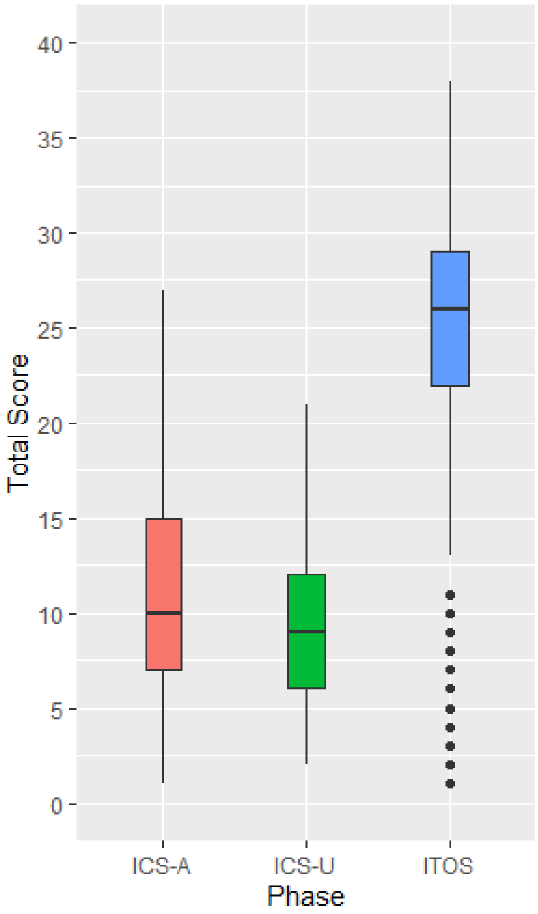


Figure 4-103. Gwm:Wc Time by Phase

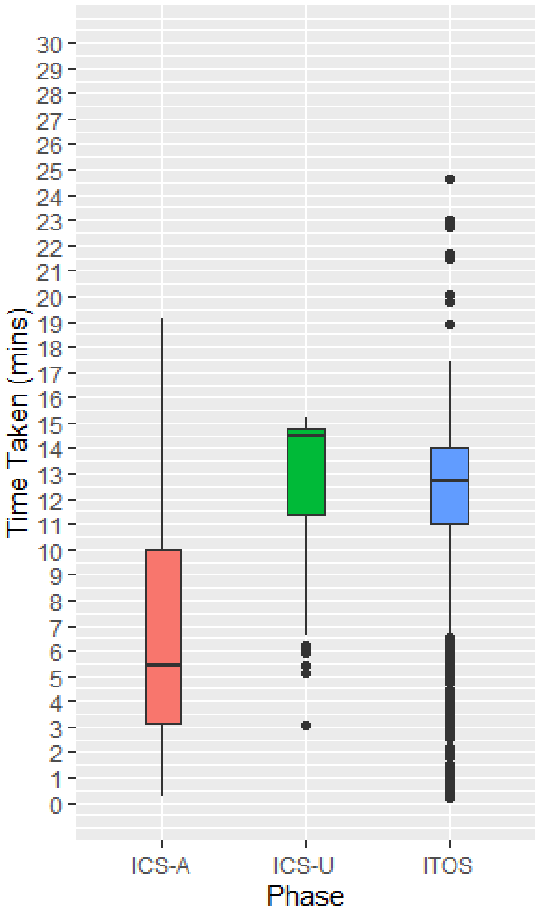


Figure 4-104. Gwm:Wc Percentage of Items Correct

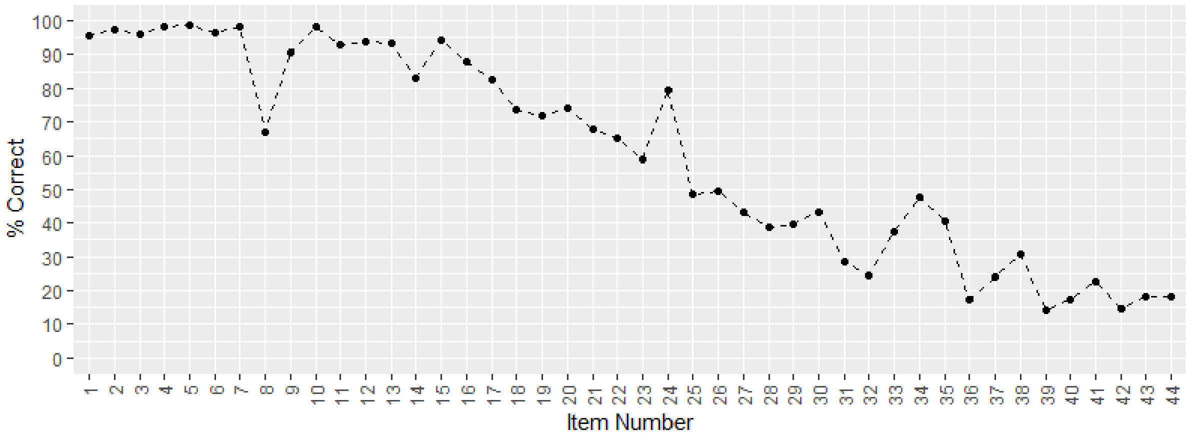


Figure 4-105. Gwm:Wc Total Score by Time Taken

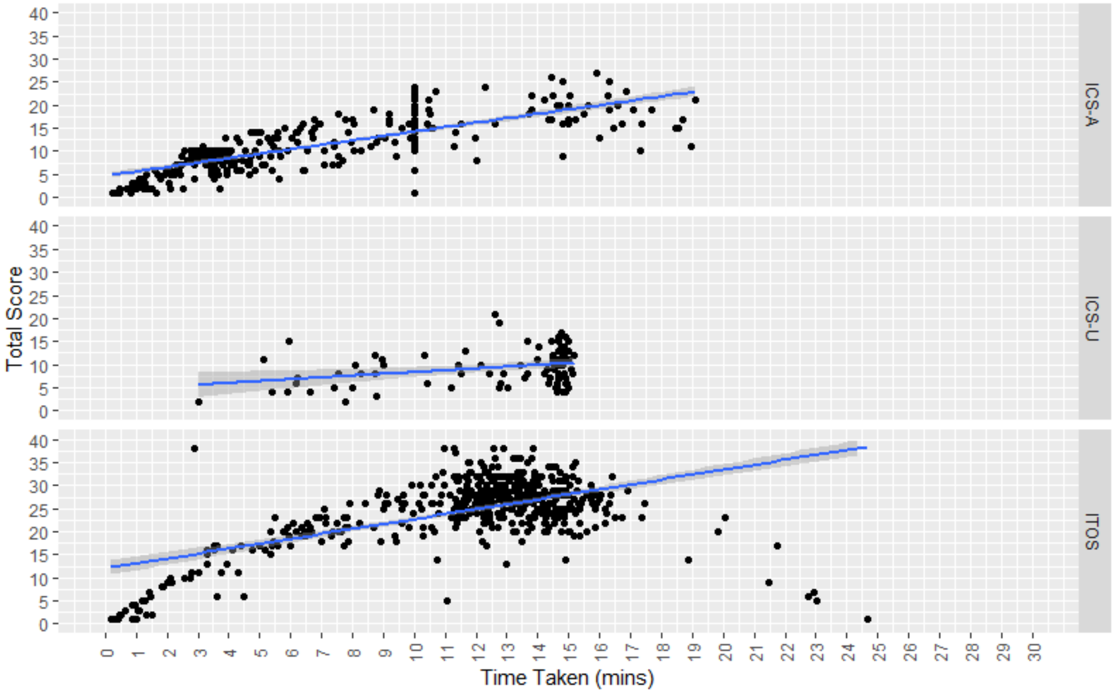


Figure 4-106. Gwm:Wc Gender by Score

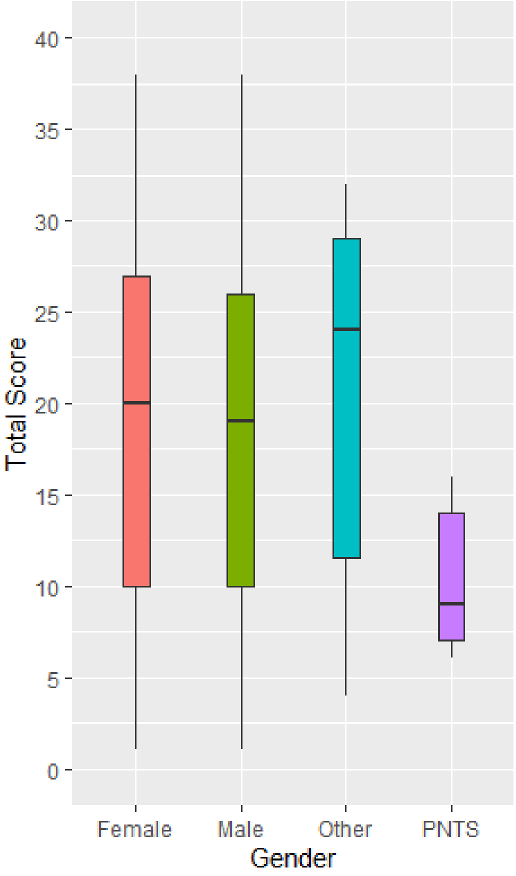
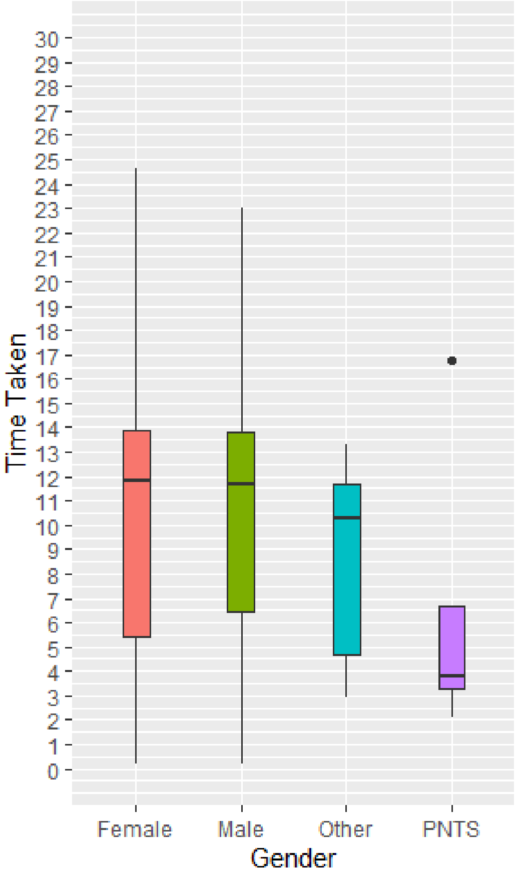


Figure 4-107. Gwm:Wc Gender by Time



4.7.2 Missing Data

Given the nature of working memory (i.e. you must be able to remember two chunks in order to remember three, which is in turn required to remember 4 parts), new items developed for the ICS (39-44) were added to the end of the ITOS items (1-38). Even though an anchor set was defined, it was not administered first for either the school aged or adult samples (i.e. items were administered sequentially for all samples).

Table 4-47 and Figure 4-108 show summary statistics for the percentage of missing data within the data set broken down by each phase of data collection.

When considering missing data percentages per participant (Table 4-48), the higher mean and median percentage of missing data by participant for the ICS Adult sample suggests a high number of participants in this sample dropped out of the study.

Figure 4-109 shows the pattern of missingness across the item sets. For each participant if an item was missing from an item set that item set was classed as incomplete. As such, blue cells in the figure below show the frequency of completed item sets.

Table 4-47. Percentage of Gwm:Wc Missing Data by Item

Phase	*n Items*	Mean	SD	Median	Min	Max
ITOS	38	11.31	6.63	9.98	1.34	22.46
ICS Adult	34	50.32	29.09	55.24	6.67	87.30
ICS School Aged	34	7.49	11.84	1.14	0	45.45

Table 4-48. Percentage of Working Memory Missing Data by Participant

Phase	*n* Participants	Mean	SD	Median	Min	Max
ITOS	521	11.31	24.13	0	0	97.37
ICS Adult	315	50.32	28.48	55.88	0	97.06
ICS School Aged	88	7.49	11.86	0	0	61.76

Note. Missing data for this table is calculated based on the items available to each participant.

Figure 4-108. Percentage of Gwm:Wc Missing Data by Phase

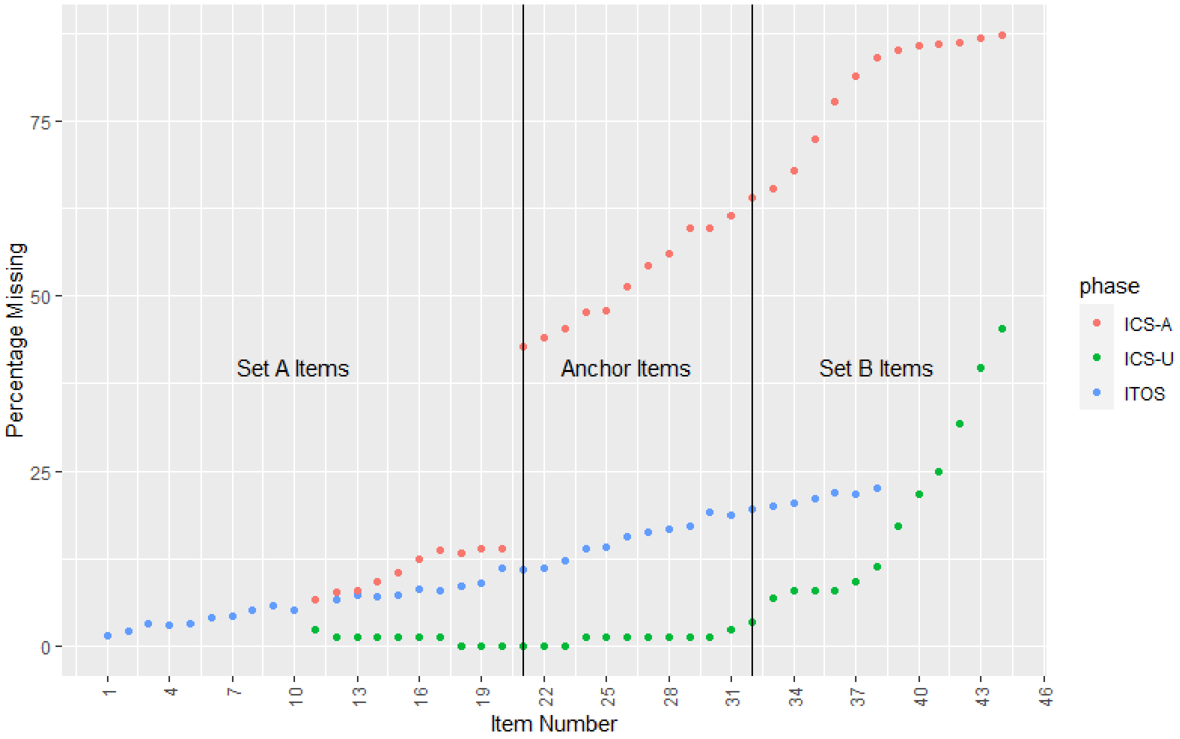
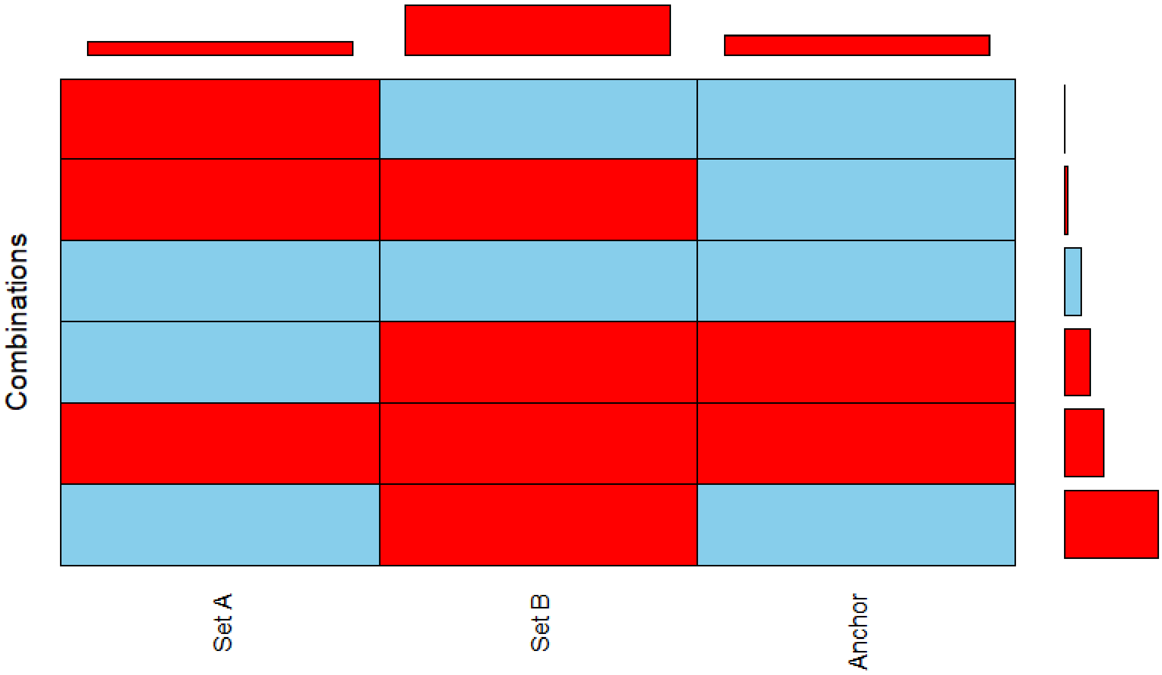


Figure 4-109. Missing Data Pattern for Gwm:Wc Sets



Note. Red cells indicate missing data while blue cells indicate present data. The bars above and beside the cells represent the proportion of participants with that pattern of data.

4.7.3 Data Imputation

Unlike the missing data analysis above, the data imputation will consider the dataset in entirety. Data was prepared for MICE in stages as detailed in the methodology and demonstrated in Table 4-49.

This has implications for the interpretation of missing data compared to the previous section. In the missing data analysis section, a participant was not counted as having missing data if they were in the ITOS sample and had completed every working memory item between 1 and 38; however, they would be considered to be missing data for items 39 through 44 (Figure 4-110). As an additional example, if they only completed items between 1 and 10, they were not identified as non-responders, however in this case they would have 100% missing data because the items they did complete were no longer in the data set (Figure 4-111). The pattern of missingness based on item sets after these alterations is represented in Figure 4-112.

After the data cleaning above, items were analysed to determine if they would be flagged as problematic for MICE. No items were flagged as problematic for the Random Forest algorithm. Items 37-44 were flagged with an outflux below .30 (Figure 4-113). These items were not removed as this is the entire Set B developed for this study. It is anticipated that further data collection will address this in future as Set B represents the largest set of missing data for the Working Memory items. The minimum correlation of .20 between items for MICE is also likely to partially address this problem.

MICE was successfully carried out with 75 individual formulas being executed for 75 variables. No problems were identified with the imputation. Figure 4-114, Figure 4-115, Figure 4-116, and Figure 4-117 show comparisons between the data before and after imputation.

Table 4-49. ICS Gwm:Wc Missing Data Cleaning

Phase	*n Items*	Mean	SD	Median	Min	Max
Initial Missing Data[1]	44	54.38	32.99	51.74	5.71	100
Full Anchor Items	44	31.91	40.91	2.58	0	100
Removed Items[2]	34	11.89	18.69	1.03	0	55.67
Low Response Rate[3]	34	11.89	18.69	1.03	0	55.67
High Missingness	34	10.90	17.99	0.53	0	54.55
Missing Age	34	10.84	17.92	0.54	0	54.35
Time Outliers	34	11.85	19.92	0	0	60.37
Score Outliers	34	12.18	20.39	0	0	61.78
Multivariate Outliers	34	12.31	20.57	0	0	62.67

[1]All ITOS participants removed due to change in item presentation; [2]Items 1-10; [3]No items had low response

Figure 4-110. Gwm:Wc Missing Data for Data Imputation Preparation

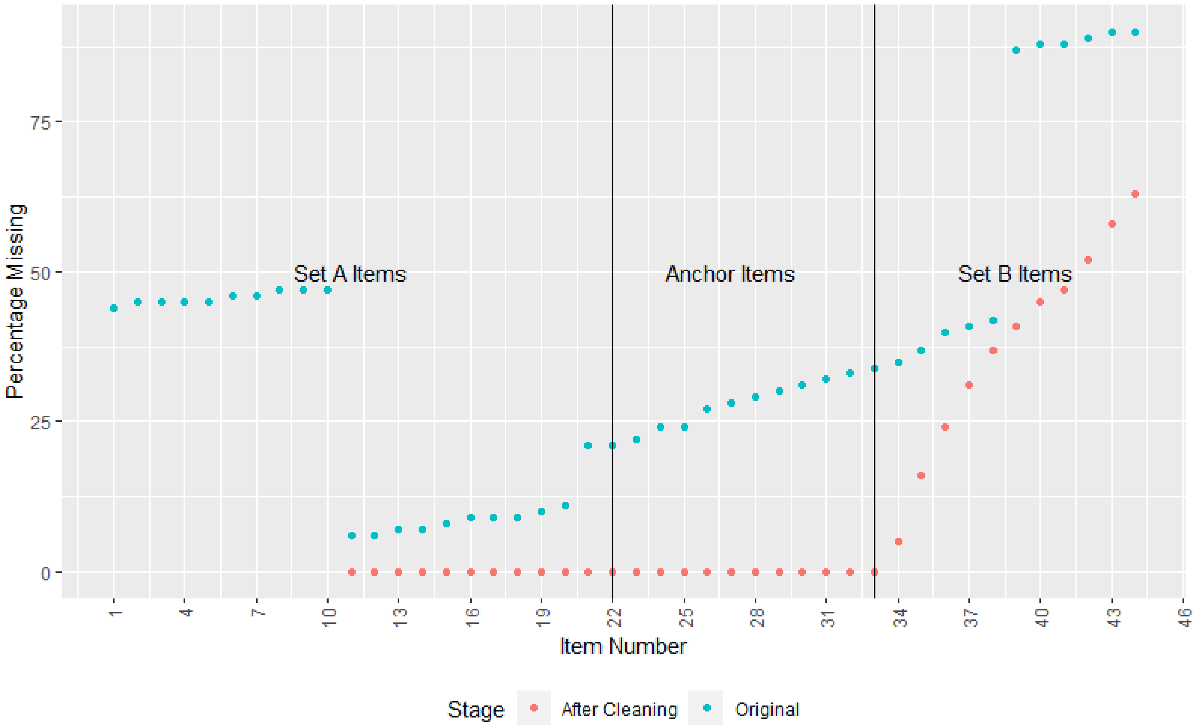
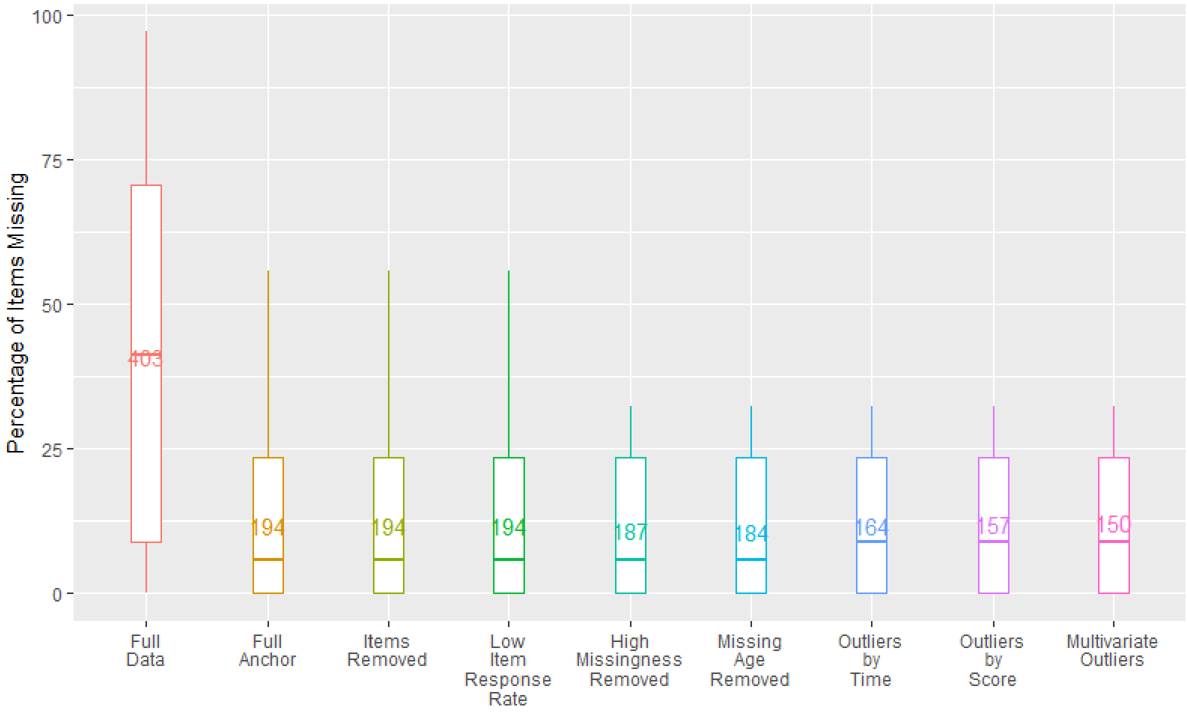
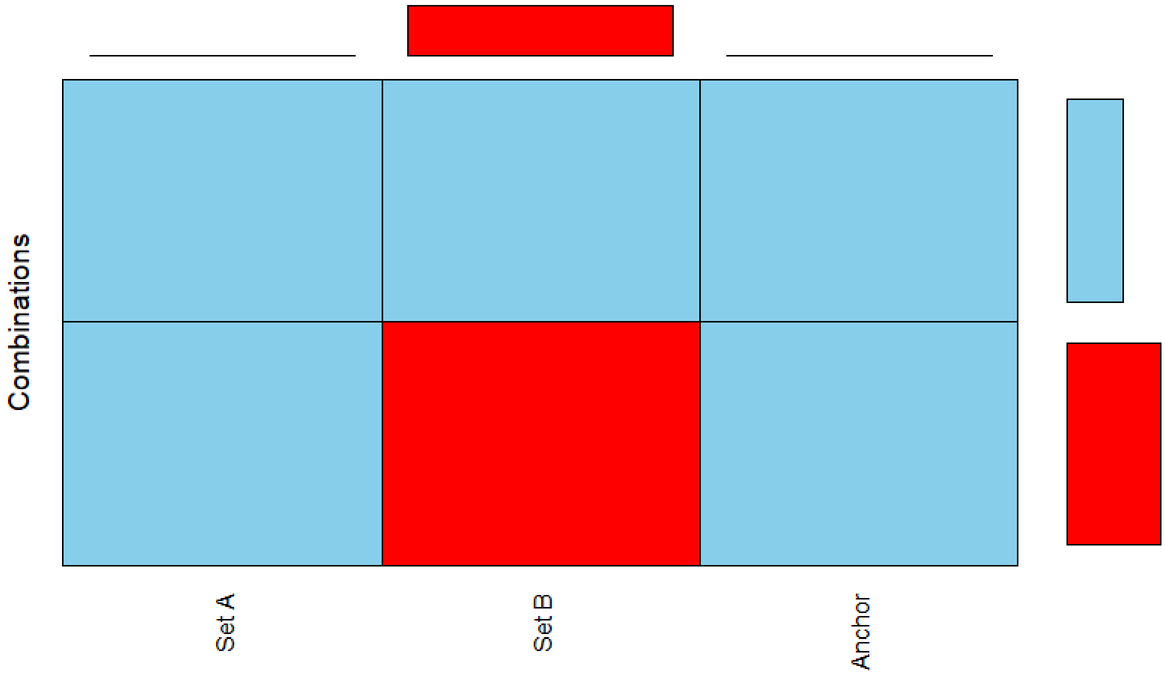


Figure 4-111. Percentage of Missing Gwm:Wc Items by Participant



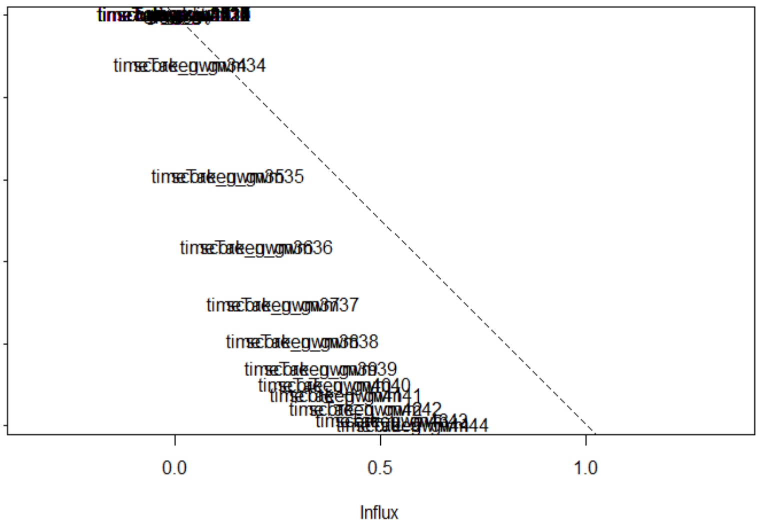
Note. All ITOS participants removed due to change in item presentation; No items had low response rates

Figure 4-112. Missing Data Pattern for Gwm:Wc with Full Anchor Items



Note. Red cells indicate missing data while blue cells indicate present data. The bars above and beside the cells represent the proportion of participants with that pattern of data.

Figure 4-113. Outflux-Influx Before Removal



Note. No outflux-influx figure after removal as no items were removed at this stage of cleaning. Items with low outflux have poor predictive power for the imputation model and items with high influx have high reliance on the imputation model. Ideally you have high outflux and low influx.

Figure 4-114. Gwm:Wc Score by Imputation

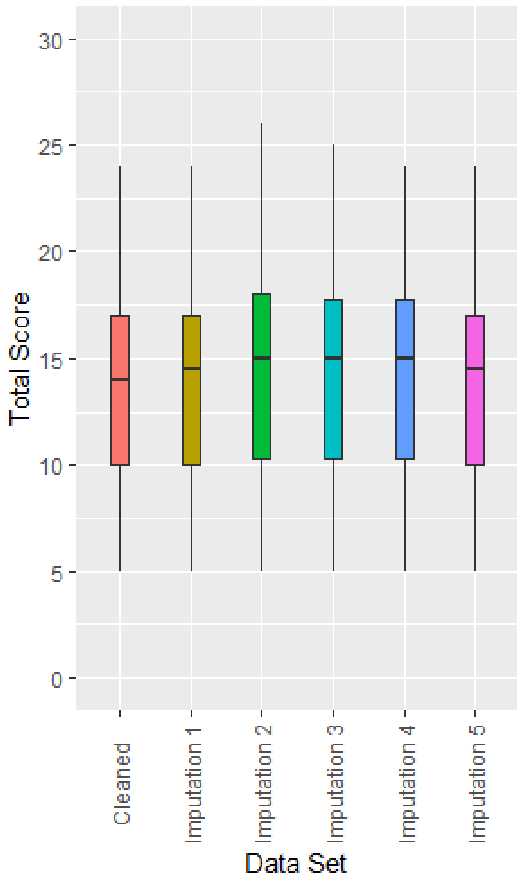


Figure 4-115. Gwm:Wc Time by Imputation

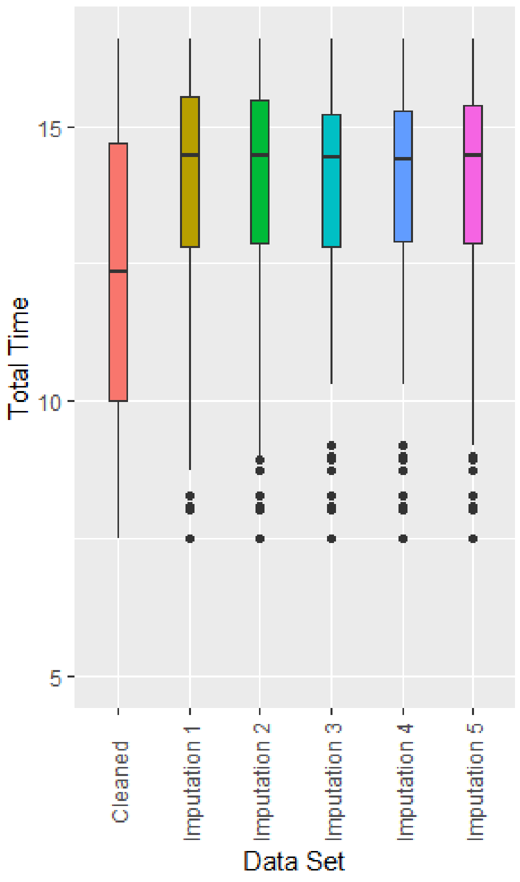


Figure 4-116. Gwm:Wc ICS Percentage of Items Correct

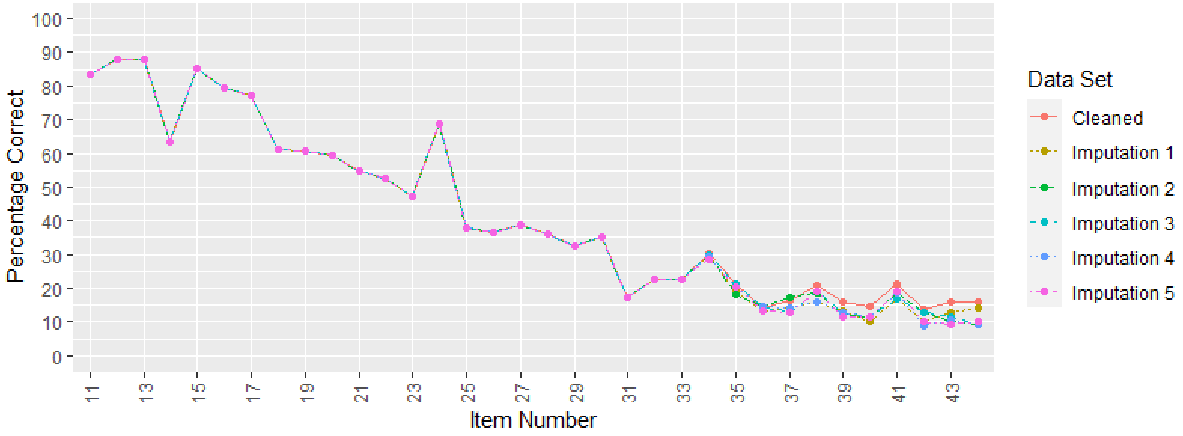
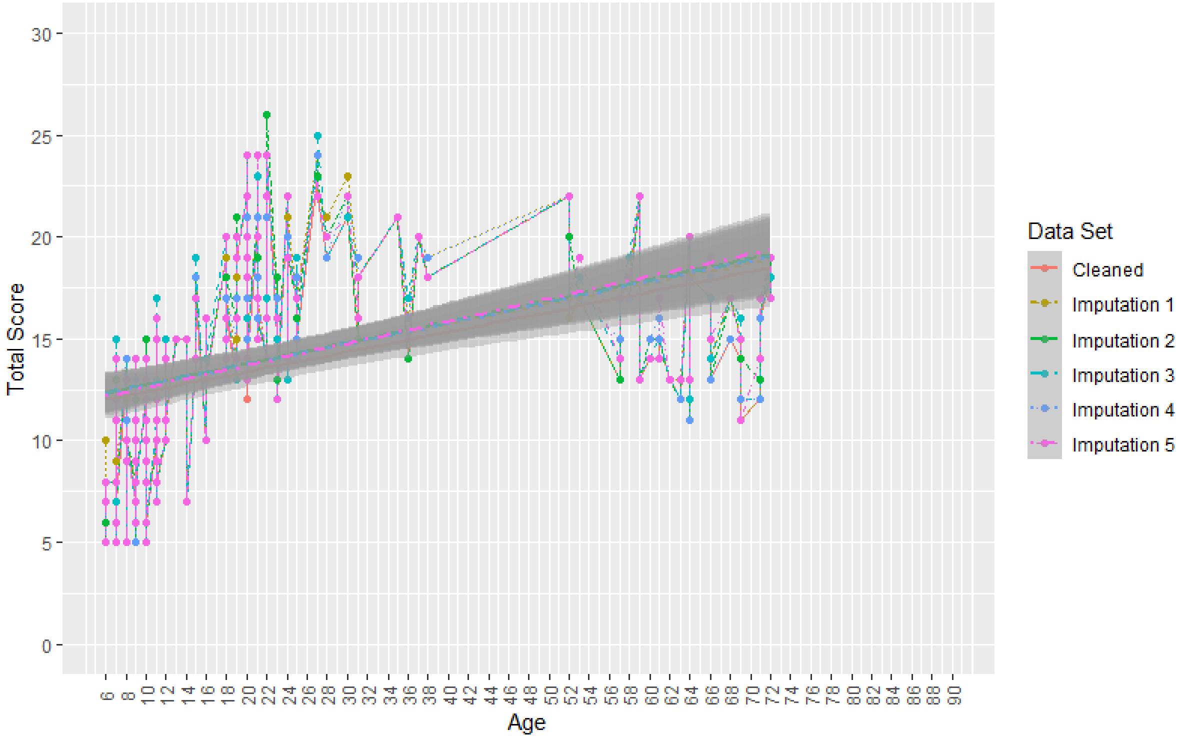


Figure 4-117. Gwm:Wc Total Score by Age



4.7.4 *Reliability*

Cronbach’s alpha for the ICS Working Memory items was calculated across each of the imputed data sets and was found to have a mean of .76. This was lower than the Cronbach’s alpha in the ITOS, likely reflecting an increased diversity in the difficulty of items. Regardless, this level of reliability can be classified as good.

4.7.5 Rasch Analysis

Rasch analysis was conducted on all five imputed data sets. For the 34-item set there was poor fit across all indices. Item fit details are in Appendix P.

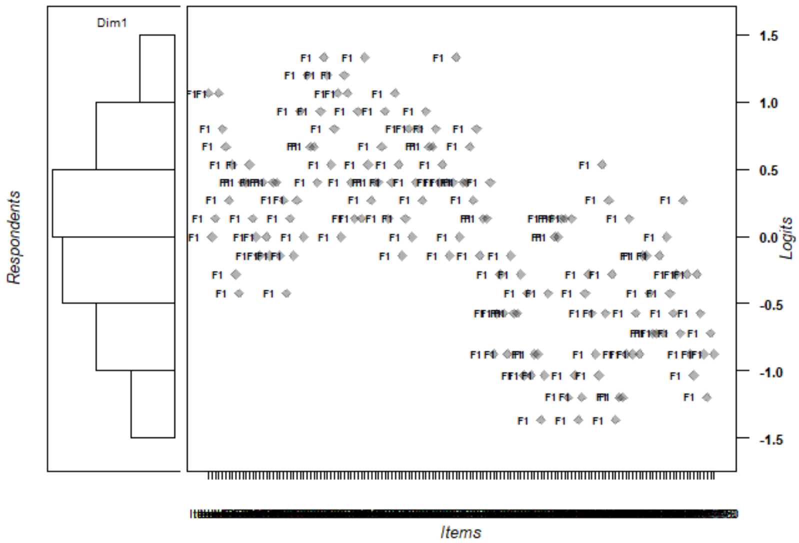
Person fit statistics were calculated (Figure 4-118), and 17 participants were found to have Zh values above +2 and below -2. Twelve of these participants were in the school aged sample. Again, a wide range of difficulty of items is believed to explain the participants that have high Zh values, likely a result of guessing.

Table 4-50. Rasch Scale Fit Statistics for Gwm:Wc 34-Item Set

Data	M2	DF	RMSEA	SRMR	TLI	CFI	Marginal Rxx
Imp 1	1213.17***	560	0.09	0.14	0.62	0.62	.84
Imp 2	1123.42***	560	0.08	0.14	0.67	0.67	.84
Imp 3	1051.75***	560	0.08	0.14	0.70	0.70	.84
Imp 4	1106.34***	560	0.08	0.14	0.68	0.68	.84
Imp 5	1209.49***	560	0.09	0.14	0.63	0.63	.84

Note. **p* < .05; ***p* < .01; ****p* < .001

Figure 4-118. Gwm:Wc 34-Item Set Wright Map



Items 32 (9 chunks), 35 (10 chunks), 36 (10 chunks), 39 (12 chunks), 40 (12 chunks), 42 (13 chunks), 43 (14 chunks) and 44 (14 chunks) demonstrated poor fit with the Rasch model. Analysis of these items showed they were difficult for participants. There did not appear to be any qualitative similarities between the items, with no pattern to the colours or numbers used. The flagged items were also mixed in the information required to be remembered. Unfortunately, as a multiple-choice response option set of items, there is a 16% chance of guessing the item correctly. Given the sequence of difficulty, the items were retained and flagged for monitoring.

4.7.6 Mokken Analysis

Thirty items were found to have Loevinger's *H* values below .30. Items fitting poorly with the Rasch model were also not fitting well with a Mokken scale. Items 38 to 44 demonstrated negative Loevinger's *H* values which suggested more errors on these items than expected with the scale ordered in sequential difficulty. It is likely the participants who scored these items incorrectly and correctly were guessing and so items 37 to 44 were removed. This improved Loevinger's *H* values.

Table 4-51. Mokken Analysis for Gwm:Wc 34 Item Set

Data Set	**Loevinger's *H***	**Standard Error**
34-items		
Imputation 1	.16	.02
Imputation 2	.17	.02
Imputation 3	.16	.02
Imputation 4	.17	.02
Imputation 5	.16	.02
26 items		
Imputation 1	.27	.02
Imputation 2	.27	.02
Imputation 3	.27	.02
Imputation 4	.27	.02
Imputation 5	.27	.02

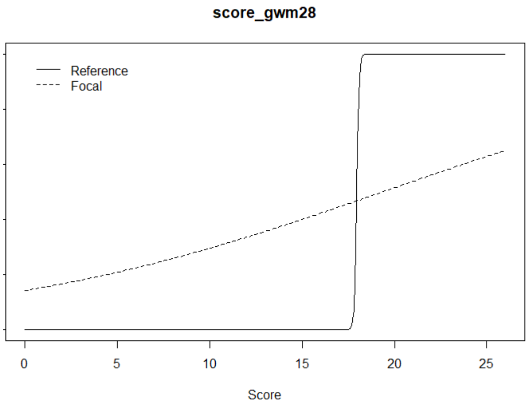
4.7.7 *Local Independence*

Local dependency was found for items 11, 12, 15, 37 and 39; substantially less local dependency compared to the ITOS. No commonalities were identified between these items and there was item would assist with other items. No items were removed.

4.7.8 *Differential Item Functioning*

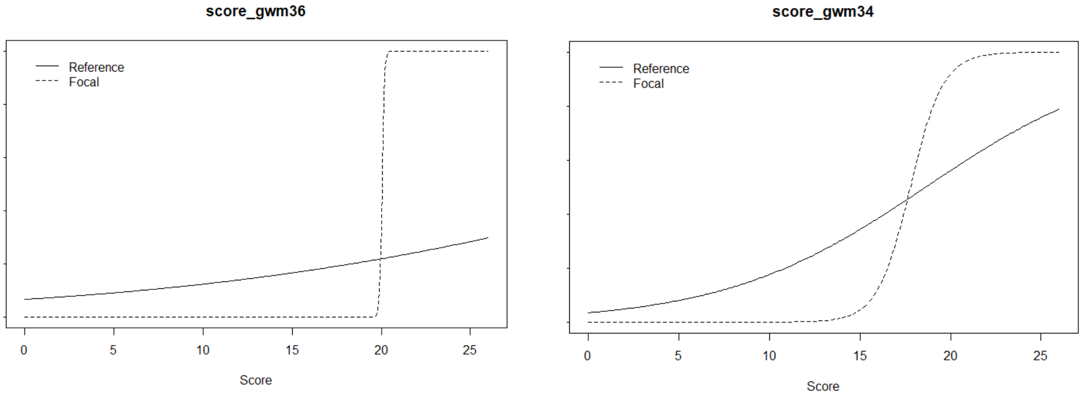
No items were flagged for DIF by gender. Item 28 was flagged for DIF by nationality, likely because of the low numbers of non-Australian participants (3 out of 9 obtained a correct answer). Items 34 and 36 were flagged for DIF by device. This DIF is likely due to some devices being used by few participants. No items were removed.

Figure 4-119. Gwm:Wc DIF by Nationality



Note. Focal group was Australian and the Reference group was non-Australian

Figure 4-120. Gwm:Wc DIF by Device



Note. The focal group was iPhone and the reference group was non-iPhone

4.7.9 Confirmatory Factor Analysis

CFA was completed with the Working Memory 26-item set. The only CFA that converged was for imputation three and with the resultant fit indices suggesting that the model had overfit due to the low sample size relative to the number of items. As with the other item sets, a higher SRMR was identified suggesting correlations of residuals. Most items had very high standardized loadings, however items 28, 32, 35 and 36 had loadings below .30. Item 32 was identified with a negative loading and was also flagged in the previous Rasch and Mokken analyses; this item was removed.

Table 4-52. CFA for Working Memory 26-Item Set

Imputation	χ^2	DF	RMSEA	SRMR	TLI	CFI
Imp 3	252.87	299	.00	.12	1.02	1.00
Imp 5	260.57	299	.00	.12	1.02	1.00

Note. **p* < .05; ***p* < .01; ****p* < .001

4.7.10 Rasch Item Parameters, ICC and Test Information

Rasch fit statistics (Table 4-53) for the 25-item set demonstrated improved Rasch fit. RMSEA, TLI and CFI were acceptable. The higher SRMR suggests that there continues to be correlation of the residuals despite removal of some misfitting items. The final parameters were calculated after item calibration using the rules from Rubin (1987) to pool parameter and standard error estimates across the five imputed data sets (Appendix Q). The items range from -2.43 θ (item 12) through to 2.20 θ (item 36) as shown in Figure 4-121. Items progress in an *almost* linear pattern of difficulty based on the number of chunks required within each item, with item 24 as the most significant exception. The highest point of information, and thus the point of highest reliability is at approximately 0 θ as shown in Figure 4-122 and Figure 4-123, respectively.

Table 4-53. Rasch Scale Fit Statistics for Gwm:Wc 25-Item Set

Data	M2	DF	RMSEA	SRMR	TLI	CFI	Marginal Rxx
Imp 1	410.42***	299	0.00	0.05	0.11	0.93	.80
Imp 2	421.40***	299	0.00	0.05	0.11	0.92	.80
Imp 3	430.01***	299	0.00	0.05	0.11	0.92	.80
Imp 4	427.54***	299	0.00	0.05	0.11	0.92	.80
Imp 5	426.56***	299	0.00	0.05	0.11	0.92	.80

Note. **p* < .05; ***p* < .01; ****p* < .001

Figure 4-121. Gwm:Wc ICS Rasch ICCs

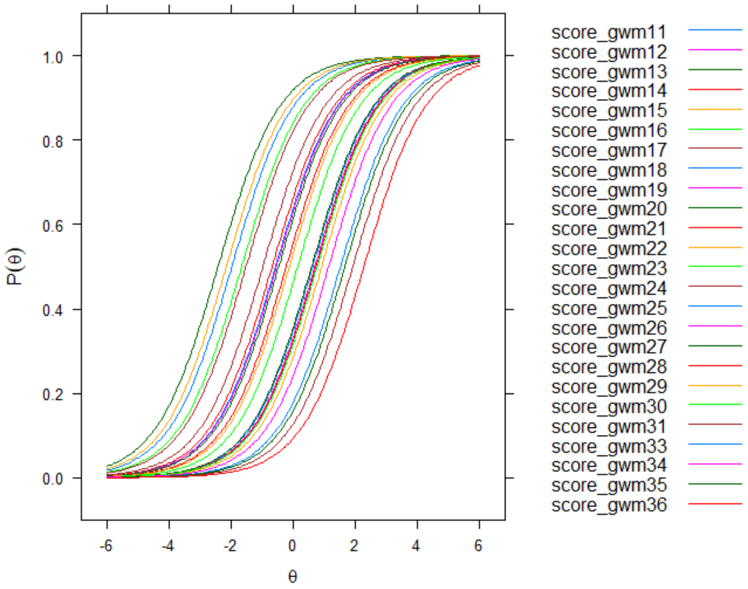


Figure 4-122. Gwm:Wc ICS Rasch TIC

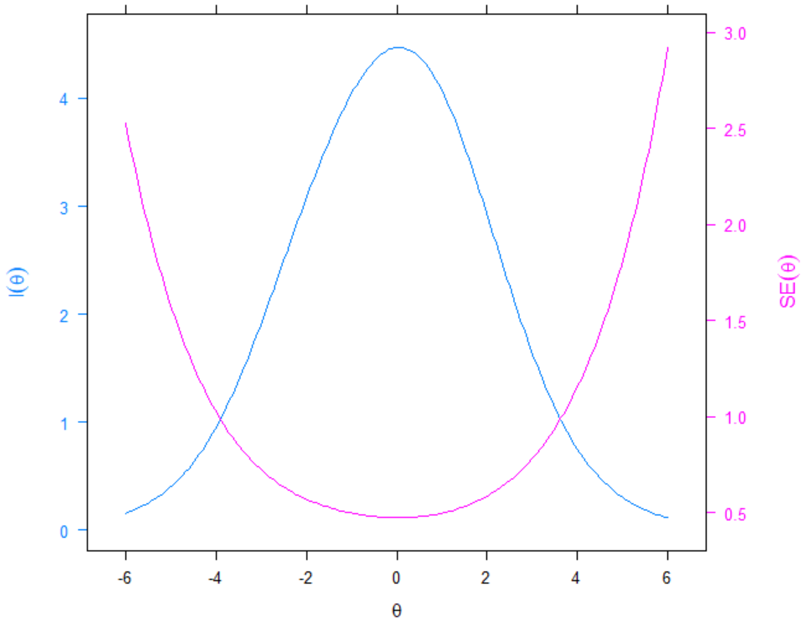
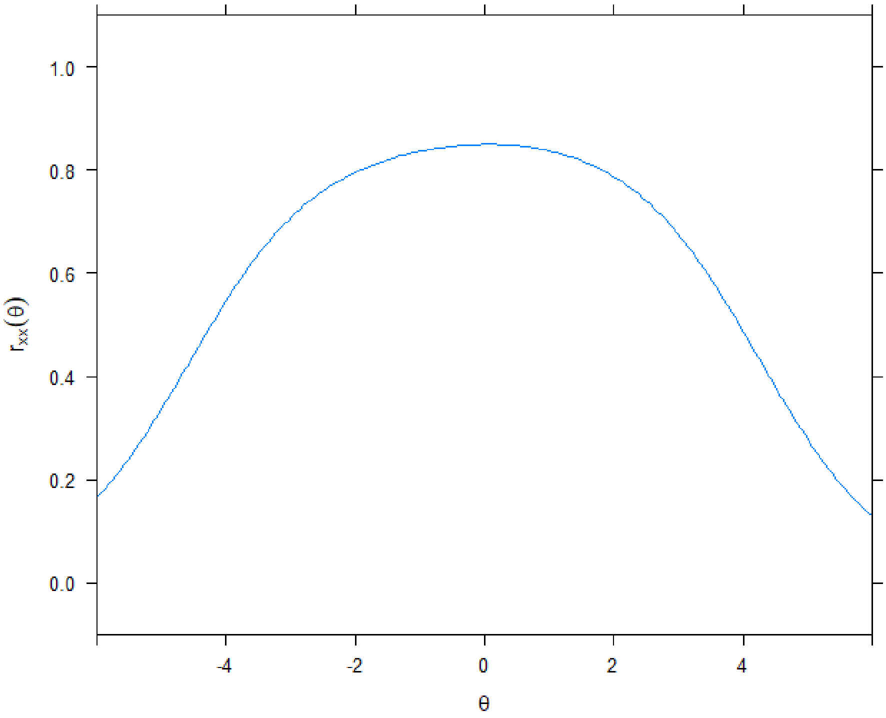


Figure 4-123. Gwm:Wc ICS Rasch Test Reliability Curve



4.8 Discussion

To develop a CAT, known item parameters which place items along the latent trait need to be established. Calibration of items involves demonstrating how well items fit with the proposed theoretical model and determining the difficulty of each item. While the ITOS attempted to demonstrate the viability of the items as measures of the theoretical construct, the current ICS study set out to continue calibration of the Lexical Knowledge, Induction, Visualisation and Working Memory item sets for use in a CAT. The goal of these item calibrations was to discover a set of items for each CHC construct under investigation to utilise in later studies, as well as to continue the iterative adjustment and development of items for a CHC-CAT screening tool. An approach was taken which integrated qualitative analysis, CHC theory, and psychometric analysis of items, rather than conservative reliance on psychometric analysis alone (such as in the ITOS).

Exploration of the data showed there were differences based on phases. This seemed to be largely due to a more significant drop out rate in the ICS-A group, as well as having different items available for the ITOS participants. The ICS-U group showed the highest mean score and time taken, likely due to being supervised. In contrast, the ICS-A showed the lowest total scores and time taken. The only exception to this appeared to be Working Memory within the ICS phase, for which adults outperformed the school aged participants. In general, these differences are caused by research design decisions rather than true differences in the abilities of the samples. Interpretation of such differences should therefore be avoided.

The items developed for the ICS proved more difficult for both adults and the school aged samples. Despite this, increased time spent by adults was generally linearly related to improved total score. In contrast, more time for school aged participants did not necessarily generate better scores, likely because of there being more variability in their actual abilities. This is particularly the case for Working Memory where there was an overall steady decline in performance as the number of stimuli increased.

Across all four item sets, several items had to be removed prior to data imputation due to significant amounts of missing data. This means they were not included in any subsequent analyses and have not yet been robustly analysed to determine if they meet the assumptions of IRT or may be useful in a CAT. Future iterations of the screening tool may consider reattempting administration of these items to gather more data. This would enable further analysis to determine whether they fit the proposed models or not.

The data imputation method utilised in this study was significantly more complex than for the ITOS. Consistent with previous research, MICE proved useful in estimating missing data. Because so many Lexical Knowledge items were removed prior to data imputation, participants' recalculated total scores for this item set dropped somewhat. Otherwise, considering the patterns of data in the cleaned and the imputed data, there were only rare deviations in the expected pattern; this tended to be in the older age groups where there was less data to reliably predict outcomes.

While still present, less local dependency was identified in the ICS compared to the ITOS. This was likely a result of more difficult items as well as more varied performance. However, local dependency is generally caused by correlation of residuals, meaning there is some element of commonality between the items outside the intended construct of interest. For a test of Lexical Knowledge (G*c:*VL), it is possible that items may also be measuring other Comprehension-Knowledge (Gc) narrow abilities such as Language Development and General Verbal Information. For Induction this local dependency may be a result of the problems presented or the two different time limits provided, tapping into other abilities. This problem was noted after the ITOS and further item development should continue to be sensitive to this problem.

While a few items were flagged for DIF, there appeared to be minimal differences in the probability of performance based on group membership (i.e. nationality and device used). Additionally, the overall score and time taken by participants was relatively consistent across genders, nationalities and devices, with participants who identified as a non-binary gender often displaying the most varied performances. Research has demonstrated there is relatively little difference between those that complete items on computers versus tablets (Kong et al., 2018). The variation in performance on specific items, as well as overall, for these groups is likely to centralise as more data is gathered.

The results indicate that the retained 47-item Lexical Knowledge, 23-item Induction, 30-item Visualisation, 25-item Working Memory sets each meet the monotonicity and unidimensionality assumptions of IRT, and that the items fit the Rasch model. Issues with local dependency were apparent in some item sets which were unexplainable by item content, and thus likely explained by another latent variable. It may be possible to address these issues by moving from four unidimensional analyses to a multidimensional analysis which includes other narrow abilities, broad abilities, and a general factor. Despite this, the number of items retained and the overall fit indices for all four item sets showed significant improvement from the ITOS.

For each psychological construct the items measure an extensive range of ability for individuals aged 6 years old to 90, particularly those with very low, low-average and average abilities. Visual inspections of the SEM curves for each item set suggests a reliability of .7 (suitable for group administration) would be achievable from a θ of -6 to 2 for Lexical Knowledge, -1 to 3 for Induction, -4 to 1 for Visualisation, and -3 to 3 for Working Memory. The items in these sets may not be useful for determining those with exceptionally high skill, particularly for Visualisation which appears to measure quite accurately for lower difficulty items but not as well for more difficult items.

The fit of the items to the Rasch model and the unidimensional CFA models was generally good, with problems only noted for the SRMR. This suggests there are more significant discrepancies between the observed correlation matrix and expected correlation matrix. This fit statistic is particularly independent of sample size (Chen, 2007). This further supports conclusions that there is residual variance unaccounted for by the four independent, unidmensional CHC narrow ability models implemented in this study. With further data collection it may be possible to overcome this challenge by adopting a hierarchical model consistent with the broader CHC literature.

Overall, it is believed the item difficulty parameters generated from the ICS can be utilised in a simulated CAT. While further research would be required to demonstrate the utility of the items in high stakes decision making, these items could be suitably implemented in research studies or group testing scenarios when the latent traits under investigation include Lexical Knowledge, Induction, Visualisation and Working Memory.

4.8.1 Limitations and Future Research

Missing data was a problem for the current study. Despite a high number of participants recruited for the ICS Adult data set, and the supervised administration for the ICS School Aged data set, there was a far greater amount of missing data for each CHC ability's Item Set B relative to what was expected based on data collected in the ITOS. The Internet sampling methodology worked well for the ITOS study, and it is unknown (outside of increases in item numbers) why there were more significant dropout rates in the ICS Adult sample. Attempts were made to address this via the use of MICE rather than ML methods, however clear impacts of the missing data are evident.

In contrast to this negative outcome of sampling methodology, the addition of more reliable data sources (i.e. ICS School Aged; ICS-U) and additional adult data with a focus on 18-25-year-olds (i.e. ICS Adult; ICS-A), appears to have provided further evidence of the reliability and validity of some of the original items. This is reflected in a wider range of items being retained for each CHC ability relative to the conclusions of the ITOS. This suggests that with iterative development of the CHC-CAT through continued item development and participant recruitment, existing items can continuously be evaluated while new items can continuously be introduced. In fact, this is the key advantage of relying on IRT as the core mechanic of the test rather than CTT perspectives.

For the CFAs in this chapter, there tended to be a pattern of slightly lower CFI and TLI indices compared to the ITOS and increases in SRMR and RMSEA indices. Previous research shows that increases in variables (as in the current study) lead to slight decreases in CFI and TLI, while smaller sample sizes (caused by the data cleaning methodology used in the current study) increase SRMR and RMSEA indices. Therefore while the fit indices are not perfect in the current study, given the focus on IRT, the CHC theoretical basis that the items were developed under, and the cumulative evidence of unidimensionality across each CHC ability throughout the ITOS and ICS suggests that the items are measuring the constructs they are intended to measure. Even for more complex items that may require multiple psychological processes, the assumption of unidimensionality is not necessarily precluded "as long as they are affected by the same underlying process" (Iramaneerat et al., 2008, p. 55), hence the lack of adherence to stringent CFA standards in this ICS versus the ITOS.

DIF was also a problem in the current study largely because of the uneven samples collected, which was much more evident in the Induction and Working Memory item sets due to the smaller sample sizes after data imputation. For gender, those that identified as Other or Prefer Not to Say tended to result in DIF being flagged for more difficult items, likely as a result of the unproctored adult test leading to low effort. For nationality, some items only had less than 10 non-Australian attempts, and thus DIF would flag this as significant changes in probability based on group membership. Similarly, while the raw data showed a high rate of diversity in devices used to access the tasks, after data cleaning very little variation remained and most people had used iPads, iPhones, or laptops; use of other devices was often erroneously flagged as DIF when a small proportion of people using a Samsung Galaxy or Google Pixel had varied performance.

4.8.2 Conclusion

The ICS demonstrates an iteratively improving set of item psychometrics with less need for new items compared to the ITOS but does suggest the need for more participant data to be collected. More participants will reduce the biases introduced in this study through removal of valuable data prior to data imputation, improve the MICE imputation of missing data by having more variables to use in prediction formulas, provide further evidence of the probability of a response, and increase the numbers of participants retained in different groups so as to conduct more robust DIF. At this point in time the item psychometrics may not yet be robust enough for implementation in a school setting for high stakes testing, but there is initial evidence of item sets that can be built upon for further research. The analyses have produced logical item parameters that align well with CHC theory.

Thus far, the development and analysis has focused on a variety of statistical tools, but all findings are based on whole sets of items being administered. That is, items are still required to be administered in a conventional manner. It has not yet been tested whether these item sets can be used more efficiently. The following chapter aims to explore this by simulating a CAT using the findings from the current chapter.

Chapter 5: Computer Adaptive Test Simulation

5.1 Introduction

A comprehensive Item Response Theory (IRT) validation alone is not evidence of the application of advanced technology this thesis aimed to achieve. However, such statistical methods form the backbone of ensuring items can be used in a reliable and valid manner in computer adaptive tests (CATs). The previous chapter resulted in sets of 47 Lexical Knowledge (G*c*:VL), 23 Induction (G*f*:I), 30 Visualisation (G*v*:Vz), 25 Working Memory (G*wm*:Wc) items that each met the monotonicity and unidimensionality assumptions of IRT, and that the items fit the Rasch model. For each item Rasch item parameters were exported. The current chapter focuses on taking the item sets from the Item Calibration Study (ICS) and utilising them in a simulated CAT to pilot the viability of implementing technological advancements in the measurement of intelligence.

CATs offer opportunities to measure psychological constructs in a more efficient way while maintaining reliability. An advantage is that a test which is adaptive in nature can measure a wider range of theoretical constructs without requiring administration of every single item. When developing a CAT, it is possible to use existing data sets to simulate items that would and would not be presented to potential examinees. This is known as a "Monte Carlo Simulation"; a broad class of methods that utilise random sampling or generation of random response patterns to evaluate the outcomes of various algorithms (Kroese et al., 2014). At their core, CATs rely on algorithms. Conducting a simulation study allows us to determine the outcomes of a CAT based on different response patterns. Having identified a set of items for each Cattell-Horn-Carroll (CHC) narrow ability under investigation in this thesis, this CAT simulation uses simulated participants that possess a wide range of theta (i.e. spectrum of difficulty; θ) to evaluate the accuracy of the CAT. The simulations used the 47 Lexical Knowledge items, 23 Induction items, 30 Visualisation items, and 25 Working Memory items retained from the ICS, each in a unidimensional CAT.

5.2 Chapter Aim

Having considered the characteristics of CATs in Chapter 2, several expectations can be established based on the outcomes of the ICS. Firstly, given the marginal reliabilities achieved for each CHC ability in the ICS (all below .90) it is likely the average number of items administered in a CAT using these items will only reduce when a higher standard error of measurement (SEM) is accepted. That is, even when an individual has an ability proportionate to the level of θ that the test is best targeted at, every item available will need to be administered in order to achieve a SEM approaching .32 (associated with a reliability of .90). It is likely that even administration of the full item set won't approach this level of SEM. In some cases, it may be acceptable to utilise a higher SEM cutoff, such as in research or group level administration (such as in a screening tool) rather than in individual clinical decision making. Further development of items and more collection of robust data will be required to obtain reliable measurement for high stakes decision making (discussed further in Chapter 7).

The second expectation is that the average items administered for each CHC ability will be reduced if the parameters are recalculated for the item set retained for school aged people only. The item difficulty parameters for the Lexical Knowledge item set ranged from extremely easy to mildly difficult when established using participants aged six to 90 years old; similar patterns, although to a less extreme extent, were noted for the other CHC abilities under investigation. By recalculating parameters using only school aged participants who answered all items retained in the ICS, the mean θ will alter and shift the parameters to the right although maintain the same order (as children have a higher probability of having lower abilities than adults). Thus, if a CAT is simulated using 5,000 simulated participants with a mean θ of 0 and a standard deviation of 1, there is likely to be a match between the recalculated item difficulty parameters and the simulated participants' 'true' θ. This would ultimately demonstrate that the items developed are more suitable for classifying the abilities of school aged participants rather than adult participants.

The third expectation is that when using these recalculated item difficulty parameters, the average bias detected would be reduced closer to zero θ. This can be predicted due to the clear peaked nature of the test discovered in the ICS as well as having a normally distributed simulated sample. While this is less than ideal to reliably measure individuals across the spectrum of ability, further items are needed to be developed that fit the assumptions of the chosen IRT model and fit the model itself (i.e. are retained after analysis).

In contrast to the third expectation, a fourth expectation would be that when broken down into groups by θ (i.e. -3 θ to -2.4 θ, -2.4 θ to -1.6 θ etc), the correlation discovered between true θ and estimated θ will be unrelated to the distribution of participants, but to the location of the items on the ability spectrum. For example, as there are more Lexical Knowledge items that are extremely easy, there is likely there to be higher correlations between the true θ and estimated θ from a CAT for simulated participants at around -3 θ. Such patterns will assist the identification of gaps in the difficulty spectrum measured by the test. Having a gap at the lower end of the difficulty spectrum is more problematic than the middle because this is a screening tool that should identify individuals with challenges.

5.3 Method

5.3.1 Participants

The participants utilised in this study are from the total sample recruited throughout the ITOS and ICS. Participants were arranged into two groups for analysis. The first group (ICS Parameters) consisted of participants who were retained after data cleaning and preparation for data imputation; these are the adult and school aged participants used in the ICS to generate the Rasch item parameters. The second group (School Aged Parameters) consisted of participants aged 18-years-old and younger that remained after data cleaning in the ICS. The resulting participant groups and their descriptive statistics are shown in the results section for each respective CHC ability.

5.3.2 Data Analysis

To determine the effectiveness of a CAT using the items retained in the ICS, a simulation of 5,000 participants with true θ scores with a mean of 0 and a standard deviation of 1 was carried out using *mirtCAT v.1.9.3* (Chalmers, 2016) package. The *parallel v.3.6.3* (R Core Team, 2020) package was used to take advantage of parallel computing; it essentially creates a set of copies of *R* running in parallel to allow computation of complex analyses in a quicker fashion.

The *mirtCAT* package allows specification of CAT characteristics to be simulated in line with those discussed in Chapter 2. An item bank with known psychometric data was established using (1) the items retained and associated Rasch parameters from the ICS outcomes, (2) a recalculation of item parameters for participants of all ages who attempted all items retained from the ICS, and (3) a recalculation of item parameters for participants aged 18 years old and younger who attempted all items retained from the ICS. The first item selected for administration in each item set was determined using Maximum Information (MI) (Chalmers, 2016). The method for estimating θ after administration of each item was expected a posteriori (EAP) estimation; in CATs this produces less bias, particularly for shorter tests (Huebner et al., 2016). The item selection method after θ estimation was set as MI. For each CHC ability simulations were carried out with the stop rule set at iteratively higher levels of acceptable SEM (and thus an iteratively lower level of reliability), ranging from .00 SEM (a reliability of 1.00) through to .71 SEM (a reliability of .50). A SEM of .00 (reliability of 1.00) is theoretically not possible, however setting this as the stop rule essentially ensures that the test finishes when the number of items administered equals the number of scale items available. The formula to calculate reliability from SEM, and vice versa, is in Equation 5-1. No optional constraints were utilised. Thus, for each CHC ability a total of 12 simulations were conducted.

Equation 5-1. SEM and Reliability

$$SEM = S_x\sqrt{1 - r_{xx}}$$

$$r = 1 - \frac{SEM^2}{s_x}$$

where s_x equals the SD of test scores (i.e. 1) and r_{xx} equals the test reliability.

For each simulation two outcome statistics were calculated. The first calculation was the average number of items to be administered simulated participants. The second was the Pearson correlation between the simulated participants' true θ and their estimated θ. A lower number of average items administered is ideal but is expected to differ based on the participants' true θ. Correspondingly, a higher Pearson correlation between true θ and estimated θ is ideal; no specific cutoff is set as the importance of the correlation changes based on the location of the participant on the latent trait.

Because of these nuances in interpretation, after these initial simulations, a simulation with a stop rule of .45 SEM (i.e. reliability of .80) was carried out using the parameters estimated on the second participant group (School Aged) for each CHC ability. For this simulation, outcome measures were calculated by splitting participants into 12 groups based on their true θ, in groups of 0.6 θ, ranging from less than -3 θ through to above 3 θ. Two additional outcome statistics were calculated for the θ group simulation: the mean bias (Equation 5-2) and root mean square deviation (RMSD) (Equation 5-3). The former demonstrates on average how much the estimated θ score was biased (or different) from the true θ. The RM*S*D is a measure of the quadratic mean of difference between the true θ and estimated θ, also known as residuals. A lower RM*S*D is desirable, as higher RM*S*D indicates variance that is unexplained.

Equation 5-2. Mean Bias

$$bias = \, mean(\theta_{true} - \, \theta_{estimated})$$

Equation 5-3. Root Mean Square Deviation

$$rmsd = \, \sqrt{mean((\theta_{true} - \theta_{estimated})^2)}$$

5.4 Lexical Knowledge Results

Table 5-1 shows the descriptive statistics for the participants in each group. Table 5-2 shows the performance of the CAT simulations for the Lexical Knowledge items. All 47-items were required to be administered when the stop rule was set at a SEM of .32 for the parameters derived from the ICS and for the school aged recalculated item parameters. Significant reductions in the number of items are noted when the SEM stop rule is set at .45. At this level, using the item parameters established in the ICS, an average of 25.58 items would be administered with a correlation of .90 between the participants' true θ and estimated θ. Table 5-3 demonstrates that while the amount of bias was lowest for those simulated participants between -0.6 θ and 0.6 θ, the best correlation was for simulated participants with a θ below -3.

Table 5-1. Gc:VL Participant Statistics

	ICS Parameters	**School Aged Parameters**
n	372	52
M Age (SD)	32.52 (21.20)	12.15 (3.82)
Gender		
n Male	176	24
n Female	190	28
n Other	2	0
n PNTS	4	0
Nationality		
Australian	347	51
Other	21	1
PNTS	4	0

Note. PNTS = Prefer not to say

Table 5-2. Gc:VL CAT Simulations

Stop Rule		**ICS Parameters**		**School Aged Parameters**	
Reliability	SEM	Average Items	Correlation θ	Average Items	Correlation θ
1.00	.00	47.00	.92	47.00	.94
.90	.32	47.00	.92	47.00	.94
.80	.45	25.58	.90	17.77	.89
.70	.55	13.02	.85	10.22	.83
.60	.63	7.58	.79	6.53	.77
.50	.71	5.09	.71	4.14	.70

Note. Items = Average Items = Average number of items administered; Correlation θ = the Pearson correlation between the true and estimated θ scores. Reliability of 1.00 set as stop rule to ensure all items administered and demonstrate correlation between full CAT and full linear administration.

Table 5-3. Gc:VL CAT Performance by Theta Groupings

Theta (θ) Group	*n*	Average Items	Correlation θ	Average Bias	RMSD
Below -3.0	9	33.67	.82	-0.72	0.81
-3 to -2.4	47	23.28	.29	-0.64	0.73
-2.4 to -1.8	140	20.27	.32	-0.44	0.57
-1.8 to -1.2	373	18.58	.28	-0.31	0.50
-1.2 to -0.6	773	17.60	.30	-0.19	0.44
-0.6 to 0.0	1086	17.15	.35	-0.06	0.38
0.0 to 0.6	1168	17.13	.34	0.07	0.38
0.6 to 1.2	816	17.39	.29	0.20	0.44
1.2 to 1.8	419	18.15	.24	0.33	0.54
1.8 to 2.4	126	19.97	.25	0.46	0.60
2.4 to 3.0	37	26.03	.41	0.70	0.83
Above 3.0	6	31.00	.31	0.95	1.01

Note. Stop rule was set as SEM of .45

5.5 Induction Results

Table 5-4 shows the descriptive statistics for the participants in each group. Table 5-5 shows performance of the CAT simulation for the Induction items. All 23-items were required to be administered when the stop rule was set at a SEM of .32 for the parameters derived from the ICS and for the school aged recalculated item parameters. Minor reductions in the number of items are noted when the SEM stop rule is set at .45; an average of 20.75 items would be administered with a correlation of .90 between the participants' true θ and estimated θ. Table 5-6 demonstrates that while the amount of bias was lowest for those simulated participants between -0.6 θ and 0.6 θ, the best correlation was for simulated participants with a θ below -3.

Table 5-4. Gf:I Participant Statistics

	ICS Parameters	School Aged Parameters
n	142	83
M Age (SD)	24.45 (21.01)	9.73 (2.49)
Gender		
n Male	59	40
n Female	81	43
n Other	1	0
n PNTS	1	0
Nationality		
Australian	134	83
Other	8	0
PNTS	0	0

Note. PNTS = Prefer not to say

Table 5-5. Gf:I CAT Simulations

Stop Rule		**ICS Parameters**		**School Aged Parameters**	
Reliability	SEM	Average Items	Correlation θ	Average Items	Correlation θ
1.00	.00	23.00	.89	23.00	.89
.90	.32	23.00	.89	23.00	.88
.80	.45	20.75	.90	18.05	.87
.70	.55	11.85	.84	8.63	.79
.60	.63	7.48	.79	4.14	.67
.50	.71	5.10	.72	2.00	.54

Note. Items = Average Items = Average number of items administered; Correlation θ = the Pearson correlation between the true and estimated θ scores. Reliability of 1.00 set as stop rule to ensure all items administered and demonstrate correlation between full CAT and full linear administration.

Table 5-6. Gf:I CAT Performance by Theta Groupings

Theta (θ) Group	*n*	Average Items	Correlation θ	Average Bias	RMSD
Below -3.0	14	23.00	.66	-1.92	1.94
-3 to -2.4	28	23.00	.26	-1.29	1.31
-2.4 to -1.8	128	22.94	.12	-0.88	0.93
-1.8 to -1.2	380	22.57	.19	-0.54	0.63
-1.2 to -0.6	751	21.19	.29	-0.27	0.46
-0.6 to 0.0	1128	18.64	.32	-0.09	0.38
0.0 to 0.6	1156	16.31	.33	0.11	0.37
0.6 to 1.2	808	15.34	.33	0.26	0.44
1.2 to 1.8	407	15.22	.25	0.48	0.60
1.8 to 2.4	154	15.64	.19	0.73	0.83
2.4 to 3.0	31	16.71	.19	0.97	1.03
Above 3.0	15	18.53	.08	1.32	1.38

Note. Stop rule was set as SEM of .45

5.6 Visualisation

Table 5-7 shows the descriptive statistics for the participants in each group. Table 5-8 shows the performance of the CAT simulation for the Visualisation items. All 30-items were required to be administered when the stop rule was set at a SEM of .32 for the parameters derived from the ICS and for the school aged recalculated item parameters. Minor reductions in the number of items are noted when the SEM stop rule was set at .45; an average of 26.57 items would be administered with a correlation of .87 between the participants true θ and estimated θ. Table 5-9 demonstrates that while the amount of bias was lowest for those simulated participants between -0.6 θ and 0.6 θ, the best correlation was for simulated participants with a θ below -3.

Table 5-7. Gv:Vz Participant Statistics

	ICS Parameters	School Aged Parameters
n	763	112
M Age (SD)	41.04 (19.01)	10.98 (3.69)
Gender		
n Male	337	57
n Female	420	54
n Other	4	1
n PNTS	2	0
Nationality		
Australian	652	108
Other	108	3
PNTS	3	1

Note. PNTS = Prefer not to say

Table 5-8. Gv:Vz CAT Simulations

Stop Rule		**ICS Parameters**		**School Aged Parameters**	
Reliability	SEM	Average Items	Correlation θ	Average Items	Correlation θ
1.00	.00	30.00	.87	30.00	.91
.90	.32	30.00	.87	30.00	.92
.80	.45	26.57	.87	14.17	.87
.70	.55	16.38	.83	6.30	.75
.60	.63	9.16	.78	3.00	.61
.50	.71	5.56	.73	1.00	.41

Note. Average Items = Average number of items administered; Correlation θ = the Pearson correlation between the true and estimated θ scores. Reliability of 1.00 set as stop rule to ensure all items administered and demonstrate correlation between full CAT and full linear administration.

Table 5-9. Gv:Vz CAT Performance by Theta Groupings

Theta (θ) Group	*n*	Average Items	Correlation θ	Average Bias	RMSD
Below -3.0	5	15.60	.71	-1.60	1.62
-3 to -2.4	29	15.21	.12	-1.07	1.13
-2.4 to -1.8	137	14.42	.26	-0.84	0.90
-1.8 to -1.2	409	14.14	.27	-0.53	0.64
-1.2 to -0.6	829	13.86	.31	-0.33	0.48
-0.6 to 0.0	1121	13.64	.27	-0.10	0.37
0.0 to 0.6	1127	13.73	.29	0.11	0.37
0.6 to 1.2	752	14.19	.23	0.34	0.48
1.2 to 1.8	417	15.80	.30	0.54	0.65
1.8 to 2.4	130	17.17	.27	0.89	0.94
2.4 to 3.0	42	20.45	.10	1.25	1.29
Above 3.0	2	30.00	NA	1.40	1.41

Note. Stop rule was set as SEM of .45

5.7 Working Memory

Table 5-10 shows the descriptive statistics for the participants in each group. Table 5-11 shows the CAT simulation performance for the Visualisation items. All 25-items were required to be administered when the stop rule was set at an SEM of .32 for the parameters derived from the ICS and for the school aged recalculated item parameters. Minor reductions in the number of items are noted when the SEM stop rule is set at .45; an average of 22.79 items would be administered with a correlation of .89 between the participants' true θ and estimated θ. Table 5-12 demonstrates that while the amount of bias was lowest for those simulated participants between -0.6 θ and 0.6 θ, the best correlation was for simulated participants with a θ below -3.

Table 5-10. Gwm:Wc Participant Statistics

	ICS Parameters	School Aged Parameters
n	150	89
M Age (SD)	24.27 (19.53)	10.49 (3.56)
Gender		
n Male	64	47
n Female	85	42
n Other	1	0
n PNTS	0	0
Nationality		
Australian	141	88
Other	9	1
PNTS	0	0

Note. PNTS = Prefer not to say

Table 5-11. Gwm:Wc CAT Simulations

Stop Rule		**ICS Parameters**		**School Aged Parameters**	
Reliability	SEM	Average Items	Correlation θ	Average Items	Correlation θ
1.00	.00	25.00	.90	25.00	.89
.90	.32	25.00	.89	25.00	.89
.80	.45	22.79	.89	22.18	.88
.70	.55	11.36	.85	12.10	.84
.60	.63	7.28	.78	7.13	.78
.50	.71	5.12	.72	5.00	.72

Note. Items = Average number of items administered; Cor. θ = the Pearson correlation between the true and estimated θ scores. Reliability of 1.00 set as stop rule to ensure all items administered and demonstrate correlation between full CAT and full linear administration.

Table 5-12. Gwm:Wc CAT Performance by Theta Groupings

Theta (θ) Group	*n*	Average Items	Correlation θ	Average Bias	RMSD
Below -3.0	3	25.00	-.94[1]	-1.59	1.66
-3 to -2.4	33	25.00	.57	-0.87	0.90
-2.4 to -1.8	127	25.00	.21	-0.62	0.71
-1.8 to -1.2	425	25.00	.27	-0.32	0.51
-1.2 to -0.6	778	24.63	.25	-0.14	0.45
-0.6 to 0.0	1155	23.28	.37	-0.03	0.42
0.0 to 0.6	1118	21.14	.33	0.06	0.41
0.6 to 1.2	789	19.36	.26	0.17	0.42
1.2 to 1.8	408	18.69	.32	0.29	0.48
1.8 to 2.4	128	19.55	.22	0.44	0.59
2.4 to 3.0	30	21.83	.31	0.56	0.72
Above 3.0	6	23.00	.86	1.02	1.07

Note. [1]This unexpected negative correlation is believed to be a “Heywood case” (Heywood, 1931). Stop rule was set as SEM of .45

5.8 Discussion

The aim of the current chapter was to demonstrate the applicability of CAT technology to novel sets of items designed to measure cognitive ability as conceptualised by CHC theory. Background literature suggests that little has been achieved in this domain in recent years, and the current chapter shows the potential to take what we have done as psychologists for 100 years and do it more efficiently without losing measurement accuracy.

Building a CAT requires known psychometric details of numerous items which measure a wide range of the latent trait spectrum. Simulation studies allow researchers to demonstrate how items would perform in a CAT if administered to real examinees. The current chapter details several CAT simulations using items developed as part of this project to demonstrate their accuracy based on different stop rules, and to identify participants who are most likely to be accurately be classified as requiring additional testing after testing. Overall, 52 simulations were executed.

As expected, a stop rule of .00 to .32 SEM required administration of all items and was unlikely to successfully complete (i.e., reach the desired SEM). This occurred regardless of whether the item parameters from the ICS were used, or whether the item parameters were recalculated using participants from the cleaned, pre-imputation data set who had completed all items retained in the ICS. This suggests the item sets in their current form are unlikely to be suitable for high stakes decision making but may be useful in group research where setting a reliability of .70 would halve administration time. Further items are required to be retained from IRT analyses to more accurately measure participants with a wide range of abilities.

Given that initial implementation of these items is anticipated to be a screening tool, a higher SEM (therefore lower reliability) may be acceptable. Considering these item sets are essentially subtests and even the most recognised high stakes cognitive ability test in the world, the *Wechsler Intelligence Scale for Children – Fifth Edition* (WISC-V), has subtest SEM statistics (based on a mean of 10 with a standard deviation of 3) of up to 1.34, which corresponds with an approximate SEM of .45 (based on a mean of 0 with a standard deviation of 1), aiming for a low SEM of .32 seems unnecessary for a screening tool. In fact, setting the stop rule for the CAT simulations at a SEM of .45 (reliability of .80) achieved at least minor reductions in the number of items administered across all CHC abilities; 12% for Working Memory, 54% for Visualisation, 22% for Induction and 64% for Lexical Knowledge. Setting a stop rule for the CAT simulations at a SEM of .55 (reliability of .70) achieved more significant reductions in the number of items administered across all CHC abilities; 52% for Working Memory, 80%

for Visualisation, 66% for Induction and 79% for Lexical Knowledge. For children this would correspond to approximately 37 items in total to measure all four narrow cognitive abilities, with an administration time of approximately 18 minutes. In contrast, the WISC-V Qi system approximates a total test time of 39 minutes for the measurement of the same abilities using the Similarities, Matrix Reasoning, Block Design and Digit Span subtests, and this also requires one-to-one administration by a psychologist.

As expected, there were more substantial reductions when the item parameters were recalculated using only school aged participants. This suggests the items are more likely to discriminate between school aged participants of high and low ability. With a more substantial sample size of school aged participants and additional items across the difficulty spectrum, it would be possible to utilise cutoff stop rules rather than SEM as a stop rule. This would essentially allow an assessor to a priori determine what level of ability they believe an examinee requires and the CAT would administer items until a certain level of classification accuracy is achieved. This is particularly useful for the purposes of screening young people in the school environment.

Due to the lower number of items available and the items largely being at the lower ends of difficulty, each item set would currently only be useful for classification between those that are of very low ability and those that have normal or above average ability. Given that the main aim of this pilot tool is to screen for potential cognitive challenges, this is suitable, but it must be acknowledged that this would not necessarily assist with identifying children with cognitive strengths because the item sets may not quite reliably differentiate between average performance and above average performance. This is demonstrated by the higher correlations found for those simulated participants with a true θ below -3 for most of the CHC abilities. This is consistent with the third and fourth expectation discussed at the beginning of this chapter; because simulated participants were generated with a mean of 0 and standard deviation of 1, there were far more participants at the mean ability level. This means there was more likely to be lower average difference between true ability levels detected by administering whole item sets versus simulated ability levels detected by administering adaptive item sets. In contrast, correlations were higher at the easiest and most difficult ends of the spectrum.

Taking the above discussion into consideration there are clear limitations and areas for further development or research. The first limitation is the number of items available after the ICS. Given that Rasch was the IRT model chosen, the information provided by each item is the same. This means more items are required across the range of abilities for simulated participants at differing θ levels to successfully meet the stop rule. If there are gaps, it is entirely feasible that some participants may be administered all available items in an item set but still not fulfil the stop rule criteria. If development of this CAT tool continues to employ the Rasch model, further item development is required. In contrast, other IRT models could be a possible alternative to determine if existing items which were removed from item sets may be retainable.

The second limitation is the number of school aged people in the sample; the difference in ability across ages differs for young people compared to adults. For example, a 6-year-old’s Lexical Knowledge is nearly always substantially different than an 18-year-old’s Lexical Knowledge, even more so than the difference in ’normal’ Lexical Knowledge between an 18-year-old and a 90-year-old. This causes problems when considering the option of a cutoff score stop rule for CATs. Because this study relies on a snowball sample, there is an innate randomness in the data which causes problems in determining if the items appropriately measure ability at different ages, and thus for our purposes, determining what an appropriate cutoff would be. For example, it may be that this sample includes very high ability 6-year-olds and very low ability 9-year-olds. One must know the average θ for these age groups to better understand how well targeted the items are. Our sample includes only one 6-year-old and one 12-year-old participant, but ten 10-year-olds and 13 18-year-olds. Further data collection is required.

As explored in this chapter, simulation studies are an important tool in the development of CATs. The simulations in this chapter demonstrate that depending on the context of testing, accepting lower levels of reliability can lead to reductions in as many as 50% of items administered. The ability to efficiently measure multiple cognitive constructs in only a short time demonstrates the applicability of IRT and CAT concepts to cognitive ability measurement in Australia. These findings suggest the items analysed in this project may pave a way forward for efficient screening of cognitive abilities in group settings such as classrooms. However, for schools, teachers and psychologists to have faith that these efficient tests are measuring the same abilities known to be important for learning, there must be a demonstration of similar performance on this test as on other mainstream tests.

Having established that each CHC ability under consideration can be measured with decreasing numbers of items as you allow for increasing levels of error, it is important to determine how well the item sets correspond to other similar tests of ability. Given the wide popularity of the WISC-V in an Australian context (James et al., 2015), the following chapter details a convergent validity study using the θ scores as an outcome from the ICS and the same participants' results on the WISC-V.

Chapter 6: Convergent Validity

6.1 Introduction

The last chapter established that the item sets used within this thesis can be administered in a CAT format to reduce the quantity of items exposed to examinees, and thus reduce the time taken to conduct testing. However, it is important to establish that such a tool is measuring what we think it is measuring. While the items developed for this CHC-CAT are based in a sound theoretical grounding, there is a vast literature base discussing the methods of establishing different types of validity, with diverse opinions on the strengths and weaknesses of each method or type. Chapter 2 introduced several concepts within psychological measurement that ultimately relate to the concepts of validity and reliability, however these are not exhaustive. The current chapter aims to briefly discuss current conceptualisations of validity, with a more specific focus on convergent validity, before conducting an analysis of the convergent validity of the developed item sets with the *Wechsler Intelligence Scale for Children – Fifth Edition* (WISC-V).

6.2 Validity

Validity in measurement of psychological constructs is a field of ongoing debate. Early work (now considered seminal in nature) conceptualised validity as “concurrent validity”, “predictive validity”, “content validity”, and “construct validity” (Cronbach & Meehl, 1955), with the first two combined into “criterion validity”. Construct validity at the time was perceived as whether the “empirical relations between test scores match theoretical relations in a nomological network” (Borsboom et al., 2004, p. 1061). Messick (1989) focused on the actual interpretations and actions taken based on test scores, including their social consequences, and the role of values in interpretation, as forms of validity. Messick went as far as to describe validity as singular with six aspects, deviating from Cronbach and Meeh's prior type conceptualisation (Messick, 1995). More recently, Borsboom et al. (2004) argue simply that a test is valid for measuring a psychological construct if (1) it exists, and (2) variations in the construction causally produce variation in the measurement outcomes. Such debates centre on issues related to causality, correlation, epistemology or ontology, and are largely beyond the scope of this thesis.

Regardless of perspectives on validity, intelligence as a psychological construct has substantial evidence. And the validity of a particular tool is something that requires cumulative evidence between theory, instrument and other outcomes (Krabbe, 2017). It is important that work done in previous chapters has assisted in establishing the validity of the item sets as measures of their proposed psychological constructs. This includes evidence in support of the proposed structure of the item sets, as well as the link between item development and underlying CHC theory.

The confirmatory factor analyses (CFAs) completed in this thesis, for example, are clearly based on a conceptualisation of cognitive ability as a reflective model. That is, the psychological construct Lexical Knowledge manifests the participant's performance on individual items. That Lexical Knowledge improves with age suggests there is temporal precedence and a causal relationship between the psychological construct and the measurement tool (Edwards & Bagozzi, 2000). This also aligns with more recent discussions of psychological constructs, such that intelligence is a natural kind of psychological construct (Fried, 2017). The validity of the item sets are intimately linked with the underlying CHC theory which has a substantial literature base (as discussed in Chapter 2); although there are contemporary challenges to the model, it is currently perceived as our best working model (Flanagan & McDonough, 2018).

Outside of establishing a clear relationship between the items developed and an underlying theory, another source of validity for the item sets can be established through convergent validity. Convergent validity is an important aspect of overarching construct validity and relies on a tool's relationship with other tools that supposedly measure the same construct (Krabbe, 2017).

6.3 Validity of Cognitive Ability Tests

Popular and comprehensive cognitive batteries demonstrate convergent validity with several tools. For example, the *Woodcock Johnson Fourth Edition* (WJ IV) (McGrew et al., 2014) considered correlations with the *Wechsler Intelligence Scale for children – Fourth Edition* (WISC-IV) (Wechsler, 2003), *Wechsler Adult Intelligence Scale – Fourth Edition* (WAIS-IV) (Wechsler, 2008), the *Wechsler Preschool and Primary Scale of Intelligence – Third Edition* (WPPSI-III) (Wechsler, 2002)*, Kaufman Assessment Battery for Children – Second Edition* (KABC-II) (Kaufman & Kaufman, 2004), the *Stanford-Binet Intelligence Scales, Fifth Edition* (SB5) (Roid, 2005), and the *Differential Abilities Scales – Second Edition* (DAS-II) (Elliott, 2006). Select correlations are in Table 6-1.

To demonstrate the convergent validity of the WISC-V, correlation analyses were conducted for the WISC-V with the WISC-IV and the WPPSI-IV (Wechsler, 2016). For the WISC-IV there were subtest correlations between .33 (Symbol Search) to .85 (Vocabulary), composite correlations of .64 (PSI) to .84 (VCI), and a Full Scale IQ correlation of .86. For the WPPSI-IV there were subtest correlations between .41 (Picture Concepts) and .77 (Symbol Search – Bug Search), composite correlations of .59 (PSI) to .81 (VCI), and a Full Scale IQ correlation of .91.

Table 6-1. WJ IV Correlations with Other Test Composites

WJ IV Score	Other Test	Correlation
General Ability Index	WISC-IV Full Scale IQ	.86
Comprehension-Knowledge	WISC-IV Verbal Comprehension Index	.75
Fluid Reasoning	WISC-IV Perceptual Reasoning Index	.70
Short-Term Working Memory	WISC-IV Working Memory Index	.72
Processing Speed	WISC-IV Processing Speed Index	.55
Visual Processing	WISC-IV Perceptual Reasoning Index	.55
General Ability Index	WAIS-IV Full Scale IQ	.84
Picture Vocabulary	WPPSI-III Verbal IQ	.46
General Ability Index	KABC-II Mental Processing Index	.72
General Ability Index	KABC-II Fluid-Crystallized Index	.77
Comprehension-Knowledge	KABC-II Knowledge Gc Index	.82
General Ability Index	SB5 Full Scale IQ	.80
Comprehension-Knowledge	SB5 Knowledge (Gc)	.68

6.4 Chapter Aim

From the discussion above it is evident that correlations can vary widely, particularly when tasks have different demands. This variation occurs despite multiple editions of the tests being developed over decades, including thousands of participants being included in normative samples. To demonstrate that the item sets developed as part of the current research and are measuring what they are proposed to measure, correlational analyses were planned between item sets and the WISC-V, like the above analyses. The items developed for this research have different task requirements than the WISC-V. For example, the Lexical Knowledge task requires identification of a picture from an audio presentation of a word, whereas in the WISC-V the Similarities subtest requires more reasoning ability by identifying the conceptual relationship between words, and the Vocabulary task requires the examinee to define the word. Therefore, while all three tasks are theoretically measuring Lexical Knowledge it would be unlikely for the CHC-CAT Lexical Knowledge task to have extremely high correlations with the WISC-V tasks. If the CHC-CAT Lexical Knowledge task is indeed an operationalisation of the same theoretical construct, more modest convergence would be expected. Therefore, it was expected that moderate correlations (between .40 and .80) would occur consistent with the correlations demonstrated between mainstream tests that also have varied task demands.

6.5 Method

6.5.1 Participants

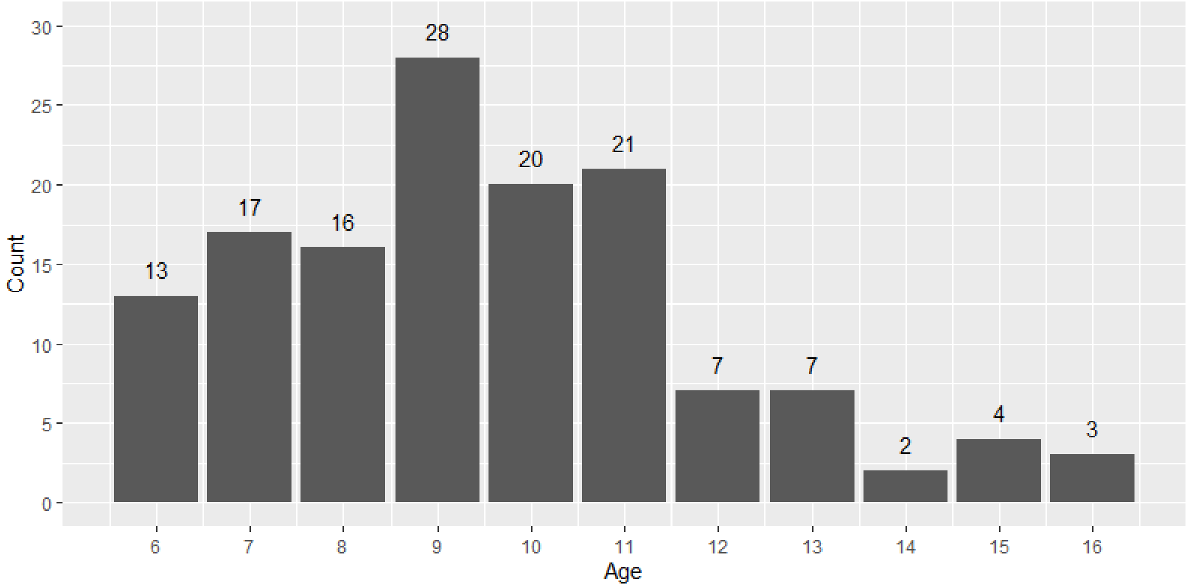
Participants for this study were school-aged students aged 6 to 16 years old. Participants were recruited via snowball recruitment; advertisements were posted online, and school principals were notified of the project. Families interested in participation contacted the researcher to be included in the research; people with suspected or confirmed intellectual disability, and those who had completed a WISC-V in the previous 24 months, were excluded due to ethical considerations.

While 192 participants provided consent to participate, due to a variety of factors only 138 participants recruited completed both necessary aspects of the project to be included in the data set. Drop out reasons included change of school, loss of contact with family, nonattendance to appointments, turning 17 prior to administration of tests, and withdrawal due to recent participation in cognitive ability testing.

The mean age of participants was 9.56 years old (*SD* = 2.38 years old), with 69 females and 69 males participating. All participants were from Australia. The distribution of participants by age is in Figure 6-1. The CHC-CAT was administered via iPad for 126 participants, on a laptop for two participants and data was missing for two participants.

As not all participants completed all individual item sets, demographics are provided in detail in the results section for each unidimensional construct under consideration.

Figure 6-1. Convergent Validity Age Distribution



The outcomes of the WISC-V are relatively consistent with expectations at a group level, as demonstrated in Table 6-2 and Table 6-3. That is, subtests centre around a mean scaled score of 10 with a standard deviation of 3, and indexes centre around a mean of 100 with a standard deviation of 15.

Table 6-2. WISC-V Subtest Performance (n = 138)

Subtest	**Mean Scaled Scale**	**Standard Deviation**
Similarities	9.94	3.23
Vocabulary	10.13	2.98
Block Design	9.05	2.69
Visual Puzzles	9.50	5.36
Matrix Reasoning	9.38	2.60
Figure Weights	9.88	3.05
Digit Span	9.54	2.91
Picture Span	9.88	3.15
Coding	11.64	5.05
Symbol Search	9.73	4.52

Table 6-3. WISC-V Index Performance (n = 138)

Index	Mean Index Score	Standard Deviation
Verbal Comprehension	100.09	14.71
Visual Spatial	94.87	13.38
Fluid Reasoning	97.88	14.28
Working Memory	97.40	13.26
Processing Speed	102.42	16.84
Full Scale IQ	98.86	14.59

6.5.2 Materials

This study consisted of two tools: the WISC-V and the CHC-CAT under investigation as part of this thesis.

The WISC-V is the most utilised cognitive ability test for children by Educational Psychologists in Australia (James et al., 2015). It is an individually administered test of cognitive ability for children aged 6 years 0 months through to 16 years 11 months (Wechsler, 2016). Including the Q-Interactive version of the test (which is administered via iPad as in this study), the test comprises of 21 subtests. Administering the 10 core subtests provides a measure of five specific cognitive areas as primary index scores: Verbal Comprehension Index, Visual Spatial Index, Fluid Reasoning Index, Working Memory Index and Processing Speed Index. Administration of the first seven subtests allows the calculation of a general ability score known as the Full Scale IQ. While other arrangements of subtests can be administered, the current study focused on the core WISC-V subtests as they provide a measurement of abilities commensurate with the proposed item sets in the CHC-CAT, as well as allowed researchers to provide feedback to participants about their results.

The CHC-CAT is the name given to the current item sets under investigation. In summary (see previous chapters for further details) the test contains four sets of items measuring Lexical Knowledge (G*c:*VL), Induction (G*f:*I), Visualisation (G*v:*Vz) and Working Memory (Gwm:Wc). For the current study items for each set were split into anchor Items, Item Set A and Item Set B, consistent with the Item Calibration Study (ICS). Items were administered in the linear order of (1) anchor Items, (2) Item Set A, and (3) Item Set B. It was expected that the anchor items would be of moderate difficulty for the participants in this study as they provided the most robust statistics and ability to discriminate ability levels in adults in the previous Item Tryout Study (ITOS). Item Set A items that from a face validity perspective were much less difficult, and Item Set B largely new items developed for the ICS and were believed to be more difficult. The factor structure of the items is explored in Chapter 3 and Chapter 4.

6.5.3 Procedure

All participants required parental consent to participate. This was emailed to the researcher and testing took place at the participant's school with the school's consent. Each participant was administered the 10 core subtests of the WISC-V and three item sets from the CHC-CAT under investigation.

Ethical concerns about the number of items being administered to young people in addition to the WISC-V meant all four item sets could not be administered (this is noted as an area for further research). Ethical concerns included reducing the amount of time participants were away from class, limiting the amount testing young people were exposed to and, due to uncertainty regarding item difficulty, inability to accurately predict the level of frustration participants may experience and the time required from them.

Both the WISC-V administration and the supervision of the CHC-CAT was carried out by provisional psychologists in their 5th year of psychology training at Monash

University (i.e. enrolled in the Master of Psychology programs). The 138 participants were randomly assigned to a pair of provisional psychologists (total of 26) and testing was completed at a time of their and their school's convenience. The administration order of the CHC-CAT and the WISC-V and the item set selection from the CHC-CAT for participants were both randomised by the provisional psychologists. Subtests from the WISC-V were administered in line with the instructions in the Administration and Scoring Manual (Joshua et al., 2016). As it was expected that the CHC-CAT would be largely automated, provisional psychologists provided little to no guidance to participants on the CHC-CAT other than input of their participant ID (to match up with their WISC-V scores) and to access the item sets. Participants were informed about the varying difficulty of items: "some items may be hard while other items may be difficult" (like the WISC-V) and to try as hard as they could.

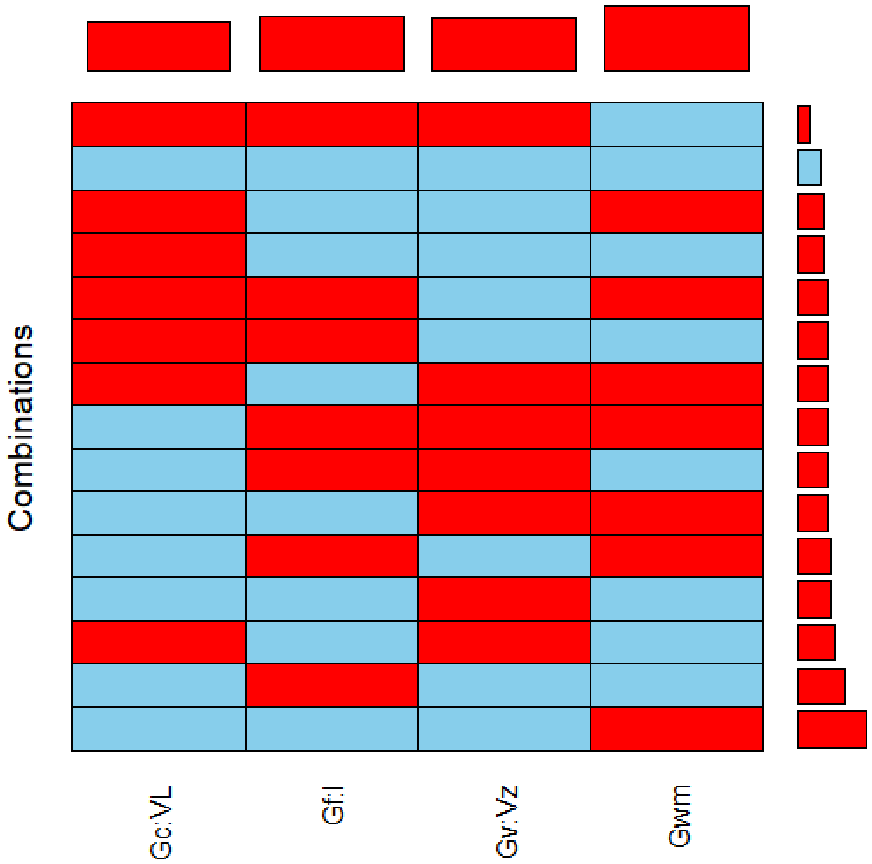
All participants who completed a WISC-V were provided a summary report of their results, written by the provisional psychologist under my supervision.

6.5.4 Data Analysis

For the unidimensional CHC abilities convergent validity was analysed by conducting Pearson product-moment correlation analysis to produce an *r* value for each bivariate correlation (Tabachnick & Fidell, 2007). There are varied guidelines for interpreting the strength of a correlation. Cohen (1988) proposes a small relationship as .10 to .29, medium relationship as .30 to .49, and a large relationship as .50 to 1.0. Hinkle et al. (2003) suggests a negligible relationship as .00 to .30, low relationship as .30 to .50, moderate relationship as .50 to .70, a high relationship as .70 to .90, and a very high relationship as .90 to 1.0. Gignac and Szodorai (2016) analysed correlations from 708 meta-analyses and found that the 25th, 50th, and 70th percentiles corresponded with correlations of approximately .10, .20, and .30 and that less than 3% of the correlations were above .5; this suggests that previous guidelines are too stringent. Therefore, the authors described .10, .20 and .30 as small, typical and relatively large, respectively. The current research will consider these varied perspectives of correlation strength.

The originally planned multidimensional Item Response Theory (IRT) analysis was unable to be carried out. As discussed above, due to ethical considerations all four item sets were very rarely administered. Figure 6-2 displays the pattern of missing data for school-aged participants based on their available θ scores as calculated in ICS using the items retained across the four individual unidimensional item set analyses.

Figure 6-2. Missing Data Pattern for IRT Results.



Note. Red cells indicate missing data while blue cells indicate present data. The bars above and beside the cells represent the proportion of participants with that pattern of data.

This missing data pattern is a result of the randomised item set selection by provisional psychologists as detailed in the procedure section in line with ethical considerations. Now that a smaller number of items has been identified for each item set it would be possible to administer these prior to any additional items from any CHC ability; due to ethical concerns about the number of items to be administered, these smaller item sets can be administered as core items prior to any other items which require further investigation, or are added to the test.

This figure suggests that it was more common for provisional psychologists to randomly select and combine the Lexical Knowledge, Induction and Visualisation item sets together (n = 25). Such a small sample does not allow for appropriate mIRT and CFA analysis. However, it is possible to calculate an approximate *g* factor value by converting each of these participants' individual CHC ability θ scores into a *Z* score (based on the mean and standard deviation of their own age group), summing these *Z* scores (i.e. sum of *z* scores) and re-evaluating the normative performance of each participant. This reflects the WISC-V process of summing subtest scaled scores into a 'sum of scaled score' and converting this into an Index score. This contrasts with methods chosen by other tools, such as the Woodcock-Johnson battery or the XBASS software, which utilises a regression weighted *g* factor (further discussed in Chapter 7). Pearson correlations were then calculated between the standardised *g* summed theta score and the WISC-V Indexes.

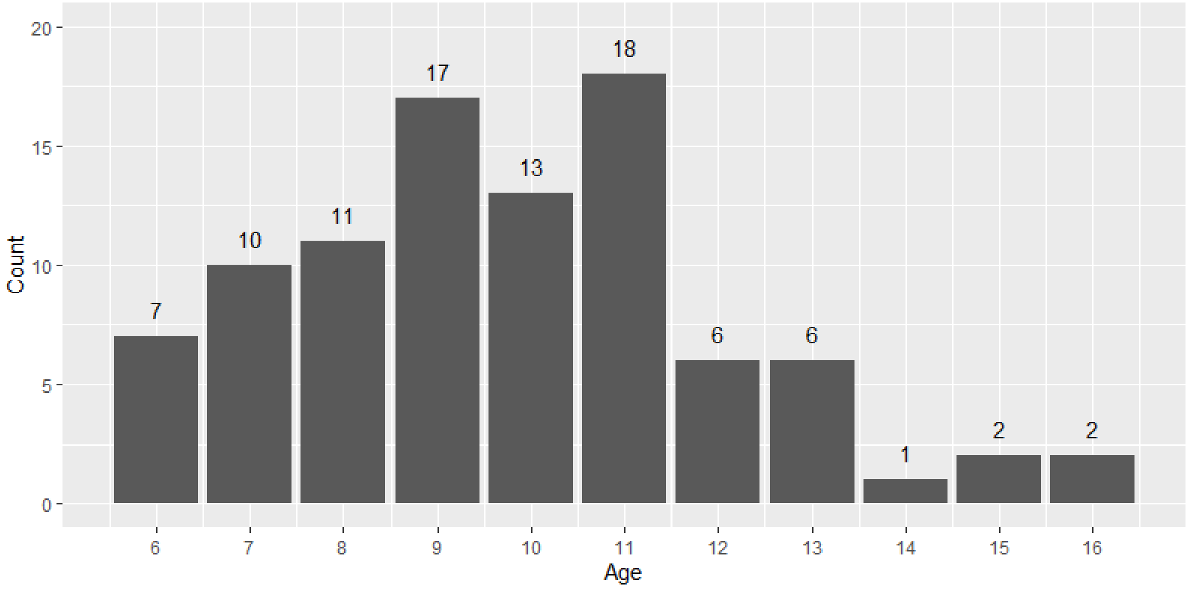
Both the unidimensional and *g* factor analysis were conducted using *R v. 3.6.3* (R Core Team, 2020) in *R Studio Integrated Development Environment v.1.2.5033* (R Studio Team, 2019). No specialised R packages were utilised for calculations, although data manipulation and visualisation packages were used to enhance ease of analysis (Kowarik

& Templ, 2016; Muller & Wickham, 2019; Revelle, 2019; Wickham, 2019; Wickham et al., 2020; Wickham & Henry, 2020). For the Convergent Validity analysis, approximately 500 lines of R code were written (github.com/jakekraska/phd).

6.6 Lexical Knowledge Results

A total of 93 participants (47 females, 46 males) with a mean age of 9.75 years old (*SD* = 2.31 years), attempted items from the Lexical Knowledge item set of the CHC-CAT. iPad was the device used for 84 participants, with seven participants using a laptop and the device variable missing for two participants. The age distribution of participants is shown in Figure 6-3.

Figure 6-3. Convergent Validity Age Distribution (Gc:VL)



The raw score correlations between the CHC-CAT Lexical Knowledge item set and the various WISC-V subtests were generally moderate to high (Table 6-4).

Figure 6-4 shows the relationship between the CHC-CAT Lexical Knowledge raw score and the WISC-V Similarities and WISC-V Vocabulary subtest raw scores.

After standardising the participants' θ scores based on their age, the relationship between the Lexical Knowledge IRT Theta *Z* score and the WISC-V Similarities and WISC-

V Vocabulary scaled scores (Figure 6-5) was low and moderate, respectively, as described by Hinkle and colleagues (2003), and relatively large as described by Gignac and Szodoarai (2016). A relationship of .5 remained when using the Verbal Comprehension Index rather than the individual WISC-V subtests (Figure 6-6).

Table 6-4. Gc:VL Raw Score Correlations

Variable	Correlation with CHC-CAT G*c*:VL Raw Score
Age	.64
WISC-V Similarities Raw Score	.57
WISC-V Vocabulary Raw Score	.64
WISC-V Block Design Raw Score	.59
WISC-V Visual Puzzles Raw Score	.45
WISC-V Matrix Reasoning Raw Score	.56
WISC-V Figure Weights Raw Score	.51
WISC-V Digit Span Raw Score	.62
WISC-V Picture Span Raw Score	.64
WISC-V Coding Raw Score	.68
WISC-V Symbol Search Raw Score	.46

Figure 6-4. Gc:VL Raw Score and WISC-V Raw Score Correlations

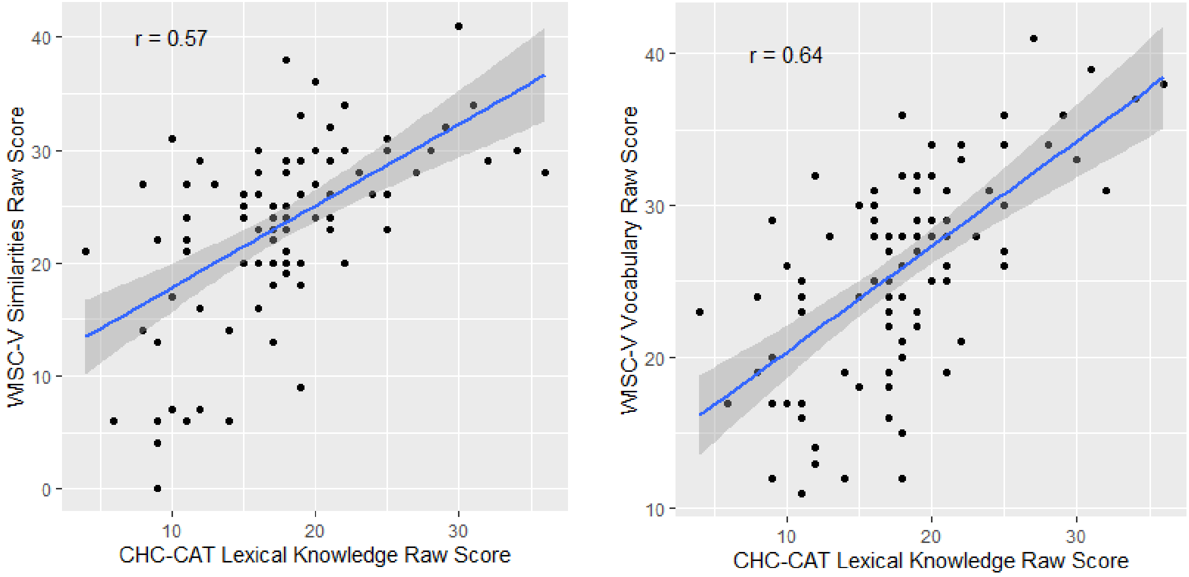


Figure 6-5. Gc:VL Theta Z Score and WISC-V Scaled Score Correlations

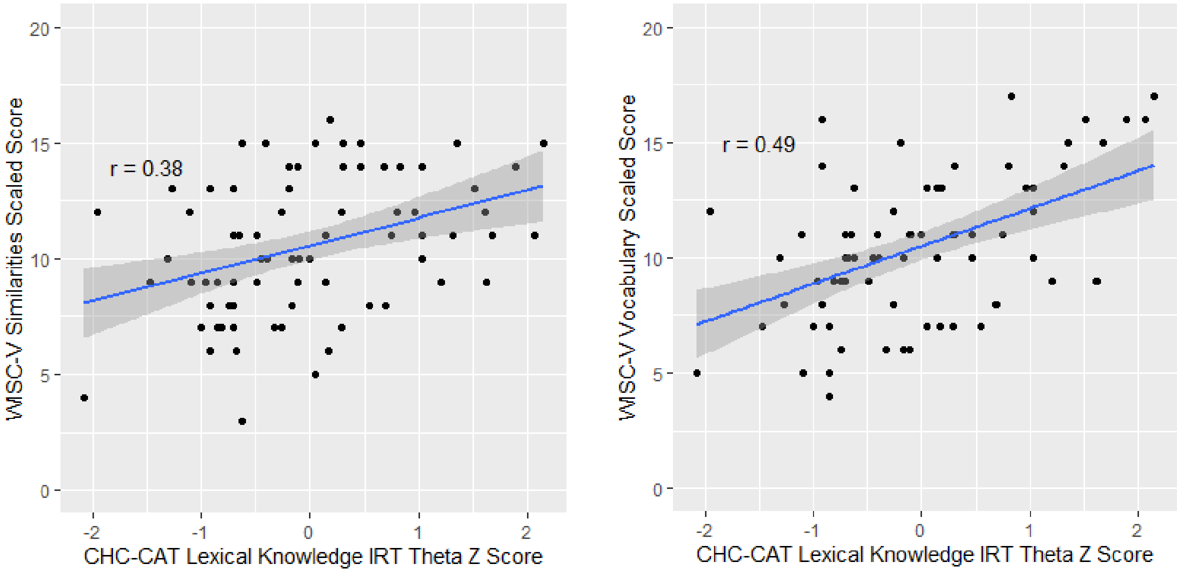
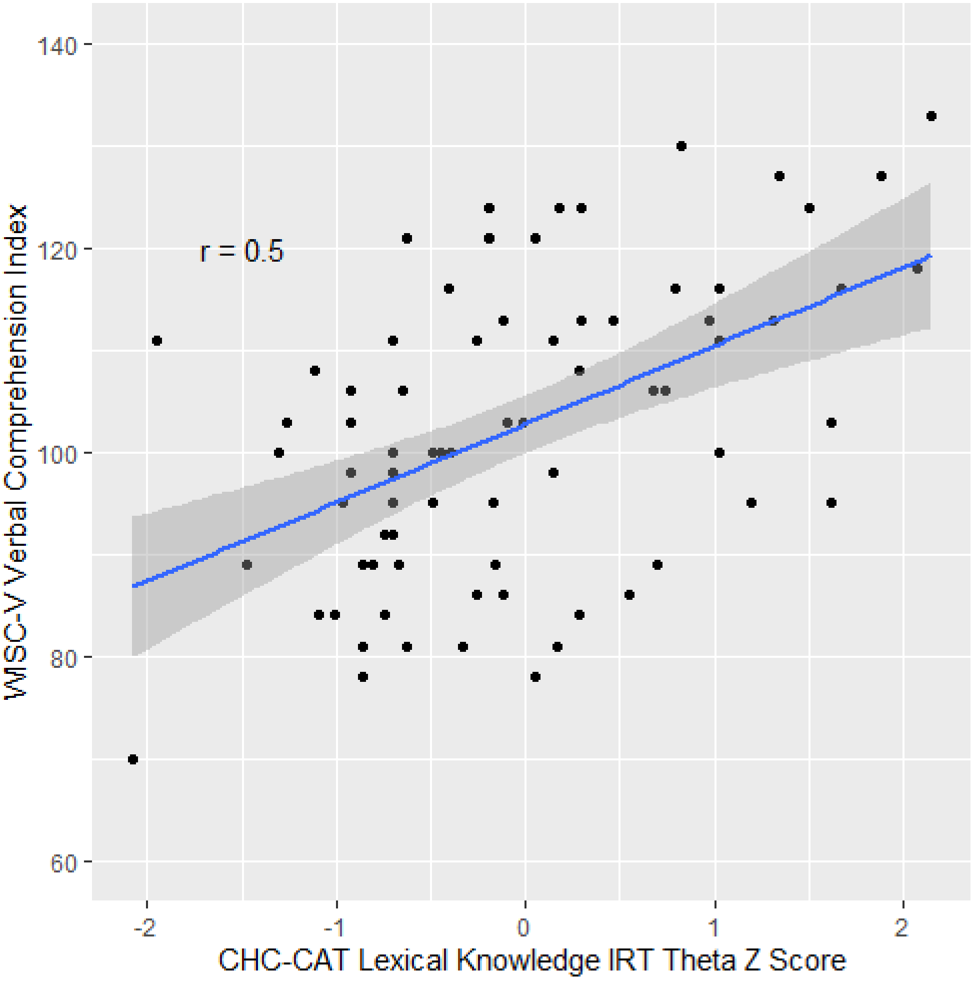


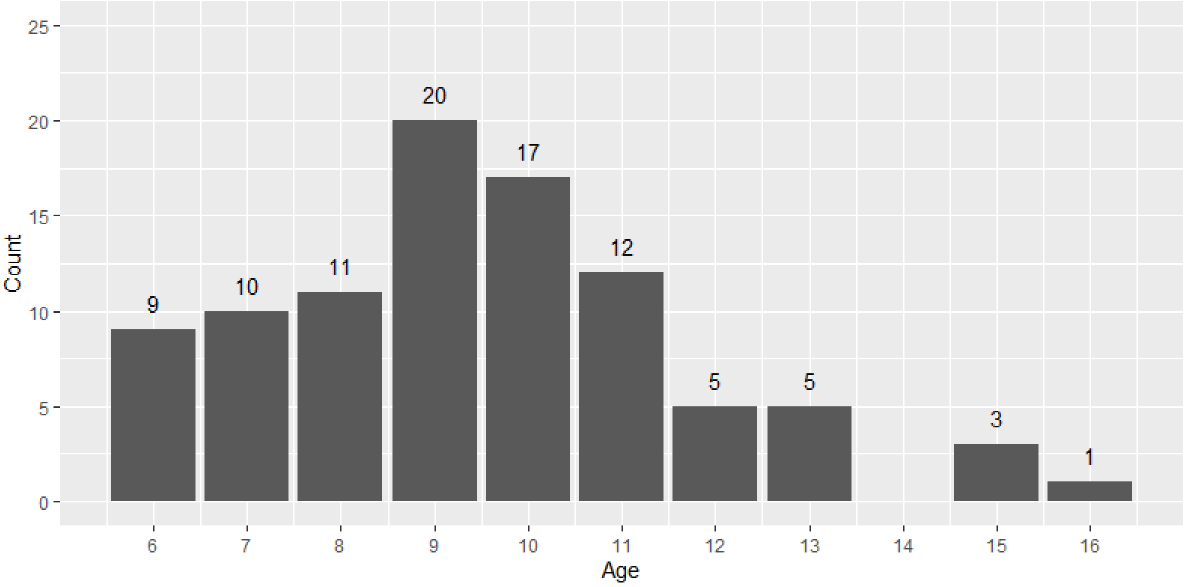
Figure 6-6. Gc:VL Theta Z Score and WISC-V VCI Correlation



6.7 Induction Results

A total of 93 participants (51 females, 42 males) with a mean age of 9.46 years old (*SD* = 2.23 years), attempted items from the Induction item set of the CHC-CAT. iPad was the device used for 86 participants, with five participants using a laptop and the device variable missing for 2 participants. The age distribution of participants is shown in Figure 6-7.

Figure 6-7. Convergent Validity Age Distribution (Gf:I)



The raw score correlations between the CHC-CAT Induction item set and the various WISC-V subtests were largely moderate (Table 6-5).

Figure 6-8 shows the relationship between the CHC-Cat Induction raw score and the WISC-V Matrix Reasoning and Figure Weights subtest raw scores.

After standardising the participants' θ scores based on their age, the relationship between the Induction IRT Theta *Z* score and WISC-V Matrix Reasoning and Figure Weights scaled scores (Figure 6-9) were both weak as described by Hinkle and colleagues (2003) but considered relatively large by Gignac and Szodoarai's (2016) standard. A relationship of .51 was found when using the Fluid Reasoning Index rather than the individual WISC-V subtests (Figure 6-10).

Table 6-5. Gf:I Raw Score Correlations

Variable	Correlation with CHC-CAT G*f*:I Raw Score
Age	.44
WISC-V Similarities Raw Score	.52
WISC-V Vocabulary Raw Score	.58
WISC-V Block Design Raw Score	.58
WISC-V Visual Puzzles Raw Score	.56
WISC-V Matrix Reasoning Raw Score	.57
WISC-V Figure Weights Raw Score	.53
WISC-V Digit Span Raw Score	.56
WISC-V Picture Span Raw Score	.51
WISC-V Coding Raw Score	.54
WISC-V Symbol Search Raw Score	.45

Figure 6-8. Gf:I Raw Score and WISC-V Raw Score Correlations

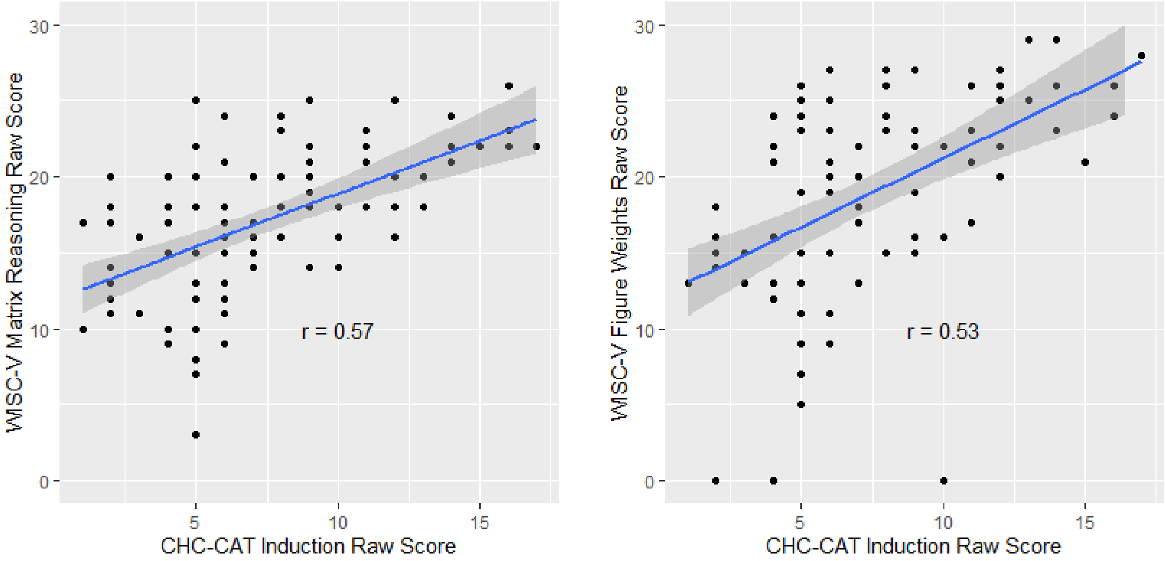


Figure 6-9. Gf:I Theta Z Score and WISC-V Scaled Score Correlations

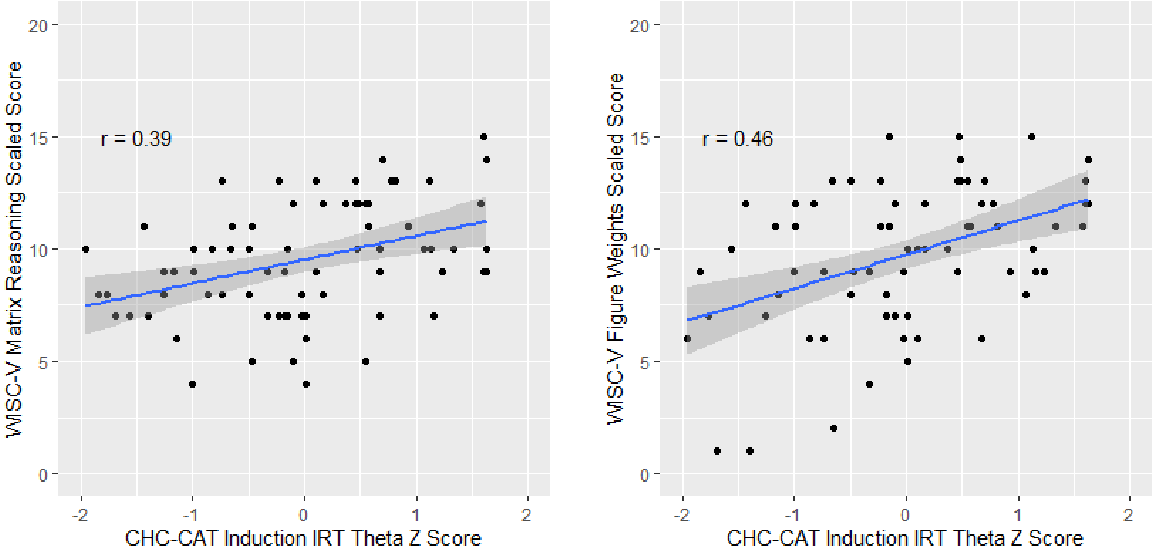
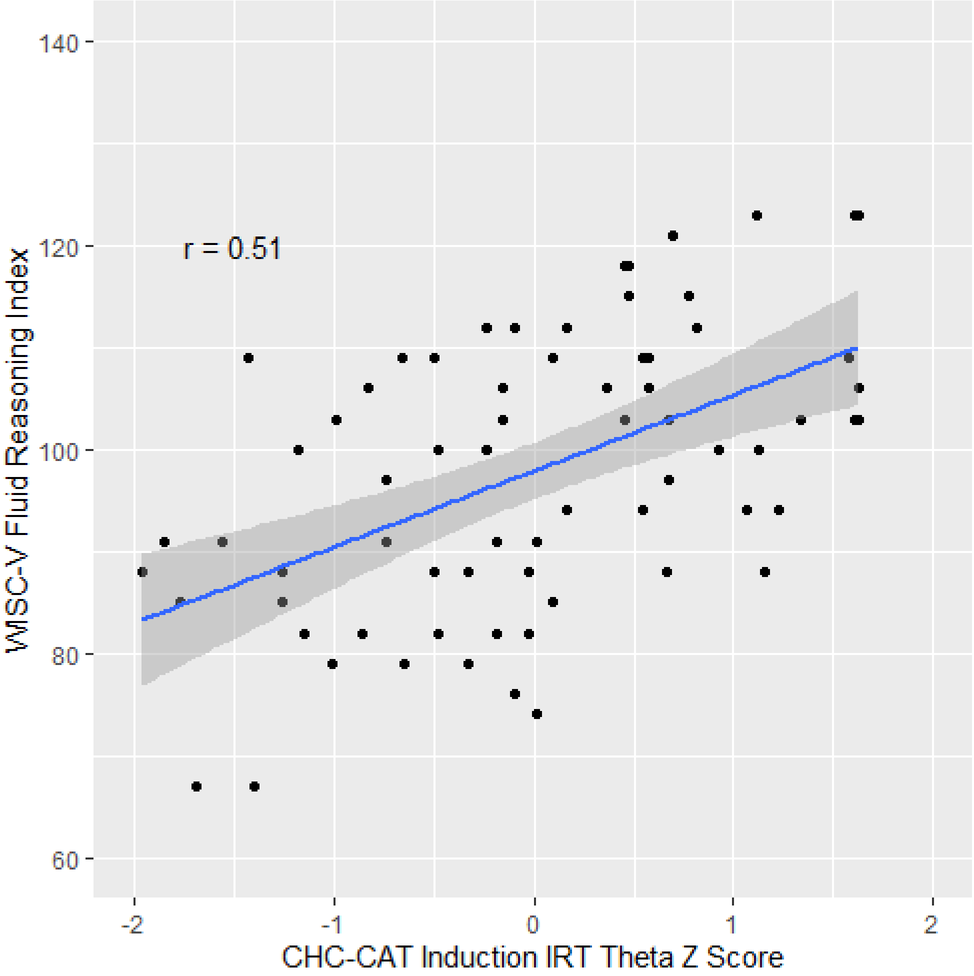


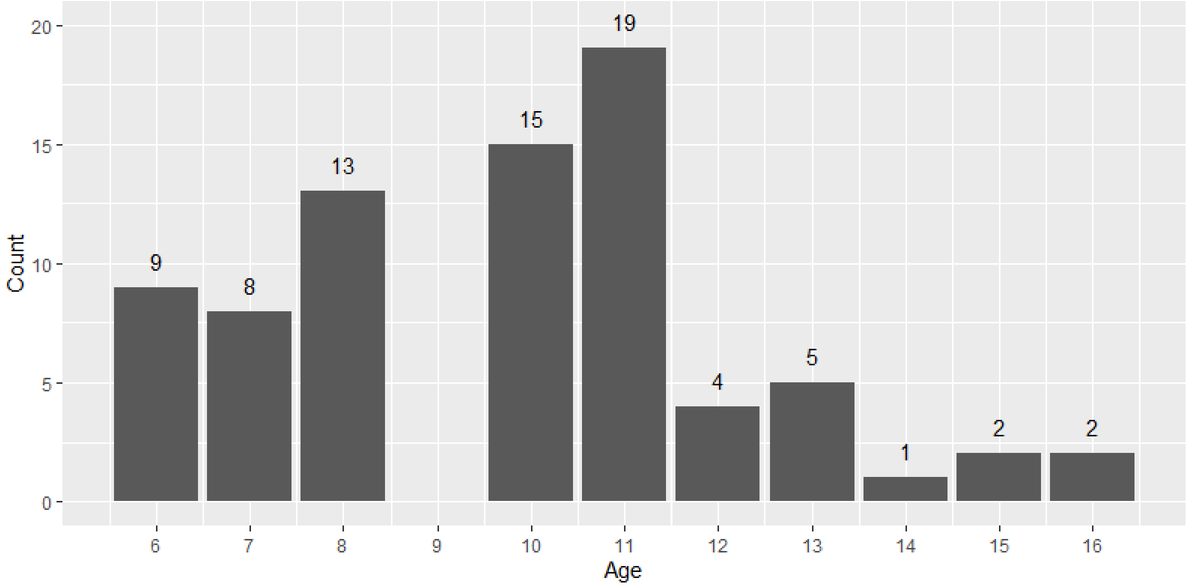
Figure 6-10. Gf:I Theta Z Score and WISC-V FRI Correlation



6.8 Visualisation Results

A total of 99 participants (48 females, 51 males) with a mean age of 9.61 years old (*SD* = 2.24 years), attempted items from the Visualisation item set of the CHC-CAT. The most common administration platform for Visualisation was an iPad, used by 89 participants. Nine participants used a laptop, and one participant did not record which device was used. The age distribution of participants is shown in Figure 6-11.

Figure 6-11. Convergent Validity Age Distribution (Gv:Vz)



The raw score correlations between the CHC-Cat Visualisation item set and the various WISC-V subtests were largely weak to moderate (Table 6-6).

Figure 6-12 shows the relationship between the CHC-CAT Visualisation raw score, and the WISC-V Block Design and WISC-V Visual Puzzles subtest raw scores.

After standardising the participants θ scores based on their age, the relationship between the Visualisation IRT Theta *Z* score and the WISC-V Block Design and Visual Puzzles scaled scores (Figure 6-13) was weak as described by Hinkle and colleagues (2003) but relatively large when considered by Gignac and Szodoarai (2016). A relationship of .31 remained when using the Visual Spatial Index rather than the individual WISC-V subtests (Figure 6-14).

Table 6-6. Gv:Vz Raw Score Correlations

Variable	**Correlation with CHC-CAT G*v:Vz* Raw Score**
Age	.33
WISC-V Similarities Raw Score	.43
WISC-V Vocabulary Raw Score	.51
WISC-V Block Design Raw Score	.46
WISC-V Visual Puzzles Raw Score	.44
WISC-V Matrix Reasoning Raw Score	.53
WISC-V Figure Weights Raw Score	.30
WISC-V Digit Span Raw Score	.41
WISC-V Picture Span Raw Score	.42
WISC-V Coding Raw Score	.33
WISC-V Symbol Search Raw Score	.22

Figure 6-12. Gv:Vz Raw Score and WISC-V Raw Score Correlations

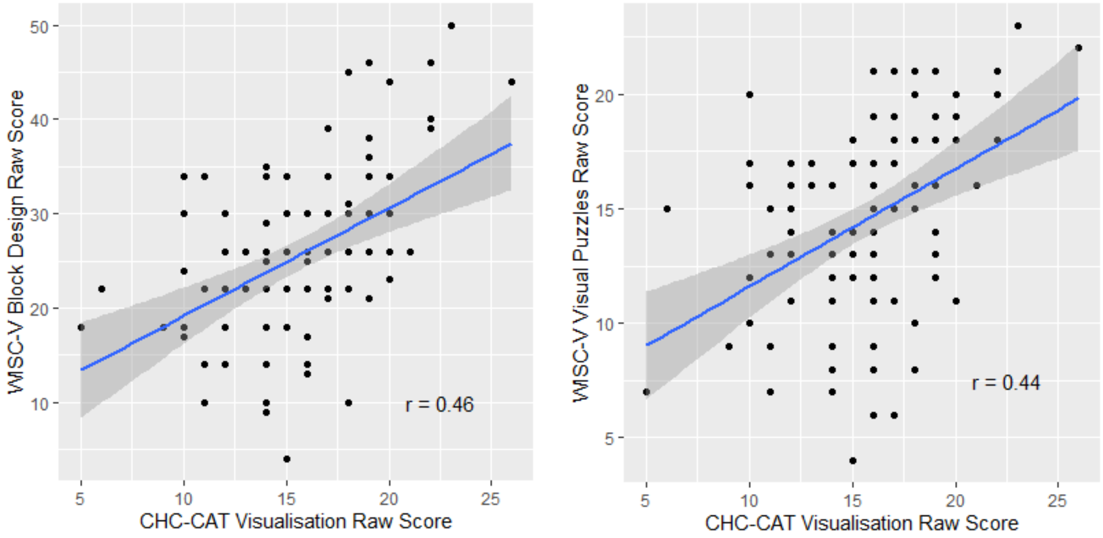


Figure 6-13. Gv:Vz Theta Z Score and WISC-V Scaled Score Correlations

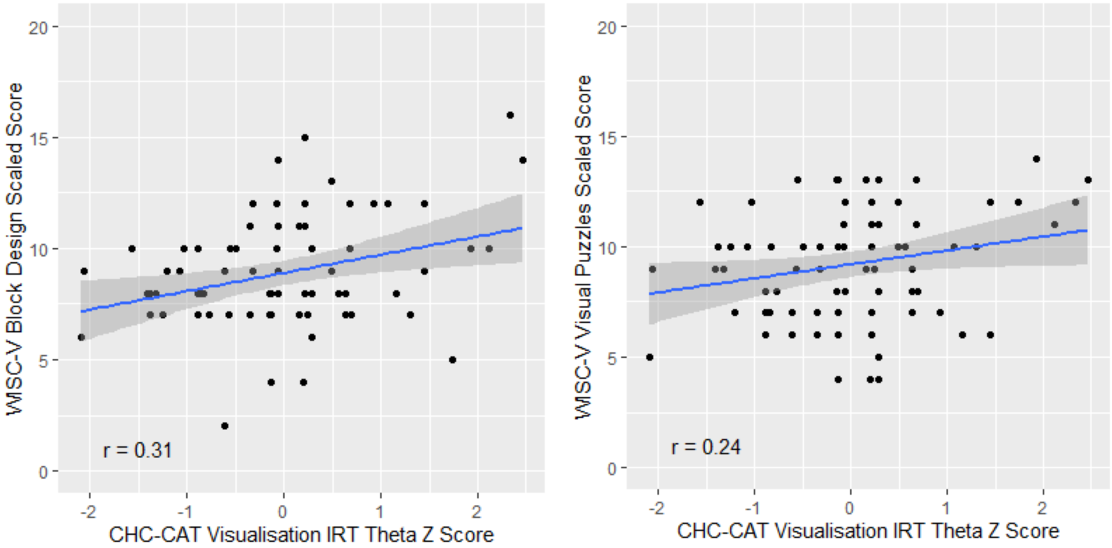
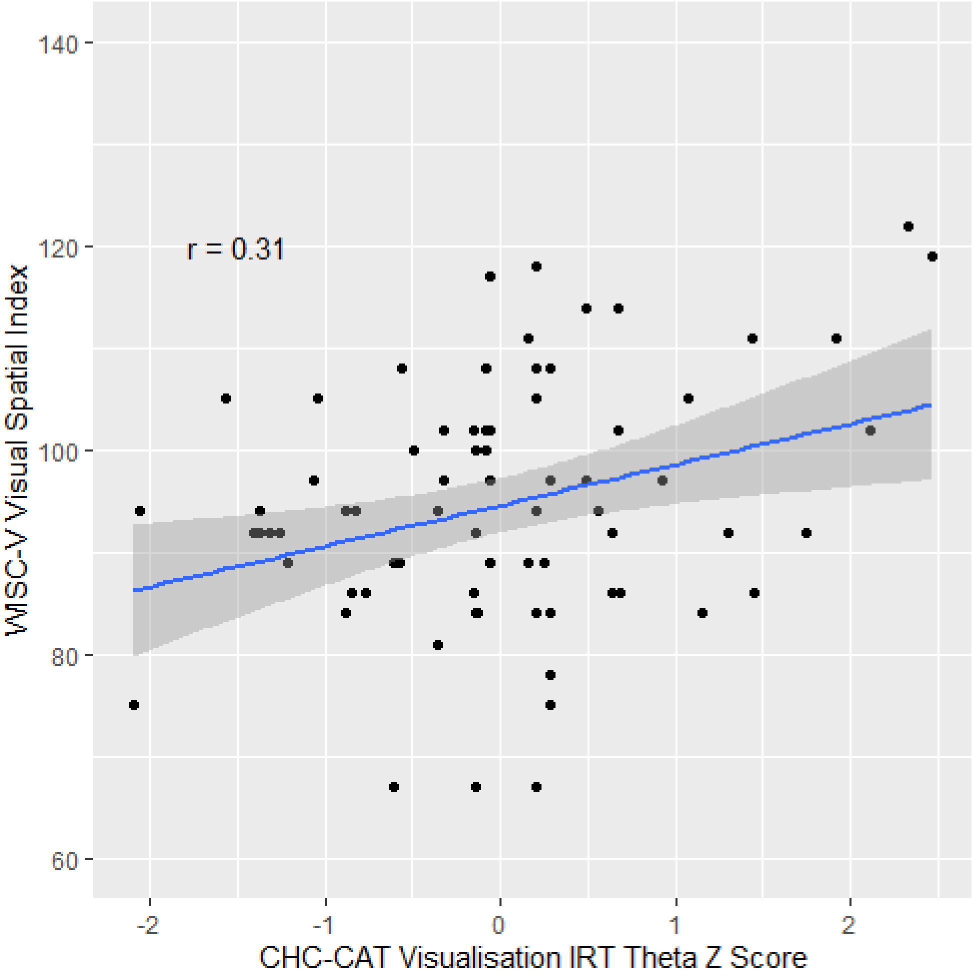


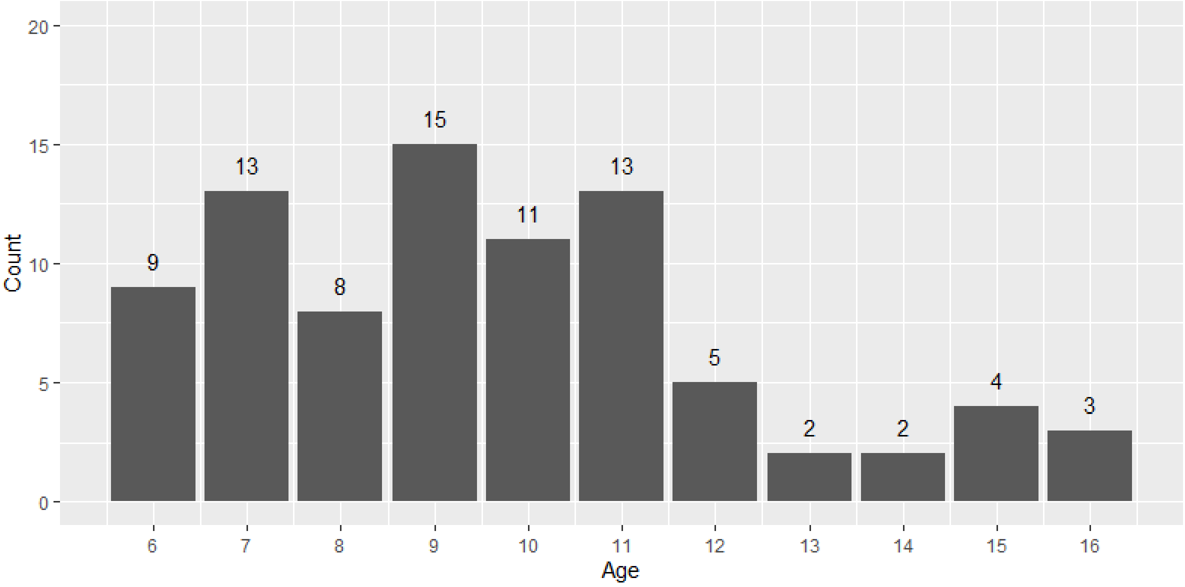
Figure 6-14. Gv:Vz Theta Z Score and WISC-V VSI Correlation



6.9 Working Memory Results

A total of 85 participants (41 females, 44 males) with a mean age of 9.64 years old (*SD* = 2.65 years), attempted items from the Working Memory item set of the CHC-CAT. iPad was the device used for 79 participants, with 6 participants using a laptop. The age distribution of participants is shown in Figure 6-15.

Figure 6-15. Convergent Validity Age Distribution (Gwm:Wc)



The raw score correlations between the CHC-CAT Working Memory item set and the various WISC-V subtests were largely moderate (Table 6-7).

Figure 6-16 shows the relationship between the CHC-CAT Working Memory raw score and the WISC-V Digit Span and Picture Span subtest raw scores.

After standardising the participants' θ scores based on their age, the relationship between the Working Memory IRT Theta *Z* score and the WISC-V Digit Span and Picture Span scaled scores (Figure 6-17) was weak as described by Hinkle and colleagues (2003) but relatively large when considered by Gignac and Szodoarai (2016). A weak relationship remained when using the Working Memory Index rather than the individual WISC-V subtests (Figure 6-18).

Table 6-7. Gwm:Wc Raw Score Correlations

Variable	Correlation with Gwm:Wc Raw Score
Age	.53
WISC-V Similarities Raw Score	.55
WISC-V Vocabulary Raw Score	.61
WISC-V Block Design Raw Score	.62
WISC-V Visual Puzzles Raw Score	.56
WISC-V Matrix Reasoning Raw Score	.67
WISC-V Figure Weights Raw Score	.61
WISC-V Digit Span Raw Score	.66
WISC-V Picture Span Raw Score	.59
WISC-V Coding Raw Score	.66
WISC-V Symbol Search Raw Score	.39

Figure 6-16. Gwm:Wc Raw Score and WISC-V Raw Score Correlations

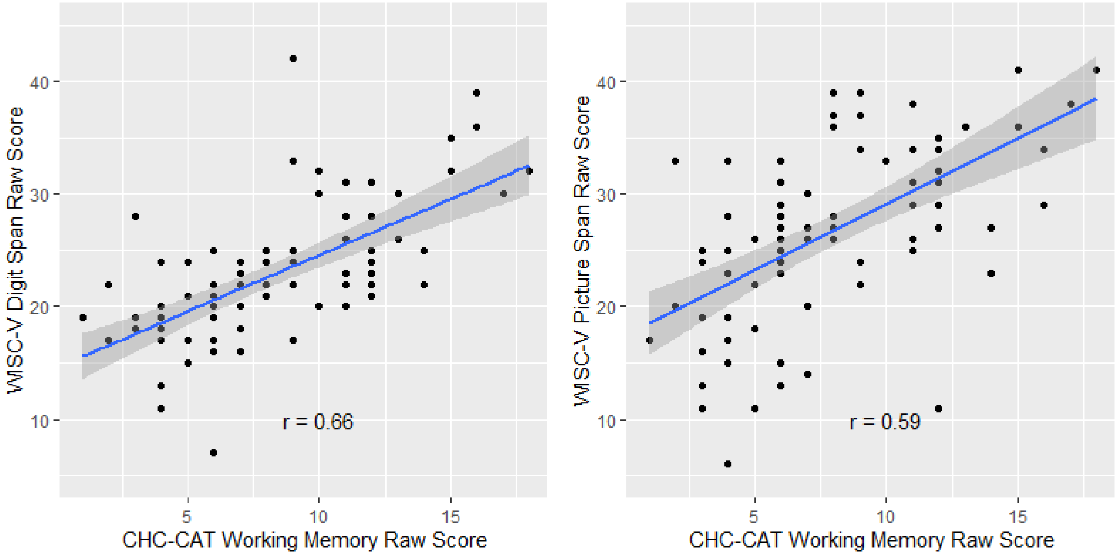


Figure 6-17. Gwm:Wc Theta Z Score and WISC-V Scaled Score Correlations

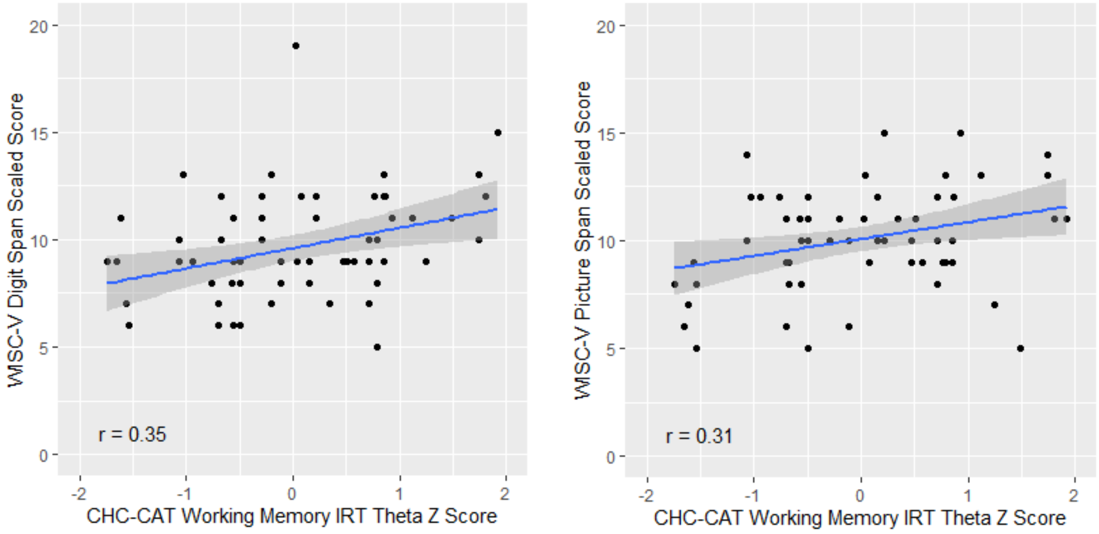
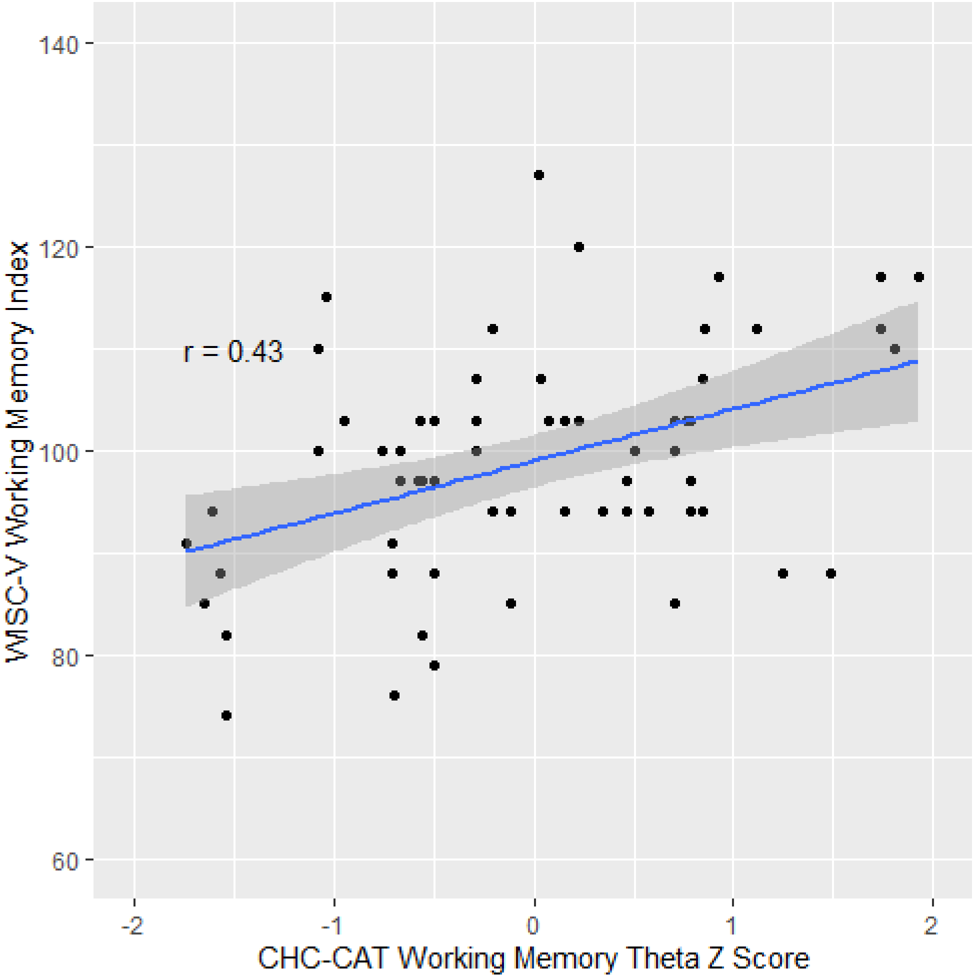


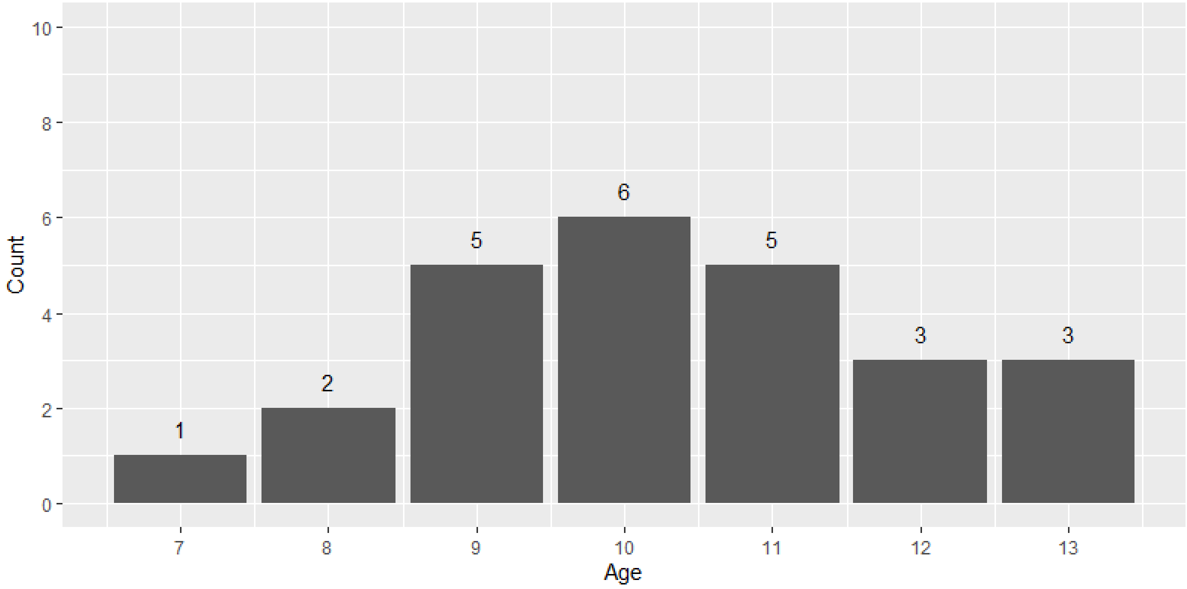
Figure 6-18. Gwm:Wc Theta Z Score and WISC-V WMI Correlation



6.10 Multidimensional Ability Results

While 138 participants underwent WISC-V assessment, due to the breakdown of participants that attempted each item set (as described in the Data Analysis section), only 25 participants (17 female, 8 male) with a mean age of 1.32 years old (*SD* = 1.63 years) were selected for the current analysis. These participants each completed the Lexical Knowledge, Induction and Visualisation items. An iPad was the most common device (*n* = 22) with two participants using a laptop and the device variable missing for one participant. The age distribution of participants is in Figure 6-19.

Figure 6-19. Convergent Validity Age Distribution (g)

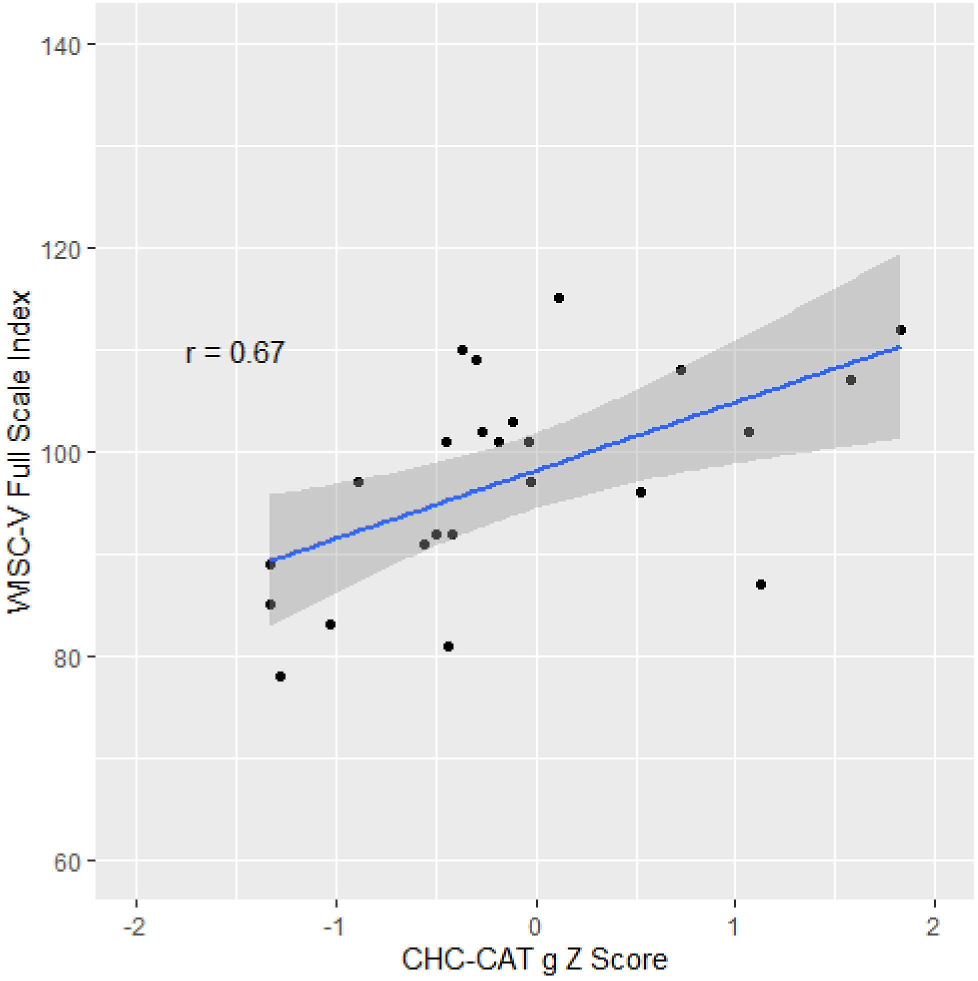


Correlations between the *g Z* score varied based on the particular WISC-V Index (Table 6-8); the strongest correlation was between the *g Z* score and the Full Scale IQ, which showed a moderate relationship as described by Hinkle and Colleagues (2003) but relatively large as described by Gignac and Szodoarai (2016) (Figure 6-20).

Table 6-8. Correlation Matrix for CHC-CAT g Z Score and WISC-V Indexes

	***g Z* Score**	**VCI**	**FRI**	**VSI**	**WMI**	**PSI**	**FSIQ**
***g Z* Score**	1.00						
VCI	.50	1.00					
FRI	.38	.67	1.00				
VSI	.57	.41	.49	1.00			
WMI	.45	.40	.40	.35	1.00		
PSI	.49	.24	.17	.16	.52	1.00	
FSIQ	.67	.84	.83	.60	.67	.42	1.00

Figure 6-20. CHC-CAT g Z Score and WISC-V Full Scale IQ Correlation



6.11 Discussion

The purpose of this chapter was to establish that the CHC-CAT measures cognitive ability consistently with contemporary conceptualisations of intelligence. As the WISC-V has been found to align with CHC theory, it was anticipated that high correlations between the CHC-CAT and WISC-V could provide initial evidence of the validity of the CHC-CAT in terms of CHC theory.

Demonstrating convergent validity provides evidence that two different tests are measuring the same theoretical construct. In psychology convergent validity is important when developing new scales or measures which focus on constructs with existing tools that are robust in nature. The current study considered the convergent validity of the CHC-CAT item sets with the WISC-V at both a raw score and normalised level. Based on the previous convergent validity studies conducted by the test developers of the WISC-V and the WJ IV, it was expected that while there would be variation in correlations (due to differing task demands), there would also be moderate relationships. The findings are summarised in Table 6-9.

When looking at raw scores, Lexical Knowledge, Induction and Working Memory item sets all had moderate relationships with their corresponding WISC-V subtest raw scores. In contrast the Visualisation raw score correlation was weak but approaching moderate. After standardising all θ scores for each CHC ability, the correlation between θ *Z* scores and the corresponding WISC-V subtest scaled scores (Table 6-9) was weak to moderate as characterised by Hinkle and colleagues (2003). When using WISC-V Index scores, the Lexical Knowledge and Induction item sets had a moderate relationship with the Verbal Comprehension Index and Fluid Reasoning Index respectively, and the

Visualisation and Working Memory item sets had had a weak relationship with the Visual Spatial Index and Working Memory Index respectively.

Table 6-9. Summary of CHC-CAT Theta Z Score and WISC-V Subtest Correlations

CHC-CAT Theta Z Score	**WISC-V Composite/Subtest**	**Correlation**
Lexical Knowledge	Verbal Comprehension Index	.50
	Similarities	.38
	Vocabulary	.49
Induction	Fluid Reasoning Index	.51
	Matrix Reasoning	.39
	Figure Weights	.46
Visualisation	Visual Spatial Index	.31
	Block Design	.31
	Visual Puzzles	.24
Working Memory	Working Memory Index	.43
	Digit Span	.35
	Picture Span	.31
g	Full Scale IQ	.67

Stronger correlations between the Lexical Knowledge and Induction item sets were likely related to higher marginal reliabilities and better CFA and Rasch fit (Chapter 4). A better Rasch fit was likely to produce more accurate θ scores; increases in the latent traits Lexical Knowledge and Induction would correspondingly produce higher θ scores in the CHC-CAT item sets, as well as higher scaled scores. That is, test performances would covary more accurately. In contrast, for the Visualisation item set marginal reliabilities were lower, and the item set did not produce as strong CFA and Rasch fit; there is likely a higher level of error in the resultant θ scores. Lower accuracy in θ scores may result in more scatter of scores, in turn reducing correlations.

This observation does not align with the outcomes of the Working Memory analysis (Section 6.9). While the CFA and Rasch fit for the Working Memory item set were relatively positive, the correlations in this study were weak. Those who performed high relative to their peers on the WISC-V Working Memory subtest did not necessarily perform well on the CHC-CAT Working Memory item set, and vice versa. Task demands for the WISC-V Working Memory subtests are very different than those within the CHC-CAT Working Memory item set. At the time of development, items included an attentional capacity component making the CHC-CAT Working Memory task more cognitive complex. For the WISC-V Digit Span subtest examinees know the type of answer they are required to provide after stimulus presentation regardless of what will be presented (e.g. 'digits backwards'). For the CHC-CAT Working Memory items, this is unknown and appears at the end of the item (e.g. "What is the number between blue and red?"). While there are some similarities between this type of task and the Picture Span task whereby there are unknown distractor stimuli, once again the examinee knows the type of answer they are required to provide for the Picture Span subtest (e.g. 'pictures in the order you were shown'). It may be that further research is required to establish convergent validity between the item sets under consideration and other tasks with cognitive complexity built into them (e.g. WJ IV).

The most promising outcome of the current study is the *g* factor correlation with the WISC-V Full Scale IQ. A moderate correlation was found, potentially demonstrating a different operationalisation of the same underlying theoretical construct (i.e. measurement of *g* via automated online testing versus face-to-face administration). It is anticipated that usage of multidimensional IRT models in future studies would further demonstrate convergent validity of the CHC-CAT with the WISC-V at a global intellectual level.

The main limitation of the current study is the sample characteristics for each individual cognitive ability. The majority of participants were aged 6 to 11, and in some cases there were specific ages missing from certain constructs (e.g. 9-year-olds missing from the Visualisation analysis). In addition to variations in participant age, as the sample was recruited in a snowball fashion there was no way to predict the abilities of each participant; given we know that Classical Test Theory (CTT) based tests are less reliable at the extreme ends of difficulties (Weiss, 2011) it is uncertain whether lower correlations in the current study are a result of lack of reliability of high and low ability participants on the WISC-V, or lack of accurate measurement of the CHC-CAT item sets. Areas for future research discussed in Chapter 5 are largely relevant to addressing these issues; further data collection of 11- to 16-year-olds would be particularly helpful.

Ultimately, the convergent validity at a unidimensional construct level appears varied for the CHC-CAT item sets. Comparing the current results (Table 6-9) to those correlations between the WJ IV and other tools (Table 6-1) suggests that outside of the visualisation task, there are some subtests that are approaching correlations seen between other tests. In particular, the *g* factor score is promising, which is useful for screening purposes. The findings, in combination with the ITOS, ICS and CAT Simulation analyses, suggest there is mounting evidence of the psychometric validity of the Lexical Knowledge and Induction item sets. Further analysis of the Working Memory item sets in comparison to other cognitively complex Working Memory tasks and further development of Visualisation items that provide a better fit to the Rasch model are both recommended. The next chapter considers possible conclusions, implications and recommendations of all the analyses conducted in amalgamation.

Chapter 7: Conclusions, Implications and Recommendations

Cognitive ability is perhaps the most researched psychological construct in history. Despite significant refinement of cognitive ability theory throughout the twentieth century culminating in the Cattell-Horn-Carroll (CHC) taxonomy of abilities, the method in which cognitive ability tests are administered has gone largely unchanged. Concerns about this slow progress and lack of innovation have been raised in the literature, yet most advancement has come in the form of theory improvements such as the advent of CHC theory, or simple transition from paper-and-pencil forms to digital forms. While technology has impacted and influenced many aspects of the psychology industry such as data collection, data analysis, conjunctive therapies, psychological intervention, and practice management, there is less evidence of continuing technological improvement in psychological assessment (Barrett, 2018). To address these concerns the current project evaluated the implementation of items developed from a CHC perspective for the purpose of Item Response Theory (IRT) analysis and Computer Adaptive Test (CAT) implementation.

Development of CAT should rely on a solid theoretical foundation for what is trying to be measured. Cognitive ability as a psychological construct has long history (Wasserman, 2018), and has been shown to impact a range of life outcomes (Evans et al., 2002; Floyd et al., 2003; Floyd et al., 2007; Lang & Kell, 2019). Theories of cognitive ability evolved throughout the twentieth century, culminating in CHC theory (Schneider & McGrew, 2018). While there are some contemporary criticisms of CHC in the literature they tend to overgeneralise, fail to acknowledge the unique variance that can be contributed by novel tasks, restrict focus to the applicability of the theory to Specific Learning Disorders, mistake probabilistic causation for deterministic causation in ability/achievement relations, and ignore the mounting evidence of ability constructs being related to neurological functions (Wasserman, 2019). Many of these challenges will probably be only addressed through data analysis of samples that are more diverse and with increased sample size, which likely will only occur with increased efficiency of testing through technological solutions. Kamphaus et al. (2012) suggested that technology may be a key influencer of intelligence test interpretation and development.

Even when implemented via technology, scoring and interpretation of psychological tests tends to rely on Classical Test Theory (CTT) which has several shortcomings. CTT relies on normative samples that may not be able to be generalised, causes difficulties when respondents do not complete all items, assumes a constant standard error of measurement (SEM) across all scores in a population, assumes all items to be of equal weight in the measurement of a psychological construct, and assumes a deterministic linear relationship between observed scores and a true score. IRT, a set of models that focus on probabilistic relationships between latent traits and observed outcomes, addresses many of these problems.

Psychological assessment tends to be dominated by pen-and-paper tests and mainstream cognitive ability tools in Australia tend to either be of consistent design with those developed over 100 years ago, or simple replications of tools on electronic screens which possess many of the same problems as other CTT based tests (Gibbons, 2017; James et al., 2015). While technological and statistical solutions such as gamification, machine learning and ecological momentary assessment have been proposed as options for advancing measurement of cognitive ability, they tend to be atheoretical or lacking in ability to ensure construct validity. While they would address the problem of stagnation in psychological measurement, they discard the importance of psychological theory in developing measures of intelligence.

CATs possess several characteristics that enhance the efficiency of psychological measurement while attempting to avoid compromising test validity or reliability. CATs use adaptive algorithms to select items to ensure precise measurement (Scalise & Allen, 2015), and match examinees' ability levels to items within tests (Gibbons, 2017) by placing them on the same scale, theta (θ), in the same way IRT does. While some tools can be interpreted using IRT (e.g. the *W* scale in the WJ IV) this still requires administration via conventional methods (i.e. linear format). CATs therefore require fewer items to be administered while maintaining high levels of reliability (Weiss, 2011). CATs have been widely utilised in achievement testing, mental health assessment and patient-reported outcomes (Fliege et al., 2005; Gibbons et al., 2016; Martin & Lazendic, 2018) but have not been as widely implemented in measurement of cognitive ability outside proprietary purposes.

Utilising CAT in the measurement of cognitive ability poses several benefits, with novel item types and presentation methodologies available to test developers (Zenisky & Sireci, 2002). CATs also allow for increases in unproctored testing while simultaneously maintaining test security by not displaying all items to examinees (Fetzer & Kantrowitz, 2011). Cognitive ability tests tend to be time consuming for both examiners and examinees, and CATs have found to reduce test time in a range of other domains (Delgado-Gomez et al., 2016; Gibbons et al., 2016). Because CATs are based on IRT, there is also increased opportunity for precision at the extreme ends of the difficulty spectrum (Weiss, 2011) increasing the validity of the test (Cappelleri et al., 2014).

Despite these potential opportunities cognitive ability CATs are not widely available to practitioners.

An evaluation of the literature in Chapter 2 demonstrated that CAT research has thus far failed to address issues of practical implementation. A significant body of methodological research has been carried out regarding different aspects of CATs, but only a very limited number of publicly available sources demonstrate the utility of CATs for cognitive ability testing. Generally, it appears that cognitive ability CATs have been largely adopted by private organisations, militaries and organisations that develop proprietary testing tools. Many of the CATs that do exist in the wider body of literature appear limited in their implementation of contemporary theories of ability (such as CHC), measure a narrow range of abilities or have not sufficiently had their validity and reliability demonstrated. Many of the studies evaluated relied on simulations or lacked details about the implementation of the CAT to be useable by other practitioners or researchers.

Based on the above discussion this thesis set out to design a CHC-CAT screening tool. The goals of this thesis were to explore the application of CATs in the measurement of CHC abilities, explore the efficiency and validity of CATs in measuring intelligence, and comparing these results to mainstream cognitive abilities. After a review of the literature (as described above), we set specific goals to evaluate and calibrate a set of items using CTT and IRT methods to identify four sets of items measuring Lexical Knowledge, Induction, Visualisation and Working Memory, simulate CATs using Rasch item parameters, and consider convergent validity with the *Wechsler Intelligence Scale for Children – Fifth Edition* (WISC-V). The results of the studies designed to address these aims are discussed below, with subsequent consideration of implications, limitations and future research directions.

7.1 Item Analysis and Calibration

The Item Tryout Study (ITOS) in this thesis demonstrated the viability of four new item sets for use in a CHC-CAT. Items for this study were derived from earlier studies that developed items based on CHC theory and consideration of how other mainstream tools have attempted to measure these constructs. It was believed that there was a strong theoretical foundation for these items. Confirmatory factor analysis (CFA), Mokken analysis and Rasch modelling was conducted on these item sets designed to measure Lexical Knowledge (G*c:*VL), Induction (G*f:*I), Visualisation (G*v:*Vz), and Working Memory (G*wm:*Wc). Stringent psychometric cutoffs were set in the analyses which resulted in the removal of a high number of items for all four narrow cognitive abilities. Lexical Knowledge items tended to be too easy for adults, Induction items were not always predictable in their ordering, possible unidimensionality issues were noted with Visualisation items, and there were design issues with the Working Memory item stimuli. As items were developed with the goal of differentiating between below average and average children, ultimately the sample used in the ITOS proved poorly targeted. Despite these concerns, the analysis demonstrated that even with conservative analysis, novel items developed using CHC theory could both meet the underlying assumptions of IRT and sufficiently fit the Rasch model. While the remaining items were unlikely to be suitable for implementation in a CAT, further opportunities for item development, data collection and changes to the analysis methodology were identified.

The subsequent Item Calibration Study (ICS) took advantage of these opportunities. Key developments within this study included development of further items based on CHC theory and related literature. This included consideration of vocabulary difficulty theories for Lexical Knowledge items, rules for Induction items, shape similarity, internal cues and number of shapes for Visualisation, and attentional control and number of information chunks for Working Memory items. Further consideration of standardised item stimuli and presentation was made in order to ensure the valid presentation of items in an online environment. Across the ITOS and the ICS a total of 107 items were developed for Lexical Knowledge, 80 items for Induction, 72 items for Visualization and 44 items for Working Memory. After utilisation of Multivariate Imputation by Chained Equations (MICE) for missing data, an advanced missing data imputation method, and a more prominent focus on Rasch analysis (relative to the ITOS), the ICS retained 47 Lexical Knowledge items, 23 Induction items, 30 Visualisation items, 25 Working Memory items. This was substantially more items than were retained in the ITOS. Improved sample targeting was evident based on age (i.e. futher evidence that older participants were able to perform better than younger participants). Retained items met the assumptions of IRT, fit the Rasch model, measured a wide range of the latent trait and met the recommendation of a minimum of 10-15 dichotomously scored items (Babcock & Weiss, 2013).

Across both the ICS and the ITOS there was not always a clear progression in ability from young to old. This was particularly evident for item sets that had, in retrospect, potential item design problems, or could have benefited from the use of a different IRT model (e.g., Visualisation). Items removed may not be replicated if samples include a larger and more diverse sample than that used in the ITOS and ICS. Many item removeal decisions were based on findings relating to item variance and difficulty, however without a larger group of children it is difficult to prove that these difficulties will stay invariant with all age groups. Further data collection is required.

The original research goal of developing sets of items to measure abilities important to learning was achieved by these two studies. Item development for testing is a complex and costly exercise. Some estimates have placed high stakes test item development at a cost of $1,000 per item and all other items at $300 (Downing, 2006b). At a total of 303 items developed across the ITOS and ICS, this could potentially be valued at $90,900 to $303,000. While there are some established psychometric difficulties with some of these items, even if we only calculate this based on the retained items from the ICS that were utilised in the CAT simulation and convergent validity study, these items are worth $37,500 to $125,000. Considering the item development investment for these items totalled $4200, notwithstanding further item generation and data collection, this has thus far been quite a cost-effective series of studies.

Across the ITOS and the ICS there is evidence of incremental improvements in psychometrics. While slightly different patterns of statistical analysis were utilised in both studies, with the increased number of participants, the utilisation of new items, and the collection of data from supervised participants, the number of items retained increased. The use of MICE in order to address unexpected missing data from the second data collection phase as well as general data cleaning, reduced the available sample for analysis significantly. Additionally, increased cognitive complexity in the newly developed items possibly led to an increase in residual variance (as demonstrated by increased SRMR). Despite this, these items demonstrated good fit with the Rasch model which enabled robust item parameters to be established for use in a CAT.

7.2 CAT Simulation

Despite the ability to vary the characteristics of a CAT, the quality of a CAT is entirely dependent on the items contained within it. If items have poor psychometrics or are lacking in theory-based reasoning, then there is a possibility that the CAT might be efficient but lacking in validity and reliability. On the other hand, if item sets are small or the test developer places strong constraints on the CAT, then there is likely to be little adaptability in the test (Reckase et al., 2019). The simulations contained in Chapter 5 made use of very few characteristics available to CATs, relying on commonly utilised item selection, θ estimation, and first item selection algorithms. The use of varying minimum SEM stop rules allowed us to demonstrate the average number of items when aiming for different levels of reliability. Ultimately, the simulations demonstrated that accepting lower levels of reliability could lead to reductions in as many as 50% of items administered. Additionally, tests were particularly suited to measuring those with lower abilities, which may be useful for classifying those with very low abilities relative to those with average or higher abilities.

7.3 Convergent Validity of the CHC-CAT and the WISC-V

The convergent validity study in Chapter 6 proved positive. All subtest raw scores had moderate relationships with their corresponding WISC-V subtests. After standardising θ scores the correlations between CHC-CAT subtests and WISC-V subtests were weak to moderate. Contrastingly, correlations between CHC-CAT subtests and WISC-V Indexes were moderately strong for Lexical Knowledge and Induction but were again weak for Visualisation and Working Memory. These findings are believed to be related to better fit and higher item counts for the former subtests, as well as item format and increased cognitive complexity in the latter subtests. These outcomes also met the expectation of a .4 to .8 correlation based on an evaluation of other tools convergent validity. While concern was noted regarding gaps in the ages of participants included in the study, the results of this chapter in conjunction with the other chapters not only provided mounting evidence of the psychometric validity and reliability of each item set, but also demonstrated further opportunities for development and research.

The originally planned multidimensional IRT analysis was unable to be carried out for two reasons: Firstly, having planned the item administration order of the ICS on the outcomes of the ITOS, the level of missing data in the second study (ICS) was unexpected. When the ITOS was completed there was only minimal missing data and this expectation carried over to the ICS; the splitting of the item sets into Item Set A, Item Set B and an anchor item set was also designed to ensure that key items were completed prior to possible drop off. Missing data was also an issue for school aged participants, as ethical considerations meant all four item sets were very rarely administered consecutively. Unfortunately, the amount of missing data in the ICS was higher than expected and meant that if the study were to utilise only participants with full item level data for the retained ICS items, then there would be less than 10 participants to run both multidimensional CFA and IRT analyses. Even if the analyses were far less conservative and attempted to simply use any participant who had attempted all four item sets, there would only have been approximately 500 participants for analysis. Additionally, several items had over 90% of missing data meaning there would have been a need to change the items retained relative to the ICS.

To address this there was an attempt to create an estimate of *g* using a sum of subtest *Z* scores and standardising (not averaging) this sum for each age group. The problem here is that the correlations between each of the CHC-CAT subtests may have poor or strong correlations which can result in significant variations of such ‘composite’ scores. As correlations between the subtests approaches 1.00, it is more likely that the standardised sum of subtest *Z* scores would align with the average of the subtest *Z* scores. Alternatively, as the correlations between subtests become lower, it is more likely that the standardised sum of subtest *Z* scores would deviate from the average of the subtest *Z* scores.

Another way to complete this in future studies would be to develop a regression weighted *g* score (like that utilised in *XBASS* and the *Woodcock-Johnson*). While this may be more psychometrically defensible in future studies (particularly if a study analyses the convergent validity between the CHC-CAT and the *Woodcock-Johnson*), the method utilised in this project was chosen because the WISC-V also uses a sum scored method. Nevertheless, correlations between the CHC-CAT *g Z* Score and the WISC-V Full Scale IQ was moderate. It is anticipated that further data collection will enable significant improvements in the estimation of a *g* factor for the CHC-CAT.

7.4 Implications

The studies carried out in this thesis have direct implications for the use of technology in the measurement of cognitive ability. While it is not yet recommended that the items be utilised for the purposes of high stakes testing, there is emerging evidence for the use of the CHC-CAT for group level testing. This would be particularly useful for group-based screening in order to assist in appropriate referral to Educational Psychologists for more comprehensive testing.

Technology use in psychological practice and research has not proliferated without some authors raising concerns about the ethical and competent use of such technologies. Concerns have typically related to email communications, use of smartphones and tablets, storage of client information on the cloud, and psychologists' personal and professional use of social media (Gamble & Morris, 2014) which is somewhat unrelated to the current project. Particularly with the early uptake of the Internet as a tool for research, academics became concerned about many factors relating to the methodology of studies (e.g. Im & Chee, 2004). Given the hosting of the CHC-CAT via Concerto, a web-based platform, and the recruitment of some participants via social media platforms, it is important to consider the ethical challenges posed by such methodologies.

There may be concerns in relation to the online features that a tool such as the CHC-CAT offers. For example, while the CHC-CAT is ultimately planned to be implemented in group-based screening under the supervision of teachers or psychologists, it *could* be administered in an unproctored format such as was done in certain parts of the current project for the purposes of data collection. In fact, many colleagues raised concerns with me throughout my candidature about the potential removal of psychologists from the measurement of cognitive ability. While this is certainly not the intention of this project, this concern is not unwarranted; Computer and Internet technologies are so diffuse that automation of job tasks is expected to impact all jobs in Australia in some way over the next 30 years (AlphaBeta, 2017). However, automation is also expected to boost Australia's national income by $2.2 trillion dollars and reduce the average Australian workload by two hours of manual labour a week.

Increasingly, with the use of virtual therapists, telecounselling, psychology apps, and automated assessment it is easy to see the potential impacts on the field of psychology. Natural language processing computer programs have been presented as "automatic, robot-like counselling" as early as the 1980s (Barak, 1999). Innes (2017) goes as far as to suggest that the threat to the employment of humans as psychologists does not necessarily rely on consciousness, self-awareness and feelings. The CAT simulations in this thesis demonstrate that with improved psychometric outcomes (which would likely be achieved with further data collection and funding) an automated test of cognitive abilities is feasible. Although completely automating this process is not desirable due to the loss of important behavioural observations and the high stakes nature of one-to-one cognitive ability testing.

Despite these concerns, many argue that technological improvements will increase productivity rather than replace workers (e.g. Hamid et al., 2017; Jha & Topol, 2016; Zammuto, 2018). Innes (2017) acknowledges that there is currently only a 0.43 per cent chance of psychologists being replaced by robots. Additionally, research in other fields suggests that artificial intelligence and machine learning will enhance worker productivity rather than cause substantial job losses. It is possible a CAT screening tool (such as developed in this thesis) is more likely to be used for enhancing referrals rather than making high stakes decisions. There is also utility for such a tool in situations where psychologists may not be available for face-to-face assessment via paper-and-pencil or iPad assessment tools. This may be due to a lack of psychologists in rural localities, or even more recently in response to the COVID-19 pandemic where there were increasing demands for teleassessment tools (Pearson, 2020b). Ultimately it is hoped a CAT that measures cognitive ability will be used within the confines for which it was designed, as well as with due regard for its psychometric strengths and weaknesses.

Another concern that may be raised in response to the design decisions in this thesis is the possible impacts of computer system differences between participants. The differential item functioning (DIF) in this study generally suggested that platform differences did not impact probability of correct response. Where there was DIF flagged, it was generally a result of there only being a very small number of participants who used a particular device type, and thus further research is required to further establish these findings.

Further, other authors have challenged such concerns. Gosling et al. (2004) addressed six preconceptions about Internet methods, showing that compared to traditional samples, Internet samples are just as diverse, do not differ in signs of maladjustment, perform similarly across presentation formats (i.e. different websites), and generally produce outcomes consistent with traditional methodology studies. They concluded "Internet methods are of at least as good quality as those provided by traditional paper-and-pencil methods" (p. 102). More recently, Daniel and colleagues (Daniel, 2012a, 2012b, 2012c, 2013a, 2013b; Daniel et al., 2014) demonstrated equivalency of the iPad version of a range of Pearson published tools and their digital counterparts. While this is specifically for one platform (i.e. iPad), the fact that there were so few issues across these studies demonstrates the viability of translating tests to different platforms when taking due consideration for image quality, connection speed, resolution and webpage formatting.

In terms of the reliance of Internet-based recruitment for studies such as, early research demonstrated this was not an issue. Hamilton and Bowers (2006) compared the US population census from 2000 with Internet data from 2003 and showed that while education and income are "generally higher among Internet users" (p. 822), ethnic and gender diversity is comparable, and all age generations are represented in the Internet data. Recent data suggests that in the US Internet use is almost ubiquitous (Pew Research Center, 2017). A similar trend can be found in Australia with 89% of adults using the Internet (Australian Bureau of Statistics, 2012, 2017a, 2017b; Australian Communications and Media Authority, 2017; Ofcom, n.d.). In Australia (population of 24.99 million) there are at least 14.7 million Internet subscribers, as well as 27.0 million mobile users with Internet connectivity (Australian Bureau of Statistics, 2018). This is reflected in the significant growth of online advertising expenditure, against shrinking TV and print media (Australian Communications and Media Authority, 2017). As has been demonstrated in this thesis, internet recruitment allows for the collection of large samples, however the resultant lower quality of data and possible missing data should be expected. This study addressed these issues via MICE. Ultimately, with appropriate design and statistical tools, Internet recruitment is a powerful avenue for data collection and has only supported the outcomes of these studies in such a short time period.

The methodology utilised in this thesis also demonstrates the ability to balance both CTT and IRT statistical methodologies in test development. In line with previous research, the psychometrics in these studies generally improved or deteriorated in response to increases or decreases in item and participant counts, as well as in response to improving the alignment of items with psychological theory. There are many moving parts when attempting to develop items, test them with samples, analyse them with competing statistical methodologies and then implement them in a useful fashion. Changes in item development can impact all the subsequent steps, but also how one plans to implement the items must be considered for earlier steps. This thesis demonstrates how iterative and cyclical item development, analysis and implementation can improve item psychometrics and result in robust item sets. Positively, IRT and CAT are suited to introduction and removal of items over time, whereas CTT requires complete revalidation of item sets when each single change is made.

7.5 Limitations

Unfortunately, the data sets in these studies possessed missing data in a pattern that meant a substantial number of participants and items had to be removed from analyses. While MICE is an advanced method of addressing missing data, it is not a silver bullet. This is a flaw both in the design of the data collection largely due to ethical considerations, and the lack of ability to gather a larger sample of supervised school aged participants. However, this data is not permanently lost; by gathering further data that is more complete it may be possible to retain these items and participants in future analyses. In fact, data collection of school aged participants is ongoing, and reanalysis of data is planned. Additionally, time limits and item sets can be more targeted towards samples in future research now that there is a better understanding of the item psychometrics from the ITOS and ICS. Having participants complete certain items that possess more missing data may improve the influx-outflux analyses and improve the Random Forest prediction of missing data.

When setting out for this research, there was a desire to make use of multidimensional IRT. As a contemporary statistical methodology that allows the implementation of hierarchical or bifactor models in an IRT framework, this would be particularly useful for items developed from a CHC perspective (a hierarchical model of intelligence). Unfortunately, as the project progressed ethical considerations limited the collection of data from young people, and expected quantities of data collection in the

ICS based on the ITOS proved incorrect, likely due to a push in advertising for young adults to reduce the mean age of adults. This resulted in an inability to analyse a hierarchical model for the CHC-CAT using the four narrow abilities under investigation from both a CFA and multidimensional IRT perspective. While the attempts to build a *g* estimate in Chapter 6 demonstrated some viability of the CHC-CAT as a measure of general intellectual ability, further data collection is warranted to enable proper analysis and ultimately a method of scoring a *g* factor.

DIF was flagged for a few items across the item sets within the ICS. Many of these were identified as being caused by small numbers of group membership. For example, in the Working Memory item set there were only nine non-Australian participants and only three of those nine obtained an incorrect answer. This is less of an issue for a test designed for use in Australia. Similar patterns of group membership were found for those that identified as non-binary gender or used a less common device when completing the test. The alternative possibility is that these differences are not due to sparse data, but due to true differences in probability based on group membership. In order to explore these findings, further data collection, particularly from diverse groups, will be required.

The current thesis relied heavily on the Rasch (or 1PL) model. The benefits of this is that items analysed are easily comparable and the item parameters are easily interpretable (i.e. only need to understand difficulty parameter). Rasch literature tends to suggest that if there is misfit, then items and data should be altered and manipulated, which was the approach taken in this thesis. There are of course drawbacks to this, whereby there is a significant loss of data.

The Rasch model is a quite conservative model. If items and scales can demonstrate that they fit the Rasch model, then it is likely that the 2PL and 3PL will also demonstrate good fit. Attempts at using the 2PL and 3PL were not attempted in this thesis as comparisons of the effectiveness of different models was not the focus of the studies in this thesis. Instead, the focus was placed on the difficulty and targeting of items. It was desireable to have stability with a more conservative model for items that are designed for high stakes testing before implementing less conservative models. Once stability has been achieved a 2PL or 3PL model could be utilised to reduce testing time and increase accuracy of an implemented CAT.

The IRT literature does explore the possibility of altering the chosen IRT model to see if perhaps another model fits. This is a particularly warranted consideration for the Visualisation subtest; having a dichotomous response option (i.e. 'same' or 'different') means there is in fact a 50% chance of guessing correctly. It was this item set that demonstrated some problems with Rasch fit and it may be that the 3PL model, which includes a guessing parameter, may have provided better fit to the data. This can be explored in a study that compares the fit of models using the same sets of items. Positively, if there was a call for changes to the models used for different item sets, this could easily be implemented by the software utilised in the current thesis; both *mIRT* in *R* and the *Concerto Platform* allow for different IRT models in each item set utilised.

7.6 Future Directions

There are some practical considerations for clinicans based on the outcomes of these studies. While the measure requires further research, if practitioners wanted to use the scales in its current form, the focus should be placed on those scales that demonstrated more robust statistics (i.e. Lexical Knowledge and Induction). It would be anticipated that at this stage, clinicians would require the support of either IT professionals in order to implement the CAT platforms, or the support of academic staff. Group testing is strongly recommended over individual testing, and only with the goal for screening or confirming concerns. The strength of CATs are that *apriori* information could be implemented; given that the CAT simulations discovered significant reductions in time administration, it would be anticipated that even further reductions in time could be achieved by teachers *apriori* estimating students ability as below average, average or above average.

There are several directions for future research using the CHC-CAT specifically. Currently, further data collection is ongoing with a focus on the retained items from the ICS to address the multidimensionality concern. Secondly, further standardisation of the Lexical Knowledge items could be achieved by slowly integrating items that have been funded and designed in a consistent manner rather than relying on open source and 'Creative Commons' licensed images. Thirdly, there is room for further item development to ensure the wide range of latent traits in each narrow ability is being measured; as demonstrated by the iterative improvements in psychometrics between the ITOS and ICS, a robust measurement of wide difficulty ranges with high reliability appears achievable with sufficient funding and research. Fourthly, there are opportunities for implementation of subtests that measure other important cognitive abilities; this may be other narrow abilities so the test can purport to measure broad abilities, or it may be a focus on completely new abilities such as Learning Efficiency (Gl), Processing Speed (Gs), Retrieval Fluency (Gr), or Auditory Processing (Ga). Fifth, CFA and IRT fit may be improved in future studies by stratifying the samples by age into groups based on clear developmental trends; this may account for inherent variability in how cognitive abilities develop as well as potential generational differences in lexical knowledge. Finally, there is opportunity to implement mixed IRT models into the CHC-CAT so that the IRT model used is most appropriate for the CHC ability being measured. This can enhance the reliability and precision of the tool. Ultimately, these future research opportunities may result in a more robust CHC-CAT that could be classified as comprehensive in its measurement, and with such improvements in reliability and validity may demonstrate the appropriate psychometrics suitable for high stakes testing.

More broadly, research into CATs as tools to measure cognitive ability is limited and thus there are varied opportunities for future research. Unfortunately, due to the use of basals and ceilings in contemporary test tools, it is difficult (but not impossible) to simulate CATs using data from these mainstream assessment batteries. As IRT is relatively robust to missing data, depending on the nature of the data retrieved it may be possible to further analyse these tests from an IRT perspective and implement the parameters in a CAT format. Although possible, it seems more likely that progress in CAT research will be via the development of new tools, rather than manipulation of existing tools.

There are also opportunities to further explore the impact of a variety of CAT characteristics on cognitive ability CATs; different stop rules, item selection methods or theta estimation techniques may enhance measurement of cognitive ability. Importantly, given some of the previously discussed concerns surrounding CATs (Newton, 2019) there are avenues for research in Australia surrounding the attitudes of psychologists and teachers toward CATs, how to best inform educational jurisdictions about the strengths and weaknesses of CATs and how to best implement them within school systems in appropriate circumstances. Ultimately, CATs that measure cognitive ability have a significant way to go in Australia before we can expect widespread implementation of technology into cognitive ability measurement; further research can aim to address this lack of innovation.

7.7 Conclusions

While significant advancements have been made in intelligence theory and statistical theory over the last 100 years, we are still conducting cognitive ability tests in much the same way. This has been characterised by some as a stagnation in the testing industry. Technology has impacted nearly all aspects of psychological practice and research, yet in Australia the use of technology in cognitive ability testing has largely been a translation of paper-based tests to iPads administered in the same linear fashion as previously designed. This thesis has explored the applicability of CATs to measure cognitive ability from the perspective of CHC theory. The project set out to measure multiple unidimensional narrow cognitive abilities using a combination of theory and technology that is rarely implemented in practice, and when done so is usually proprietary in nature. The studies in this thesis are transparent in design and align well with principles of Open Science, ensuring opportunities for other researchers to replicate this research or build upon it to develop practical tools. Ultimately this PhD thesis demonstrates the viability of CATs in the measurement of cognitive abilities.

Outcomes of the studies in this thesis were positive. Iterative development and implementation of items showed improved psychometrics and with further funding and research it will be possible to develop a robust test of cognitive abilities that is automated and adaptive. The CHC-CAT can be effectively used as a group based adaptive screening tool that measures four unidimensional constructs of Lexical Knowledge, Induction, Visualisation and Working Memory. The subtests demonstrated weak to moderate correlations with similar tests within the WISC-V, and a derived *g* factor demonstrated moderate correlation with the WISC-V Full Scale IQ, suggesting that the increase in cognitive complexity of items did not compromise the validity of items. Accepting lower levels of reliability for group-based testing or research purposes can effectively halve the number of items administered on average, and results in a potential average administration time of 18 minutes, less than half of the time needed to administer the comparable subtests within the WISC. Although there are many contexts that could benefit from a cognitive ability CAT, given that no such cognitive ability tool appears to exist for schools to screen their students, this is believed to be an ideal environment for implementation due to the ability to classify between those students with cognitive deficits and those with average abilities. The advantages of theoretically driven, precise and efficient measurement ultimately outweigh the sometimes-complex nature of CATs.

The purpose of this thesis was to stimulate research on CATs that measure cognitive abilities. It has been demonstrated that even with a novel set of items, a psychometrically defensible CAT can be implemented in the measurement of CHC abilities. We are now one step closer to a practical solution to such a domain, rather than the focus on theoretical and methodological foundations that previous literature has taken. There are many avenues for further development and research from the findings in these studies. This thesis identified and contributes to gaps in the literature on CATs in relation to measurement of cognitive ability in a novel way. It provides a platform for future, rigorous and diverse studies. The CHC-CAT in this research, or likeminded tests, have the potential to revolutionise the measurement of cognitive ability in our society.

References

Abu-Hamour, B., & Al Hmouz, H. (2018). Cattell-Horn-Carroll broad cognitive ability profiles for dyslexia and intellectual disability. *International Journal of Inclusive Education*, 1-17. https://doi.org/10.1080/13603116.2018.1534999

ACARA. (2016). *Tailored test development studies*. Retrieved 24th May 2020 from https://www.nap.edu.au/online-assessment/research-and-development/tailored-test-development-studies

Adobe Corporate Communications. (2017). *Flash & The Future of Interactive Content*. Retrieved 31st May 2020 from https://theblog.adobe.com/adobe-flash-update/

Aguado, D., Vidal, A., Olea, J., Ponsoda, V., Barrada, J. R., & Abad, F. J. (2018). Cheating on unproctored Internet test applications: An analysis of a verification test in a real personnel selection context. *The Spanish Journal of Psychology, 21*, E62, Article E62. https://doi.org/10.1017/sjp.2018.50

AlphaBeta. (2017). *The Automation Advantage*. AlphaBeta. https://alphabeta.com/wp-content/uploads/2017/08/The-Automation-Advantage.pdf

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. https://doi.org/10.1007/BF02293814

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1). https://journals.lww.com/lww-medicalcare/Fulltext/2004/01001/Controversy_and_the_Rasch_Model__A_Characteristic.2.aspx

Arendasy, M. E., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and Individual Differences, 22*, 112-117. https://doi.org/10.1016/j.lindif.2011.11.005

Arendasy, M. E., & Sommer, M. (2017). Reducing the effect size of the retest effect: Examining different approaches. *Intelligence, 62*, 89-98. https://doi.org/10.1016/j.intell.2017.03.003

Arendasy, M. E., Sommer, M., & Mayr, F. (2011). Using automatic item generation to simultaneously construct German and English versions of a word fluency test. *Journal of Cross-Cultural Psychology, 43*(3), 464-479. https://doi.org/10.1177/0022022110397360

ASC. (2020). *Computerized Adaptive Testing*. Retrieved 31st May 2020 from https://assess.com/adaptive-testing/

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American, 225*(2), 82-91. https://doi.org/10.1038/scientificamerican0871-82

Australian Bureau of Statistics. (2012). *Educational Attainment*. Australian Government,. Retrieved 30th January from http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1301.0~2012~Main%20Features~Educational%20attainment~110

Australian Bureau of Statistics. (2017a). *Household Income and Wealth Levels*. Australian Government,. Retrieved 30th January from http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/6523.0~2015-16~Main%20Features~Household%20Income%20and%20Wealth%20Levels~5

Australian Bureau of Statistics. (2017b). *Population by Age and Sex, Australia, States and Territories*. Retrieved 30th January from http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/3101.0Feature%20Article1Jun%202017

Australian Bureau of Statistics. (2018). *Internet Activity*. Australian Government. Retrieved 27th July 2020 from https://www.abs.gov.au/statistics/industry/technology-and-innovation/internet-activity-australia/latest-release

Australian Communications and Media Authority. (2017). *Communications Report 2016-17*. ACMA.

Australian Research Data Commons. (2020). *ARDC's Nectar Research Cloud*. Retrieved 14th March from https://nectar.org.au/cloudpage/

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research, 20*(1), 40-49. https://doi.org/10.1002/mpr.329

Babcock, B., & Weiss, D. J. (2013). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing, 1*, 1-18. https://doi.org/10.7333/1212-0101001

Baker, F. B., & Kim, S.-H. (2017). *The Basics of Item Response Theory Using R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-54205-8

Balas-Timar, D. V., & Balas, V. E. (2009). Ability estimation in CAT with Fuzzy logic. 4th International Symposium on Computational Intelligence and Intelligent Informatics, Egypt.

Barak, A. (1999). Psychological applications on the internet: A discipline on the threshold of a new millennium. *Applied and Preventive Psychology, 8*(4), 231-245. https://doi.org/10.1016/S0962-1849(05)80038-1

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5-37. https://doi.org/10.1016/j.jsp.2009.10.001

Barrett, P. (2018). The EFPA test-review model: When good intentions meet a methodological thought disorder. *Behavioral Sciences, 8*(1), 5. https://doi.org/10.3390/bs8010005

Bartram, D. (2000). Internet recruitment and selection: Kissing frogs to find princes. *International Journal of Selection and Assessment, 8*(4), 261-274. https://doi.org/10.1111/1468-2389.00155

Bass, M., Morris, S., & Neapolitan, R. (2015). Utilizing multidimensional computer adaptive testing to mitigate burden with patient reported outcomes. AMIA Annual Symposium Proceedings,

Beaujean, A. A. (2012). *BaylorEdPsych: R Package for Baylor University Educational Psychology Quantitative Courses.* In (Version 0.5) https://CRAN.R-project.org/package=BaylorEdPsych

Beaujean, A. A. (2014). *Latent Variable Modeling Using R: A Step-By-Step Guide*. Routledge.

Beaujean, A. A., & Benson, N. F. (2019). The one and the many: Enduring legacies of Spearman and Thurstone on intelligence test score interpretation. *Applied Measurement in Education, 32*(3), 198-215. https://doi.org/10.1080/08957347.2019.1619560

Bench, J., Jacobs, K., & Furlonger, B. (2019). On differentiating auditory processing disorder (APD) from attention deficit disorder (ADD): an illustrative example using the Cattell-Horn-Carroll (CHC) model of cognitive abilities. *International Journal of Audiology*, 1-6. https://doi.org/10.1080/14992027.2019.1682199

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. https://doi.org/10.1037//0033-2909.107.2.238

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Addison-Wesley.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17*(3), 303-316. https://doi.org/10.1177/0049124189017003004

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

Bonifay, W. (2020). *Multidimensional item response theory*. SAGE Publications.

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science, 5*(1), 184-186. https://doi.org/10.1177/2167702616657069

Borsboom, D., Mellenbergh, G. J., & Heerden, J. v. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071. https://doi.org/10.1037/0033-295X.111.4.1061

Boyd, M. (2017). *Enterprise adoption of APIs is driven by internal integration needs*. Retrieved 31st May 2020 from https://thenewstack.io/integration-drives-api-uptake-enterprise/

Brehmer, Y., Westerberg, H., Bellander, M., Furth, D., Karlsson, S., & Backman, L. (2009). Working memory plasticity modulated by dopamine transporter genotype. *Neuroscience letters, 467*(2), 117-120. https://doi.org/10.1016/j.neulet.2009.10.018

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230-258. https://doi.org/10.1177/0049124192021002005

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology, 7*(1116). https://doi.org/10.3389/fpsyg.2016.01116

Burren, S. v., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*, 1-67. https://www.jstatsoft.org/v45/i03/

Burton, L. J., & Fogarty, G. J. (2003). The factor structure of visual imagery and spatial abilities. *Intelligence, 31*(3), 289-318. https://doi.org/10.1016/S0160-2896(02)00139-3

Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and success of psychology students in higher education. *Personality and Individual Differences, 29*, 1057-1068. https://doi.org/10.1016/S0191-8869(99)00253-6

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2016). Factor structure of the Wechsler Intelligence Scale for Children-Fifth Edition: Exploratory factor analyses with the 16 primary and secondary subtests *Psychological Assessment 28*(8), 975-986. https://doi.org/10.1037/pas0000238

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2017). Structural validity of the Wechsler Intelligence Scale for Children-Fifth Edition: Confirmatory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment, 29*(4), 458-472. https://doi.org/10.1037/pas0000358

Canivez, G. L., Watkins, M. W., & McGill, R. J. (2018). Construct validity of the Wechsler Intelligence Scale For Children – Fifth UK Edition: Exploratory and confirmatory factor analyses of the 16 primary and secondary subtests. *British Journal of Educational Psychology, 89*(2), 195-224. https://doi.org/10.1111/bjep.12230

Canivez, G. L., & Youngstrom, E. A. (2019). Challenges to the Cattell-Horn-Carroll theory: Empirical, clinical, and policy implications. *Applied Measurement in Education, 32*(3), 232-248. https://doi.org/10.1080/08957347.2019.1619562

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics, 36*(5), 648-662. https://doi.org/10.1016/j.clinthera.2014.04.006

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. Cambridge University Press.

Cattell, R. B. (1971). *Abilities: Their structure, growth and action*. Houghton Mifflin.

Cattell, R. B. (1982). *The inheritance of personality and ability: Research methods and findings*. Academic Press.

Cavanaugh, K. J. (2018). *Predicting score change: an empirical investigation of cheating on unproctored employment tests* [Dissertation, Old Dominion University].

Chad, H. V. I., Herman, A., Jeremy, D. M., & Philip, S. D. (2017). A Meta-Analysis of the Interactive, Additive, and Relative Effects of Cognitive Ability and Motivation on Performance. *Journal of Management, 44*(1), 249-279. https://doi.org/10.1177/0149206317702220

Chalmers, P. (2017). *CAT simulation with customized item selection*. Retrieved 1st June 2020 from https://philchalmers.github.io/mirtCAT/html/sim-unidimensional_custom.html

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generative adaptive and non-adaptive test interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software, 71*(5), 1-39. https://doi.org/10.18637/jss.v071.i05

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80*(1), 1-20. https://doi.org/10.1007/s11336-014-9401-5

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Chen, J.-Q., & Gardner, H. (2018). Assessment from the Perspective of Multiple-Intelligences Theory: Principles, Practices and Values. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary Intellectual Assessment* (4th ed., pp. 164-173). The Guilford Press.

Chen, S.-K., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement, 58*(4), 569-595. https://doi.org/10.1177/0013164498058004002

Chen, S.-K., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1997). The effect of population distribution and method of theta estimation on computerized adaptive testing (CAT) using the Rating Scale Model. *Educational and Psychological Measurement, 57*(3), 422-439. https://doi.org/10.1177/0013164497057003004

Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement, 33*(8), 644-645. https://doi.org/10.1177/0146621608329892

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2016). *lordif: Logistic Ordinal Regression Differential Item Functioning using IRT.* In (Version 0.3-3) https://CRAN.R-project.org/package=lordif

Christensen, H. (2014). Social media: The new e-mental health tool. *InPsych, 36*(3), 12-13. https://www.psychology.org.au/inpsych/2014/june/christensen

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q(3): Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178-194. https://doi.org/10.1177/0146621616677520

Cliniko. (2020). Retrieved 20th September 2020 from https://www.cliniko.com/

Cohen, J. W. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, R. J., & Swerdlik, M. E. (2005). *Psychological Testing and Assessment* (7th ed.). McGraw-Hill.

Cokely, E. T., Galesic, M., Shchulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgement and Decision Making, 7*(1), 25-47. https://doi.org/10.1037/e683152011-109

Cole, J. C. (2008). How to deal with missing data: Conceptual overview and details for implementing two modern methods. In J. W. Osborne (Ed.), *Best Practices in Quantitative Methods*. Sage Publications.

Colwell, N. M. (2013). Test anxiety, computer-adaptive testing, and the common core. *Journal of Education and Training Studies, 1*, 50-60. https://doi.org/10.11114/jest.vli2.101

Condon, D. M. (2015). An organizational framework for the psychological individual differences: Integrating the affective, cognitive, and conative domains. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 76*(5-B(E)).

Cornish, D., & Dukette, D. (2009). *The essential 20: Twenty components of an excellent health care team*. RoseDog Books.

Coutinho, V., Câmara-Costa, H., Kemlin, I., Billette de Villemeur, T., Rodriguez, D., & Dellatolas, G. (2017). The Discrepancy between Performance-Based Measures and Questionnaires when Assessing Clinical Outcomes and Quality of Life in Pediatric Patients with Neurological Disorders. *Applied Neuropsychology: Child, 6*(4), 255-261. https://doi.org/10.1080/21622965.2016.1146141

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. https://doi.org/10.1037/h0040957

Culbertson, M. (2011). *oscats Commits*. Retrieved 31st May 2020 from https://code.google.com/archive/p/oscats/source/default/commits

Daniel, M. H. (2012a). Equivalence of Q-interactive administered cognitive tasks: CVLT-II and selected D-KEFS subtests. In *Q-interactive Technical Report 3*. Pearson.

Daniel, M. H. (2012b). Equivalence of Q-interactive administered cognitive tasks: WAIS-IV. In *Q-interactive Technical Report 1*. Pearson.

Daniel, M. H. (2012c). Equivalence of Q-interactive administered cognitive tasks: WISC-IV. In *Q-interactive Technical Report 2*. Pearson.

Daniel, M. H. (2013a). Equivalence of Q-interactive and paper administrations of cognitive tasks: Selected NEPSY-II and CMS subtests. In *Q-interactive Technical Report 4*. Pearson.

Daniel, M. H. (2013b). Equivalence of Q-interactive and paper scoring of acdaemic tasks: Selected WIAT-III subtests. In *Q-interactive Technical Report 5*. Pearson.

Daniel, M. H., Wahlstrom, D., & Zhang, O. (2014). *Equivalence of Q-interactive and Paper Administrations of Cognitive Tasks: WISC-V*. Pearson. Retrieved 27th September 2020 from https://www.pearsonclinical.co.uk/Sitedownloads/q-interactive/research-documents/wisc-v-equivalency-report.pdf

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 600 million words, 1990-present.* Retrieved 3rd April 2020 from https://www.english-corpora.org/coca/

Davies, M. (2018). *The 14 Billion Word iWeb Corpus*. Retrieved 3rd April 2020 from https://www.english-corpora.org/iWeb/

Dawadi, P. N., Cook, D. J., & Schmitter-Edgecombe, M. (2013). Automated cognitive health assessment using smart home monitoring of complex tasks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 43*(6), 1302-1313. https://doi.org/10.1109/TSMC.2013.2252338

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

de Beer, M. (2005). Development of the Learning Potential Computerised Adaptive Test (LPCAT). *South African Journal of Psychology, 35*(4), 717-747. https://doi.org/10.1177/008124630503500407

De Luca, C. R., Wood, S. J., Anderson, V., Buchanan, J.-A., Proffitt, T. M., Mahony, K., & Pantelis, C. (2003). Normative data from the Cantab. I: Development of executive function over the lifespan. *Journal of Clinical and Experimental Neuropsychology, 25*(2), 242-254. https://doi.org/10.1076/jcen.25.2.242.13639

De Marco, M., Beltrachini, L., Biancardi, A., Frangi, A., & Venneri, A. (2017). Machine-learning support to individual diagnosis of mild cognitive impairment using multimodal MRI and cognitive assessments. *lzheimer Disease & Associated Disorders, 31*, 278-286. https://doi.org/10.1097/WAD.0000000000000208

Delgado-Gomez, D., Baca-Garcia, E., Aguado, D., Courtet, P., & Lopez-Castroman, J. (2016). Computerized Adaptive Test vs. decision trees: Development of a support decision system to identify suicidal behavior. *Journal of Affective Disorders, 206*, 204-209. https://doi.org/10.1016/j.jad.2016.07.032

Department of Employment, S., Small and Family Business. (2019). *Psychologists and Psychotherapists*. Australian Government. Retrieved 4th July from https://joboutlook.gov.au/Occupation?search=quiz&code=2723

DeVellis, R. F. (2006). Classical test theory. *Medical Care, 44*(11), S50-S59. http://www.jstor.org.ezproxy.lib.monash.edu.au/stable/41219505

Devine, J., Fliege, H., Kocalevent, R., Mierke, A., Klapp, B. F., & Rose, M. (2016). Evaluation of computerized adaptive tests (CATs) for longitudinal monitoring of depression, anxiety, and stress reactions. *Journal of Affective Disorders, 190*, 846-853. https://doi.org/10.1016/j.jad.2014.10.063

Dombrowski, S. C., Beaujean, A. A., McGill, R. J., Benson, N. F., & Schneider, W. J. (2019). Using exploratory bifactor analysis to understand the latent structure of multidimensional psychological measures: An example featuring the WISC-V. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(6), 847-860. https://doi.org/10.1080/10705511.2019.1622421

Dombrowski, S. C., Canivez, G. L., & Watkins, M. W. (2018). Factor structure of the 10 WISC-V primary subtests across four standardization age groups. *Contemporary School Psychology, 22*(1), 90-104. https://doi.org/10.1007/s40688-017-0125-2

Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2018). An alternative conceptualization of the theoretical structure of the Woodcock-Johnson IV Tests of Cognitive Abilities at school age: A confirmatory factor analytic investigation. *Archives of Scientific Psychology, 6*(1), 1-13. https://doi.org/10.1037/arc0000039

Dombrowski, S. C., McGill, R. J., & Morgan, G. B. (2019). Monte Carlo Modeling of Contemporary Intelligence Test (IQ) Factor Structure: Implications for IQ Assessment, Interpretation, and Theory. *Assessment*, 1073191119869828. https://doi.org/10.1177/1073191119869828

Dombrowski, S. C., & Watkins, M. (2013). Exploratory and higher order factor analysis of the WJ-III full test battery: A school-aged analysis. *Psychological Assessment, 25*, 442-455. https://doi.org/10.1037/a0031335

Downing, S. M. (2006a). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Lawrence Erlbaum.

Downing, S. M. (2006b). Twelve Steps for Effective Test Development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development*. Lawrence Erlbaum Associates.

Downing, S. M., & Haladyna, T. M. (1997). Test Item Development: Validity Evidence From Quality Assurance Procedures. *Applied Measurement in Education, 10*(1), 61-82. https://doi.org/10.1207/s15324818ame1001_4

Drasgow, F., & Olson-Buchanan, J. B. (1999). *Innovations in computerized assessment*. Lawrence Erlbaum Associates Publishers.

Duda, I., & Walter, T. (2012). A software framework for e-testing. IADIS International Conference e-Learning 2012, Lisbon, Spain.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155-174. https://doi.org/10.1037/1082-989X.5.2.155

Elliott, C. (2006). *Differential Ability Scales, Second Edition (DAS-II)*. Harcourt Assessment.

Embretson, S. E. (1995). The role of working memory capacity and general control processes in intelligence. *Intelligence, 20*(2), 169-189. https://doi.org/10.1016/0160-2896(95)90031-4

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349. https://doi.org/10.1037/1040-3590.8.4.341

Enders, C. K. (2010). *Applied Missing Data Analysis*. The Guilford Press.

Engle, R. W., & Kane, M. J. (2004). Executive Attention, Working Memory Capacity, and a Two-Factor Theory of Cognitive Control. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 145-199). Elsevier Science.

Evans, J. J., Floyd, R. G., McGrew, K. S., & Leforgee, M. H. (2002). The relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and reading achievement during childhood and adolescence. *School Psychology Review, 31*(2), 246-262. https://doi.org/10.1080/02796015.2002.12086154

Evans, T. R., Hughes, D. J., & Steptoe-Warren, G. (2019). A conceptual replication of emotional intelligence as a second-stratum factor of intelligence. *Emotion*. https://doi.org/10.1037/emo0000569

Farsides, T., & Woodfield, R. (2003). Individual differences and undergraduate academic success: the roles of personality, intelligence, and application. *Personality and Individual Differences, 34*, 1225-1243. https://doi.org/10.1016/S0191-8869(02)00111-3

Fenollar-Cortés, J., & Watkins, M. (2018). Construct validity of the Spanish Version of the Wechsler Intelligence Scale for Children Fifth Edition (WISC-V Spain). *7*, 150-164. https://doi.org/10.1080/21683603.2017.1414006

Fetzer, M., & Kantrowitz, T. (2011). Implementing Computer Adaptive Tests: Successes and Lessons Learned. In *Technology-Enhanced Assessment of Talent* (pp. 380-393). https://doi.org/10.1002/9781118256022.ch15

Finch, H., & French, B. F. (2019). A comparison of estimation techniques for IRT models with small samples. *Applied Measurement in Education, 32*(2), 77-96. https://doi.org/10.1080/08957347.2019.1577243

Flak, M. M., Hol, H. R., Hernes, S. S., Chang, L., Engvig, A., Bjuland, K. J., Pripp, A., Madsen, B. O., Knapskog, A. B., Ulstein, I., Lona, T., Skranes, J., & Løhaugen, G. C. (2019). Adaptive computerized working memory training in patients with mild cognitive impairment. A randomized double-blind active controlled trial. *Frontiers in Psychology, 10*(Apr). https://doi.org/10.3389/fpsyg.2019.00807

Flanagan, D. P., & Alfonso, V. C. (2017). *Essentials of WISC-V Assessment*. Wiley.

Flanagan, D. P., Alfonso, V. C., Mascolo, J. T., & Sotelo-Dynega, M. (2012). Use of ability tests in the identification of Specific Learning Disabilities within the context of an operational definition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment* (3rd ed., pp. 463-669). The Guilford Press.

Flanagan, D. P., & McDonough, E. M. (Eds.). (2018). *Contemporary Intellectual Assessment*. Guilford Publications.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of Cross-Battery Assessment* (3rd ed.). John Wiley & Sons.

Flanagan, D. P., & Schneider, W. J. (2016). Cross-Battery Assessment? XBA PSW? A case of mistaken identity: A commentary on Kranzler and colleagues' “Classification agreement analysis of Cross-Battery Assessment in the identification of specific learning disorders in children and youth”. *International Journal of School & Educational Psychology, 4*(3), 137-145. https://doi.org/10.1080/21683603.2016.1192852

Fleming, S. (2018). *Cognitive assessment and fluid reasoning: moving to the future with item response theory and computerized adaptive testing* Monash University]. Australia.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research, 14*(10), 2277-2291. https://doi.org/10.1007/s11136-005-6651-9

Floyd, R. G., Evans, J. J., & McGrew, K. S. (2003). Relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and mathematics achievement across the school-age years. *Psychology in the Schools, 40*(2), 155-171. https://doi.org/10.1002/pits.10083

Floyd, R. G., Keith, T. Z., Taub, G. E., & McGrew, K. S. (2007). Cattell-Horn-Carroll cognitive abilities and their effects on reading decoding skills: g has indirect effects, more specific abilities have direct effect. *School Psychology Quarterly, 22*(2), 200-233. https://doi.org/10.1037/1045-3830.22.2.200

Floyd, R. G., McGrew, K. S., & Evans, J. J. (2008). The relative contributions of the Cattell-Horn-Carroll cognitive abilities in explaining writing achievement during childhood and adolescence. *Psychology in the Schools, 45*(2), 132-144. https://doi.org/10.1002/pits.20284

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression* (2nd ed.). Sage.

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effects of adaptive testing on test taking motivation. *Diagnostica, 55*, 20-28. https://doi.org/10.1026/0012-1924.55.1.20

Fried, E. I. (2017). What are psychological constructs? On the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health Psychology Review, 11*(2), 130-134. https://doi.org/10.1080/17437199.2017.1306718

Gamble, N., & Morris, Z. (2014). Ethical and competent practice in the online age. *InPsych, 36*(3), 18-19. https://www.psychology.org.au/inpsych/2014/june/gramble

Gardner, W., Shear, K., Kelleher, K. J., Pajer, K. A., Mammen, O., Buysse, D., & Frank, E. (2004). Computerized adaptive measurement of depression: a simulation study. *BMC psychiatry, 4*(1), 13. https://doi.org/10.1186/1471-244x-4-13

Georgiadou, E., Triantafilou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment, 5*(8). www.jtla.org

Gershon, R. C., & Cook, K. (2011). Use of computer adaptive testing in the development of machine learning algorithms. *Pain Medicine, 12*(10), 1450-1452. https://doi.org/10.1111/j.1526-4637.2011.01235.x

Gibbons, C. J. (2017). Turning the page on pen-and-paper questionnaires: Combining ecological momentary assessment and computer adaptive testing to transform psychological assessment in the 21st century. *Frontiers in Psychology, 7*(1933), 1-4. https://doi.org/10.3389/fpsyg.2016.01933

Gibbons, C. J., Bower, P., Lovell, K., Valderas, J., & Skevington, S. (2016). Electronic quality of life assessment using computer-adaptive testing. *Journal of Medical Internet Research, 18*(9), e240. https://doi.org/10.2196/jmir.6053

Giesinger, J. M., Kuster, M. S., Holzner, B., & Giesinger, K. (2013). Development of a computer-adaptive version of the forgotten joint score. *The Journal of Arthroplasty, 28*(3), 418-422. https://doi.org/10.1016/j.arth.2012.08.026

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78. https://doi.org/https://doi.org/10.1016/j.paid.2016.06.069

Gomes, H., Molholm, S., Christodoulou, C., Ritter, W., & Cowan, N. (2000). The development of auditory attention in children. *Frontiers in Bioscience, 5*(1), 108-120. https://doi.org/10.2741/gomes

Gómez-Pérez, E., & Ostrosky-Solís, F. (2006). Attention and memory evaluation across the life span: heterogeneous effects of age and education. *Journal of Clinical* *and Experimental Neuropsychology, 28*(4), 477-494. https://doi.org/10.1080/13803390590949296

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies?: A comparative analysis of six preconcenptions about internet questionnaires. *American Psychologist, 59*(2), 93-104. https://doi.org/10.1037/0003-066X.59.2.93

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *BMJ: British Medical Journal, 331*(7524), 1064-1065. https://doi.org/10.1136/bmj.38636.593461.68

Haley, S. M., Ni, P., Hambleton, R. K., Slavin, M. D., & Jette, A. M. (2006). Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. *Journal of Clinical Epidemiology, 59*(11), 1174-1182. https://doi.org/10.1016/j.jclinepi.2006.02.010

Hamid, O. H., Smith, N. L., & Barzanji, A. (2017). Automation, per se, is not job elimination: How artificial intelligence forwards cooperative human-machine coexistence. 2017 IEEE 15th International Conference on Industrial Informatics,

Hamilton, R. J., & Bowers, B. J. (2006). Internet recruitment and E-Mail interviews in qualitative studies. *Qualitative Health Research, 16*(6), 821-835. https://doi.org/10.1177/1049732306287599

Hart, D. L., Cook, K. F., Mioduski, J. E., Teal, C. R., & Crane, P. K. (2006). Simulated computerized adaptive test for patients with shoulder impairments was efficient and produced valid measures of function. *Journal of Clinical Epidemiology, 59*(3), 290-298. https://doi.org/10.1016/j.jclinepi.2005.08.006

Hart, D. L., Mioduski, J. E., Werneke, M. W., & Stratford, P. W. (2006). Simulated computerized adaptive test for patients with lumbar spine impairments was efficient and produced valid measures of function. *Journal of Clinical Epidemiology, 59*(9), 947-956. https://doi.org/10.1016/j.jclinepi.2005.10.017

Hartmann, P. (2006). Spearman's law of diminishing returns: A look at age differentiation. *Journal of Individual Differences, 27*(4), 199-207. https://doi.org/10.1027/1614-0001.27.4.199

Hausler, J. (2006). Adaptive success control in computerized adaptive testing. *Psychology Science, 48*(4), 436-450. https://psycnet.apa.org/record/2007-03313-004

Hausler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science, 50*(1), 75-87.

Hawn, C. (2009). Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health Affairs, 28*(2), 361-368. https://doi.org/10.1377/hlthaff.28.2.361

He, Y., Zaslavsky, A. M., Landrum, M. B., Harrington, D. P., & Catalano, P. (2010). Multiple imputation in a large-scale complex survey: a practical guide. *Statistical methods in medical research, 19*(6), 653-670. https://doi.org/10.1177/0962280208101273

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology, 13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

HealthKit. (2018). Retrieved 6th February from https://www.healthkit.com/

Heng, A. (2018). *Towards a computerised adaptive test of visual processing: a pilot item development study using Rasch analysis* Monash University]. Australia.

Heywood, H. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 134*(824), 486-501. https://doi.org/10.1098/rspa.1931.0209

Hides, L. (2014). Are SMARTapps the future of youth mental health? *InPsych, 36*(3), 16-17. https://www.psychology.org.au/inpsych/2014/june/hides

Hines, S. (2018). *The development and validation of an automatic-item generation measure of cognitive ability* [Dissertation, Louisiana Tech University]. USA. https://digitalcommons.latech.edu/dissertations/71/

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (Vol. 663). Houghton Mifflin.

Hinzman, M., & Reed, D. (2018). *Teaching Sight Words as a Part of Comprehensive Reading Instruction* Iowa Reading Research Center. Retrieved 31st December 2018 from https://iowareadingresearch.org/blog/teaching-sight-words

Hlas, A. C., Neyers, K., & Molitor, S. (2017). Measuring student attention in the second language classroom. *Language Teaching Research, 23*(1), 107-125. https://doi.org/10.1177/1362168817713766

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Routledge.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Articles*, 2.

Horn, J. L. (1986). Some thoughts about intelligence. In R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence? Contemporary viewpoints on its nature and definition* (pp. 91-96). Ablex.

Horn, J. L. (1988). Thinking about human abilities. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 645-685). Plenum Press. https://doi.org/10.1007/978-1-4613-0893-5_19

Horn, J. L., & Blankson, A. N. (2012). Foundations for better understanding of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (3rd ed., pp. 73-98). Guilford.

Howe, W., & Dailey, D. (2015). *Woodcock Johnson IV Australasian Adaptation - Validity Study*. The Riverside Publishing Company.

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Huang, Y.-M., Lin, Y.-T., & Cheng, S.-C. (2009). An adaptive testing system for supporting versatile educational assessment. *Computers & Education, 52*(1), 53-67. https://doi.org/10.1016/j.compedu.2008.06.007

Huebner, A. R., Wang, C., Quinlan, K., & Seubert, L. (2016). Item exposure control for multidimensional computer adaptive testing under maximum likelihood and expected a posteriori estimation. *Behavior Research Methods, 48*(4), 1443-1453. https://doi.org/10.3758/s13428-015-0659-z

huhs, & yoonani. (2014). *IRT-Computerized Adaptive Testing*. Retrieved 31st May 2020 from https://sourceforge.net/projects/irt-cat/

Im, E.-O., & Chee, W. (2004). Recruitment of research participants through the Internet. *Computers, Informatics, Nursing, 22*(5), 289-297. https://doi.org/10.1097/00024665-200409000-00009

Innes, M. (2017). Projecting the future impact of advanced technologies: Will a robot take my job? *InPsych, 39*(2), 34-35. https://www.psychology.org.au/inpsych/2017/april/innes

Institute of Objective Measurement. (n.d.). *Rasch Dichotomous Model vs. One-parameter Logistic Model*. Retrieved 22nd June 2020 from https://www.rasch.org/rmt/rmt193h.htm

Iramaneerat, C., Smith, E. V., & Smith, R. M. (2008). An introduction to rasch measurement. In J. W. Osborne (Ed.), *Best Practices in Quantitative Measurement* (pp. 50-70). Sage Publications.

Jacobs, K. (2015). *Advances in Cognitive Assessment: Cattell-Horn-Carroll Theory and Cross-Battery Assessment* APS Educational and Developmental Psychology Conference, University of Melbourne, Parkvill.

Jacobs, K., & Costello, S. (2013). An initial investigation of an Australian Adaption of the Multidimensional Aptitude Battery - II. *The Australian Educational and Developmental Psychologist, 30*(1), 84-102. https://doi.org/10.1017/edp.2013.9

Jacobs, K., Watt, D., & Roodenburg, J. (2013). Why can't Jonny read? Bringing theory into cognitive assessment. *InPsych, 35*(December), 16-17. https://www.psychology.org.au/inpsych/2013/december/jacobs

James, L., Jacobs, K., & Roodenburg, J. (2015). Adoption of the Cattell-Horn-Carroll Model of cognitive abilities by Australian psychologists. *Australian Psychologist, 50*, 194-202. https://doi.org/10.111/ap.12110

Jha, S., & Topol, E. J. (2016). Adapting to artificial intelligence: radiologists and pathologists as information specialists. *Journal of American Medical Association, November*, 1-2. https://doi.org/10.1001/jama.2016.17438

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2019). *semTools: Useful tools for structural equation modeling.* In (Version 0.5-2) https://CRAN.R-project.org/package=semTools

Joshua, N., Wilson, C., & Bendrups, N. (2016). *Administration and Scoring Manual*. PsychCorp.

Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2012). A History of Intelligence Test Interpretation. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment* (pp. 56-71). The Guilford Press.

Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2018). A History of Intelligence Test Interpretation. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary Intellectual Assessment* (pp. 56-72). The Guilford Press.

Kantrowitz, T. M., & Dainis, A. M. (2014). How secure are unproctored pre-employment tests? Analysis of inconsistent test scores. *Journal of Business and Psychology, 29*(4), 605-616. https://doi.org/10.1007/s10869-014-9365-6

Kantrowitz, T. M., Dawson, C. R., & Fetzer, M. S. (2011). Computer adaptive testing (CAT): a faster, smarter, and more secure approach to pre-employment testing. *Journal of Business and Psychology, 26*(2), 227. https://doi.org/10.1007/s10869-011-9228-3

Kathleen, E. M., Wetzel, C. D., James, R. M., & David, J. W. (1984). Relationship between corresponding armed services vocational aptitude battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 8*(2), 155-163. https://doi.org/10.1177/014662168400800203

Kaufman, A. S., & Kaufman, N. L. (2004). *Manual for the Kaufman Assessment Battery for Children - Second Edition (KABC-II)*. American Guidance Service.

Kavanagh, D. (2014). Six reasons to integrate e-health into psychology practice. *InPsych, 36*(3), 14-15. https://www.psychology.org.au/inpsych/2014/june/kavanagh

Keith, T. Z., & Reynolds, M. R. (2012). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, tests and issues* (3rd ed., pp. 758-799). Guilford Press.

Kennedy, E. (2018). *The development of computer-administered items to improve the measurement of working memory* Monash University]. Australia.

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217. https://doi.org/10.1207/s15327957pspr0203_4

Kim, S.-H., Kwak, M., Bian, M., Feldberg, Z., Henry, T., Lee, J., Ölmez, İ. B., Shen, Y., Tan, Y., Tanaka, V., Wang, J., Xu, J., & Cohen, A. S. (2020). Item response models in psychometrika and psychometric textbooks. *Frontiers in Education, 5*(63). https://doi.org/10.3389/feduc.2020.00063

Kimura, T. (2017). The impacts of computer adaptive testing from a variety of perspectives. *Journal of educational evaluation for health professions, 14*, 12-12. https://doi.org/10.3352/jeehp.2017.14.12

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375. https://doi.org/10.1207/s15324818ame0204_6

Klenberg, L., Korkman, M., & Lahti-Nuuttila, P. (2001). Differential development of attention and executive functions in 3- to 12-year-old Finnish children. *Developmental Neuropsychology, 20*(1), 407-428. https://doi.org/10.1207/S15326942DN2001_6

Kong, X., Davis, L. L., McBride, Y., & Morrison, K. (2018). Response time differences between computers and tablets. *Applied Measurement in Education, 31*(1), 17-29. https://doi.org/10.1080/08957347.2017.1391261

Konsztowicz, S., Xie, H., Higgins, J., Mayo, N., & Koski, L. (2011). Development of a method for quantifying cognitive ability in the elderly through adaptive test administration. *International Psychogeriatrics, 23*(7), 1116-1123. https://doi.org/10.1017/S1041610211000615

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 5802-5805. https://doi.org/10.1073/pnas.1218772110

Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software, 74*(7), 1-16. https://doi.org/10.18637/jss.v074.i07

Krabbe, P. F. M. (2017). *The Measurement of Health and Health Status*. Academic Press. https://doi.org/10.1016/C2013-0-19200-8

Kranzler, J. H., Floyd, R. G., Benson, N., Zaboski, B., & Thibodaux, L. (2016). Cross-Battery Assessment pattern of strengths and weaknesses approach to the identification of specific learning disorders: Evidence-based practice or pseudoscience? *International Journal of School & Educational Psychology, 4*(3), 146-157. https://doi.org/10.1080/21683603.2016.1192855

Kraska, J. (2013). *Childhood personality as a moderator between cognitive ability and academic achievement* [Unpublished Masters Thesis]. Monash University.

Kroese, D. P., Brereton, T. J., Taimre, T., & Botev, Z. I. (2014). Why the Monte Carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics, 6*, 386-392. https://doi.org/10.1002/wics.1314

Kyrios, M., & Thomas, N. (2014). Psychology and the Internet: Where are we and where to from here? . *InPsych, 36*(3), 8-11. https://www.psychology.org.au/inpsych/2014/june/kyrios

Laidra, K., Pullmann, H., & Allik, J. (2006). Personality and intelligence as predictors of academic achievement: A cross-sectional study from elementary to secondary school. *Personality and Individual Differences, In Press*. https://doi.org/10.1016/j.paid.2006.08.001

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior, 12*, 119-131. https://doi.org/10.1016/s0022-5371(73)80001-5

Lang, J. W. B., & Kell, H. J. (2019). General mental ability and specific abilities: Their relative importance for extrinsic career success. *Journal of Applied Psychology, Advance online publication*. https://doi.org/10.1037/apl0000472

Langsrud, Ø. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing, 13*(2), 163-167. https://doi.org/10.1023/A:1023260610025

Lebedeva, E., Gallant, S., Tsai, C.-E., & Koski, L. (2015). Improving the measurement of cognitive ability in geriatric patients. *Dementia and Geriatric Cognitive Disorders, 40*, 148-157. https://doi.org/10.1159/000381536

Lecerf, T., & Canivez, G. L. (2018). "Complementary exploratory and confirmatory factor analyses of the French WISC-V: Analyses based on the standardization sample": Correction to Lecerf and Canivez (2018). *Psychological Assessment, 30*(8), 1009. https://doi.org/10.1037/pas0000638

Legree, P. J., Fischl, M. A., Gade, P. A., & Wilson, M. (1998). Testing word knowledge by telephone to estimate general cognitive aptitude using an adaptive test. *Intelligence, 26*(2), 91-98. https://doi.org/10.1016/S0160-2896%2899%2980056-7

Lemann, E. R., Davis, A. S., Finch, W. H., & Pierson, E. E. (2019). Evaluating the relation between CHC cognitive factors and selected components of executive functioning. *Journal of Pediatric Neuropsychology, 5*(4), 152-162. https://doi.org/10.1007/s40817-019-00073-3

Lewis, T. F. (2017). Evidence regarding the internal structure: confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development, 50*(4), 239-247. https://doi.org/10.1080/07481756.2017.1336929

Liao, W.-W., & Ho, R.-G. (2011). Designing a Virtual Item Bank Based on the Techniques of Image Processing. *Turkish Online Journal of Educational Technology, 10*(4), 93-106.

https://www.researchgate.net/publication/289593149_Designing_a_virtual_item_bank_based_on_the_techniques_of_image_processing

Ligtvoet, R., van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement, 70*(4), 578-595. https://doi.org/10.1177/0013164409355697

Lilienfeld, S. O., Sauvigné, K. C., Lynn, S. J., Cautin, R. L., Latzman, R. D., & Waldman, I. D. (2015). Fifty psychological and psychiatric terms to avoid: a list of inaccurate, misleading, misused, ambiguous, and logically confused words and phrases. *Frontiers in Psychology, 6*, 1100. https://doi.org/10.3389/fpsyg.2015.01100

Lillard, A. S., & Peterson, J. (2011). The immediate impact of different types of television on young children's executive function. *Pediatrics, 128*(4), 644. https://doi.org/10.1542/peds.2010-1919

Linacre, J. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*, 328.

Lis, P. (2018). *5.0.beta.7.2*. The University of Cambridge Psychometrics Centre. Retrieved 23rd June from https://github.com/campsych/concerto-platform/tree/5.0.beta.7

Lis, P. (2020). *Changelog*. The University of Cambridge Psychometrics Centre. Retrieved 23rd June from https://github.com/campsych/concerto-platform/blob/master/CHANGELOG.md

Little, R. J. A., & Rubin, D. B. (2020). *Statistical Analysis with Missing Data*. Wiley.

Liu, C., Han, K. T., & Li, J. (2019). Compromised item detection for computerized adaptive testing. *Frontiers in Psychology, 10*, 829-829. https://doi.org/10.3389/fpsyg.2019.00829

Loe, B. S., Stillwell, D., & Gibbons, C. (2017). Computerized Adaptive Testing provides reliable and efficient depression measurement using the CES-D. *Journal of Medical Internet Research, 19*(9), e302. https://doi.org/10.2196/jmir.7453

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of" scale analysis" and factor analysis. *Psychological Bulletin, 45*(6), 507-529. https://doi.org/10.1037/h0055827

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*(2), 157-162. https://doi.org/10.1111/j.1745-3984.1986.tb00241.x

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR serious games, 4*(2), e11-e11. https://doi.org/10.2196/games.5888

Lunz, M. E., Bergstrom, B. A., & Gershon, R. C. (1994). Computer adaptive testing. *International Journal of Educational Research, 21*(6), 623-634. https://doi.org/10.1016/0883-0355(94)90015-9

Lyonette, C., Atfield, G., Baldauf, B., & Owen, D. (2019). *Research on the Educational Psychologist Workforce*. Government Social Research.

Maas, H. L. J. v. d., Kan, K.-J., & Borsboom, D. (2014). Intelligence is what intelligence test measures. Seriously. *Journal of Intelligence, 2*, 12-15. https://doi.org/10.3390/jintelligence2010012

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149. https://doi.org/10.1037/1082-989x.1.2.130

Magis, D., Beland, S., Tuerlinckx, F., & Boeck, P. D. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847-862. https://doi.org/10.3758/BRM.42.3.847

Maguire, J. (2018). *Past and present measures of crystallised intelligence inform the development of an adaptive cognitive ability scale* Monash University]. Australia.

Mair, P. (2018). *Modern Psychometrics with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-93177-7

Makransky, G., & Glas, C. A. (2013). The applicability of multidimensional computerized adaptive testing for cognitive ability measurement in organizational assessment. *International Journal of Testing, 13*(2), 123-139. https://doi.org/10.1080/15305058.2012.672352

Martin, A. J., & Lazendic, G. (2018). Computer-Adaptive Testing: Implications for Students' Achievement, Motivation, Engagement, and Subject Test Experience. *Journal of Educational Psychology, 110*(1), 27-48. https://doi.org/10.1037/edu0000205

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. https://doi.org/10.1007/BF02296272

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713. https://doi.org/10.1007/s11336-005-1295-9

McCann Associates. (2019). Retrieved 31st May 2020 from http://www.mccanntesting.com/

McGill, R. J. (2017). Re(Examining) relations between CHC broad and narrow cognitive abilities and reading achievement. *Journal of Educational and Developmental Psychology, 7*(1), 265-282. https://doi.org/10.5539/jedp.v7n1p265

McGrew, K. S. (2009a). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1-10. https://doi.org/10.1016/j.intell.2008.08.004

McGrew, K. S. (2009b). *Evolution of CHC Theory of Intelligence and Assessment.* Retrieved 23rd April 2018 from http://www.iapsych.com/chctimeline/chctimeline2.pdf

McGrew, K. S. (2016). *The CHC model of human cognitive abilities--a proposed revision (v2.3): Has Glr been incorrectly conceptualized since 1997?* . IQs Corner. Retrieved 23rd April from http://www.iqscorner.com/2016/06/the-chc-model-of-human-cognitive.html

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). Technical Manual. In *Woodcock-Johnson IV*. Riverside.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H., & Yarkoni, T. (2016). How open science helps researchers succeed. *eLife, 5*, e16800. https://doi.org/10.7554/eLife.16800

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Jorunal of Applied Psychology, 97*(5), 1016-1031. https://doi.org/10.1037/a0027934

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11. https://doi.org/10.3102/0013189x018002005

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066x.50.9.741

Miciak, J., Taylor, W. P., Stuebing, K. K., & Fletcher, J. M. (2016). Simulation of LD identification accuracy using a pattern of processing strengths and weaknesses method with multiple measures. *Journal of Psychoeducational Assessment, 36*(1), 21-33. https://doi.org/10.1177/0734282916683287

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. De Gruyter, The Hague.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve methoden, 12*(37), 97-117.

Muller, K., & Wickham, H. (2019). *tibble: Simple Data Frames.* In (Version 2.1.3) https://CRAN.R-project.org/package=tibble

Müller, K., & Wickham, H. (2019). *tibble: Simple Data Frames.* In (Version 2.1.3) https://CRAN.R-project.org/package=tibble

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176. https://doi.org/10.1177/014662169201600206

Naglieri, J. A., & Otero, T. M. (2018). Redifining intelligence with the Planning, Attention, Simultaneous, and Successive Theory of Neurocognitive Processes. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary Intellectual Assessment* (3rd ed., pp. 195-218). The Guilford Press.

Nesnidol, S., & Highhouse, S. (2018). Why does the public sector resist unproctored internet testing? *Personnel Assessment and Decisions, 4*(2). https://doi.org/10.25035/pad.2018.02.002

Neural Assembly. (2019). *Research Based*. Retrieved 9th July 2020 from https://www.cogmed.com/working-memory/research

Newton, D. (2019). *Today's Tests Are Failing Students*. Retrieved 8th June 2020 from https://www.forbes.com/sites/dereknewton/2019/05/22/todays-tests-are-failing-students/

Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *The patient, 7*(1), 23-35. https://doi.org/10.1007/s40271-013-0041-0

Nikolaou, I., Georgiou, K., & Kotsasarlidou, V. (2019). Exploring the relationship of a gamified assessment with performance. *The Spanish Journal of Psychology, 22*. https://doi.org/10.1017/sjp.2019.5

Nydick, S. J., & Weiss, D. J. (2010). *Accepting the null: Determining no change within the adaptive measurement of change* 2010 International Association for Computerized Adaptive Testing, Arnhem, The Netherlands.

Ofcom. (n.d.). *Distribution of internet users in Australia as of August 2015, by age group*. Retrieved 6th March from https://www-statista-com.ezproxy.lib.monash.edu.au/statistics/259828/age-distribution-of-internet-users-in-australia/

Oppl, S., Reisinger, F., Eckmaier, A., & Helm, C. (2017). A flexible online platform for computerized adaptive testing. *International Journal of Educational Technology in Higher Education, 14*(1), 2. https://doi.org/10.1186/s41239-017-0039-0

Pallant, J. (2011). *SPSS Survival Manual* (4th ed.). Allen & Unwin.

Pearson. (2013). *Q-interactive*. Pearson. Retrieved 23rd April from https://www.pearsonclinical.com.au/products/view/520

Pearson. (2020a). *Pearson VUE*. Retrieved 31st May 2020 from https://home.pearsonvue.com/#

Pearson. (2020b). *Telepractice Today*. Retrieved 7th June 2020 from https://www.pearsonassessments.com/professional-assessments/digital-solutions/telepractice/about.html

Pelánek, R., & Jarušek, P. (2015). Student modeling based on problem solving times. *International Journal of Artificial Intelligence in Education, 25*(4), 493-519. https://doi.org/10.1007/s40593-015-0048-x

Petersen, M. A., Gamper, E.-M., Costantini, A., Giesinger, J. M., Holzner, B., Johnson, C., Sztankay, M., Young, T., & Groenvold, M. (2016). An emotional functioning item bank of 24 items for computerized adaptive testing (CAT) was established. *Journal of Clinical Epidemiology, 70*, 90-100. https://doi.org/10.1016/j.jclinepi.2015.09.002

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of experimental psychology, 58*(3), 193. https://doi.org/10.1037/h0049234

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value in Health, 18*, 25-34. https://doi.org/10.1016/j.jval.2014.10.005

Pew Research Center. (2017). *Internet/Broadband Fact Sheet*. Author. Retrieved 30th January from http://www.pewinternet.org/fact-sheet/internet-broadband/

Press, G. (2013). *A very short history of data science*. Forbes. Retrieved 12th January from https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

Primi, R., Nakano, T. d. C., & Wechsler, S. M. (2018). Using four-parameter item response theory to model human figure drawings. *Avaliação Psicológica, 17*, 473-483. https://doi.org/10.15689/ap.2018.1704.7.07

Prometric. (2020). Retrieved 31st May 2020 from https://www.prometric.com/

R Core Team. (2018). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* In https://www.R-project.org/

R Core Team. (2019). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* In (Version 3.6.2) https://www.R-project.org/

R Core Team. (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* In (Version 3.6.3) https://www.R-project.org/

R Studio Team. (2015). *RStudio: Integrated Development for R.* In (Version 1.1.456) RStudio, Inc. http://www.rstudio.com/

R Studio Team. (2019). *RStudio: Integrated Development for R.* In (Version 1.2.5033) RStudio, Inc. http://www.rstudio.com/

Raiford, S. E., & Coalson, D. L. (2014). *Essentials of WPPSI-IV Assessment*. Wiley.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Rässler, S., & Riphahn, R. T. (2006). Survey Item Nonresponse and its Treatment. In O. Hübler & J. Frohn (Eds.), *Modern Econometric Analysis: Surveys on Recent Developments* (pp. 215-230). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-32693-6_15

Ravelle, W. (2018). *psych: Procedures for Personality and Psychological Research.* In (Version 1.8.12) https://CRAN.R-project.org/package=psych

Ravelle, W. (2019). *psych: Procedures for Personality and Psychological Research.* In (Version 1.9.12) https://CRAN.R-project.org/package=psych

Reckase, M. D., Ju, U., & Kim, S. (2019). How adaptive is an adaptive test: are all adaptive tests adaptive? *Journal of Computerized Adaptive Testing, 7*(1). https://doi.org/10.7333%2Fjcat.v7i1.69

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: not much more than g. *Journal of Applied Psychology, 79*(4), 518-524. https://doi.org/10.1037/0021-9010.79.4.518

Reichenberg, R. (2018). Dynamic Bayesian Networks in educational measurement: Reviewing and advancing the state of the field. *Applied Measurement in Education, 31*(4), 335-350. https://doi.org/10.1080/08957347.2018.1495217

Reise, S. P., Ventura, J., Keefe, R. S. E., Baade, L. E., Gold, J. M., Green, M. F., Kern, R. S., Mesholam-Gately, R., Nuechterlein, K. H., Seidman, L. J., & Bilder, R. (2011). Bifactor and item response theory analyses of interviewer report scales of cognitive impairment in schizophrenia. *Psychological Assessment, 23*(1), 245-261. https://doi.org/10.1037/a0021501

Revelle, W. (2019). *psych: Procedures for Personality and Psychological Research.* In (Version 1.9.12) https://CRAN.R-project.org/package=psych

Revicki, D. A., & Cella, D. F. (1997). Health status assessment for the twenty first century: item response theory, item banking and computer adaptive testing. *Quality of Life Research, 6*, 595-600. https://doi.org/10.1023/a:1018420418455

Revlin, R. (2012). *Cognition: Theory and Practice*. Worth Publishers.

Rhemtulla, M., & Hancock, G. R. (2016). Planned missing data designs in educational psychology research. *Educational Psychologist, 51*(3-4), 305-316. https://doi.org/10.1080/00461520.2016.1208094

Rhemtulla, M., & Little, T. (2012). Tools of the trade: planned missing data designs for research in cognitive development. *Journal of Cognition and Development 13*(4), 425-438. https://doi.org/10.1080/15248372.2012.717340

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 3*(4), 369-379. https://doi.org/10.1080/10705519609540052

Robbins, H., & Munro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics, 22*, 400-407. https://doi.org/10.1214/aoms/1177729586

Roberts, G., Quach, J., Spencer-Smith, M., Anderson, P. J., Gathercole, S., Gold, L., Sia, K.-L., Mensah, F., Rickards, F., Ainley, J., & Wake, M. (2016). Academic outcomes 2 years after working memory training for children with low working memory: a randomized clinical trial. *JAMA pediatrics, 170*(5), e154568. https://doi.org/10.1001/jamapediatrics.2015.4568

Roberts, R. D., Goff, G. N., Anjoul, F., Kyllonen, P. C., Pallier, G., & Stankov, L. (2000). The armed services vocational aptitude battery (ASVAB): Little more than acculturated learning (Gc)!? *Learning and Individual Differences, 12*(1), 81-103. https://doi.org/10.1016/S1041-6080(00)00035-2

Roid, G. H. (2005). *Stanford-Binet Intelligence Scales Fifth Edition (SB5)*. Pro-Ed.

Roid, G. H., & Pomplun, M. (2012). The Stanford-Binet Intelligence Scales, Fifth Edition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment* (pp. 249-268). The Guilford Press.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. http://www.jstatsoft.org/v48/i02/

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys* John Wiley.

Rudner, L. M. (1998). *An On-line, Interactive, Computer Adaptive Testing Tutorial*. Retrieved 1st June 2020 from http://edres.org/scripts/cat

Ruff, H. A., & Lawson, K. R. (1990). Development of sustained, focused attention in young children during free play. *Developmental Psychology, 26*(1), 85-93. https://doi.org/10.1037/0012-1649.26.1.85

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information & Management, 54*, 189-203. https://doi.org/10.1016/j.im.2016.06.005

Samejima, F. (1997). Graded Response Model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85-100). Springer New York. https://doi.org/10.1007/978-1-4757-2691-6_5

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation*. American Psychological Association.

Scalise, K. (2009). New electronic technologies for facilitating differentiated instruction. *Journal on School Educational Technology, 4*(4), 39-45. https://doi.org/10.26634/jsch.4.4.532

Scalise, K., & Allen, D. D. (2015). Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. *British Journal of Mathematical and Statistical Psychology, 68*(3), 478-496. https://doi.org/10.1111/bmsp.12057

Schenker, N., Raghunathan, T. E., Chiu, P.-L., Makuc, D. M., Zhang, G., & Cohen, A. J. (2006). Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association, 101*(475), 924-933. https://doi.org/10.1198/016214505000001375

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll Model of Intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment* (3rd ed., pp. 99-144). The Guilford Press.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary Intellectual Assessment: Theories, Tests and Issues* (4th ed.). Guilford Press.

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika, 66*(1), 79-97. https://doi.org/10.1007/BF02295734

Segall, D. O., & Moreno, K. E. (1999). Development of the Computerized Adaptive Testing Version of the aRmed Services Vocational Aptitude Battert. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in Computerized Assessment*. Lawrence Erlbaum Associates.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology, 179*(6), 764-774. https://doi.org/10.1093/aje/kwt312

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*(2), 149-157. https://doi.org/10.1177/014662169201600204

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Jorunal of Educational Measurement, 28*(3), 237-247. https://doi.org/10.1111/j.1745-3984.1991.tb00356.x

Smits, N., van der Ark, L. A., & Conijn, J. M. (2017). Measurement versus prediction in the construction of patient-reported outcome questionnaires: can we have our cake and eat it? *Quality of Life Research*. https://doi.org/10.1007/s11136-017-1720-4

Sommer, M., Arendasy, M. E., Schutzhofer, B., & Knessel, G. (2018). Comparing the effectiveness of different methods to reduce the effect size of the practice effect in traffic psychological assessment. *Transportation Research Part F: Traffic Psychology and Behaviour, 58*, 955-968. https://doi.org/10.1016/j.trf.2018.06.042

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. Macmillan.

Spearman, C. (1938). Propose explanation of individual differences of ability by "sampling". *British Journal of Psychology, 30*(1), 1-16. https://doi.org/10.1037/h0061267

Stahl, S. A., & Jacobson, M. G. (1986). Vocabulary difficulty, prior knowledge, and text comprehension. *Journal of Reading Behavior, XVIII*(4), 309-323. https://doi.org/10.1080/10862968609547578

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: a new generation of personality tests to support military selection and classification decisions. *Military Psychology, 26*(3), 153-164. https://doi.org/10.1037/mil0000044

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods, 15*(3), 463-487. https://doi.org/10.1177/1094428112444611

Sternberg, R. J. (2003). Implications of the theory of successful intelligence for career choice and development. *Journal of Career Assessment, 11*(2), 136-152. https://doi.org/10.1177/1069072703011002002

Sternberg, R. J. (2018). The Triarchic Theory of Successful Intelligence. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary Intellectual Assessment* (3rd ed., pp. 174-194). The Guilford Press.

Stone, A. A., & Shhiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine, 16*, 199-202. https://doi.org/10.1093/abm/16.3.199

Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A., & Clausen, J. A. (1950). *Measurement and prediction*. Princeton University Press.

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2014). Minimum sample size requirements for Mokken scale analysis. *Educational and Psychological Measurement, 74*(5), 809-822. https://doi.org/10.1177/0013164414529793

Stucky, B. D., Edelen, M. O., Sherbourne, C. D., Eberhart, N. K., & Lara, M. (2014). Developing an item bank and short forms that assess the impact of asthma on quality of life. *Respiratory Medicine, 108*, 252-263. https://doi.org/10.1016/j.rmed.2013.12.008

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Pearson Education.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education, 2*, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Thamsborg, L. H., Petersen, M. A., Aaronson, N. K., Chie, W.-C., Costantini, a., Holzner, B., Leeuw, I. M. V.-d., Young, T., & Groenvold, M. (2015). Development of a lack of appetite item bank for computer adaptive testing (CAT). *Support Care Cancer, 23*, 1541-1548. https://doi.org/10.1007/s00520-014-2498-3

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In *Differential item functioning.* (pp. 67-113). Lawrence Erlbaum Associates, Inc.

Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. *Psychological Assessment, 31*(12), 1442-1455. https://doi.org/10.1037/pas0000597

Thomas, M. L., Brown, G. G., Patt, V. M., & Duffy, J. R. (2020). Latent variable modeling and adaptive testing for experimental cognitive psychopathology research. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164420919898

Thompson, G. (2017). Computer adaptive testing, big data and algorithmic approaches to education. *British journal of sociology of education, 38*(6), 827-840. https://doi.org/10.1080/01425692.2016.1158640

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*(1). http://pareonline.net/getvn.asp?v=16&n=1

Thurstone, L. L. (1948). Psychological implications of factor analysis. *The American Psychologist, 3*(9), 402-408. https://doi.org/10.1037/h0058069

Tighe, P., Laduzenski, S., Edwards, D., Ellis, N., Boezaart, A. P., & Aygtug, H. (2011). Use of machine learning theory to predict the need for femoral nerve block following ACL repair. *Pain Medicine, 12*(10), 1566-1575. https://doi.org/10.1111/j.1526-4637.2011.01228.x

Tong, T., & Chignell, M. (2014). *Developing a serious game for cognitive assessment: choosing settings and measuring performance* Proceedings of the Second International Symposium of Chinese CHI, Toronto, Ontario, Canada. https://doi-org.ezproxy.lib.monash.edu.au/10.1145/2592235.2592246

Torres Irribarri, D., & Freund, R. (2014). *Wright Map: IRT item-person map with ConQuest integration.* In (Version 1.2.2) http://github.com/david-ti/wrightmap

Tu, D., Han, Y., Cai, Y., & Gao, X. (2018). Item selection methods in multidimensional computerized adaptive testing with polytomously scored items. *Applied Psychological Measurement, 42*(8), 677-694. https://doi.org/10.1177/0146621618762748

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10. https://doi.org/10.1007/bf02291170

Twinword. (2016). *How to check English word difficulty?* Retrieved 30th December 2018 from https://www.twinword.com/blog/how-to-check-english-word-difficulty/

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1-19. https://doi.org/10.18637/jss.v020.i11

Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*(5), 1-27. https://doi.org/10.18637/jss.v048.i05

W3C. (2014). *HTML5: A vocabulary and associated APIs for HTML and XHTML*. Retrieved 31st May 2020 from https://www.w3.org/TR/2018/SPSD-html5-20180327/

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of school outcomes based on early language production and socioeconomic factors. *Child development, 65*(2), 606-621. https://doi.org/10.2307/1131404

Wang, Y.-C., Hart, D. L., Cook, K. F., & Mioduski, J. E. (2010). Translating shoulder computerized adaptive testing generated outcome measures into clinical practice. *Journal of Hand Therapy, 23*(4), 372-383. https://doi.org/10.1016/j.jht.2010.06.001

Warne, R. T. (2015). Five reasons to put the g back into giftedness: An argument for applying the Cattell-Horn-Carroll Theory of Intelligence to Gifted Education Research and Practice. *Gifted Child Quarterly*. https://doi.org/10.1177/0016986215605360

Wasserman, J. D. (2012). The History of Intelligence Assessment. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment* (pp. 3-55). The Guilford Press.

Wasserman, J. D. (2018). A History of Intelligence Assessment: The Unfinished Tapestry. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary Intellectual Assessment* (pp. 3-55). The Guilford Press.

Wasserman, J. D. (2019). Deconstructing CHC. *Applied Measurement in Education, 32*(3), 249-268. https://doi.org/10.1080/08957347.2019.1619563

Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Williams & Wilkins.

Wechsler, D. (2002). *WPPSI-III: Technical and interpretative manual*. The Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children - Fourth Edition (WISC-IV)*. Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale Fourth Edition (WAIS-IV)*. Psychological Corporation.

Wechsler, D. (2014). *Wechsler Preschool and Primary Scale of Intelligence - Fourth Edition Australian and New Zealand Standardised Edition*. Pearson.

Wechsler, D. (2016). *Wechsler Intelligence Scale for Children, Fifth Edition: Australian and New Zealand Standardised Edition*. Pearson.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting & Clinical Psychology, 53*(6), 774-789. https://doi.org/10.1037/0022-006x.53.6.774

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 1-23. https://doi.org/10.2458/jmm.v2i1.12351

Wettstein, M. T., Benjamin; Kuzma, Elzbieta; Wahl, Hans-Werner (2017). The interplay between personality and cognitive ability across 12 years in middle and late adulthood: evidence for reciprocal associations *Psychology and Aging, 32*(3), 259-277. https://doi.org/10.1037/pag0000166

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software, 21*(12), 1-20. http://www.jstatsoft.org/v21/i12/

Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations.* In (Version 1.4.0) https://CRAN.R-project.org/package=stringr

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *dplyr: A Grammar of Data Manipulation.* In (Version 0.8.3) https://CRAN.R-project.org/package=dplyr

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A Grammar of Data Manipulation.* In (Version 0.8.4) https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2019). *tidyr: Tidy Messy Data.* In (Version 0.8.3) https://CRAN.R-project.org/package=tidyr

Wickham, H., & Henry, L. (2020). *tidyr: Tidy Messy Data.* In (Version 1.0.2) https://CRAN.R-project.org/package=tidyr

Wilson, K., & Korn, J. H. (2007). Attention during Lectures: Beyond Ten Minutes. *Teaching of Psychology, 34*(2), 85-89. https://doi.org/10.1080/00986280701291291

Witruk, E. (2019). Working memory in Cantonese and German speaking dyslexic children. *Health Psychology Report, 7*(4), 305-315. https://doi.org/10.5114/hpr.2019.88663

Wools, S., Molenaar, M., & Otter, D. H.-d. (2019). The validity of technology enhanced assessments - threats and opportunities. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*. Springer Open.

Wouters, H., de koning, I., Zwinderman, A. H., van Gool, W. A., Schmand, B., Buiter, M., & Lindeboom, R. (2009). Adaptive cognitive testing in cerebrovascular disease and vascular dementia. *Dementia and Geriatric Cognitive Disorders, 28*(5), 486-492. https://doi.org/10.1159/000250593

Wouters, H., Zwinderman, A. H., van Gool, W. A., Schmand, B., & Lindeboom, R. (2009). Adaptive cognitive testing in dementia. *International journal of methods in psychiatric research, 18*(2), 118-127. https://doi.org/10.1002/mpr.283

Xie, Y. (2015). *Dynamic Documents with R and knitr* (2nd ed.). Chapman and Hall. https://yihui.org/knitr/

Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai archives of psychiatry, 26*(3), 171-177. https://doi.org/10.3969/j.issn.1002-0829.2014.03.010

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213. http://www.jstor.org.ezproxy.lib.monash.edu.au/stable/1435043

Yoes, M. (1995). *An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model*. Assessment Systems Corporation.

Zammuto, M. (2018). *How artificial intelligence is changing the CIO's role*. CIO. Retrieved 17th March from https://www.cio.com/article/3261545/artificial-intelligence/how-artificial-intelligence-is-changing-the-cios-role.html

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337-362. https://doi.org/10.1207/S15324818AME1504_02

Žitný, P., Halama, P., Jelínek, M., & Květon, P. (2012). Validity of cognitive ability testscomparison of computerized adaptive testing with paper and pencil and computer-based forms of administrations. *Studia Psychologica, 54*(3), 181-194.

Appendices

Appendix A Lexical Knowledge ITOS CFA Parameters and Mokken Analysis

	55 Items		53 Items				51 Items				6 items			
Item	***B***	**Stan. *B***	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**
1	0.014	0.048	-	-	-	-	-	-	-	-	-	-	-	-
2	0.016***	0.125	0.016***	0.125	.214	0	-	-	-	-	-	-	-	-
3	0.011*	0.065	-	-	-	-	-	-	-	-	-	-	-	-
4	0.032***	0.258	0.032***	0.258	.439	1	0.032***	0.258	0.442	1	-	-	-	-
5	0.035***	0.267	0.035***	0.267	.425	1	0.035***	0.266	0.431	1	-	-	-	-
6	0.040***	0.283	0.040***	0.283	.433	1	0.040***	0.283	0.438	1	-	-	-	-
7	0.052***	0.356	0.052***	0.356	.520	1	0.052***	0.356	0.527	1	-	-	-	-
8	0.058***	0.420	0.058***	0.420	.638	1	0.058***	0.420	0.648	1	-	-	-	-
9	0.062***	0.498	0.062***	0.497	.817	1	0.062***	0.498	0.832	1	-	-	-	-
10	0.071***	0.525	0.071***	0.525	.801	1	0.071***	0.525	0.814	1	-	-	-	-
11	0.085***	0.515	0.085***	0.515	.672	1	0.085***	0.515	0.678	1	-	-	-	-
12	0.100***	0.572	0.100***	0.572	.702	1	0.100***	0.572	0.709	1	-	-	-	-
13	0.110***	0.595	0.110***	0.595	.695	1	0.110***	0.596	0.701	1	-	-	-	-
14	0.120***	0.649	0.120***	0.649	.755	1	0.120***	0.649	0.760	1	-	-	-	-
15	0.131***	0.668	0.131***	0.668	.732	1	0.131***	0.668	0.739	1	-	-	-	-
16	0.131***	0.600	0.131***	0.600	.625	1	0.131***	0.600	0.628	1	-	-	-	-
17	0.135***	0.677	0.135***	0.677	.722	1	0.136***	0.678	0.728	1	-	-	-	-
18	0.140***	0.700	0.140***	0.700	.746	1	0.140***	0.700	0.749	1	-	-	-	-
19	0.156***	0.755	0.156***	0.755	.780	1	0.156***	0.756	0.785	1	-	-	-	-
20	0.168***	0.823	0.168***	0.823	.849	1	0.168***	0.823	0.854	1	-	-	-	-
21	0.174***	0.792	0.174***	0.792	.779	1	0.174***	0.792	0.783	1	-	-	-	-
22	0.184***	0.821	0.184***	0.821	.792	1	0.184***	0.821	0.797	1	-	-	-	-
23	0.200***	0.786	0.200***	0.786	.720	1	0.200***	0.786	0.724	1	-	-	-	-
24	0.204***	0.864	0.204***	0.864	.809	1	0.204***	0.864	0.811	1	-	-	-	-
25	0.207***	0.875	0.207***	0.875	.818	1	0.207***	0.875	0.821	1	-	-	-	-
26	0.210***	0.888	0.210***	0.888	.831	1	0.210***	0.888	0.834	1	-	-	-	-
27	0.215***	0.876	0.215***	0.876	.808	1	0.215***	0.877	0.811	1	-	-	-	-
28	0.219***	0.899	0.219***	0.899	.833	1	0.219***	0.899	0.835	1	-	-	-	-
29	0.221***	0.859	0.221***	0.859	.783	1	0.221***	0.859	0.787	1	-	-	-	-
30	0.227***	0.928	0.227***	0.928	.863	1	0.227***	0.928	0.866	1	-	-	-	-
31	0.223***	0.733	0.223***	0.733	.693	1	0.223***	0.733	0.697	1	-	-	-	-

	55 Items		53 Items				51 Items				6 items			
Item	***B***	**Stan. *B***	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**
32	0.233***	0.891	0.233***	0.891	.820	1	0.233***	0.891	0.825	1	-	-	-	-
33	0.236***	0.854	0.236***	0.854	.790	1	0.236***	0.854	0.794	1	-	-	-	-
34	0.212***	0.530	0.212***	0.530	.612	1	0.211***	0.530	0.621	1	-	-	-	-
35	0.240***	0.858	0.240***	0.858	.797	1	0.240***	0.858	0.800	1	-	-	-	-
36	0.242***	0.897	0.242***	0.897	.832	1	0.242***	0.896	0.835	1	-	-	-	-
37	0.242***	0.912	0.242***	0.912	.847	1	0.242***	0.912	0.849	1	-	-	-	-
38	0.242***	0.820	0.242***	0.821	.776	1	0.242***	0.820	0.780	1	-	-	-	-
39	0.221***	0.568	0.221***	0.568	.664	1	0.221***	0.567	0.667	1	0.254***	0.652	0.520	1
40	0.242***	0.836	0.242***	0.836	.787	1	0.242***	0.836	0.791	1	-	-	-	-
41	0.241***	0.808	0.241***	0.808	.771	1	0.241***	0.808	0.776	1	-	-	-	-
42	0.205***	0.479	0.205***	0.479	.594	1	0.205***	0.479	0.612	1	-	-	-	-
43	0.245***	0.894	0.245***	0.894	.840	1	0.245***	0.894	0.841	1	-	-	-	-
44	0.242***	0.780	0.242***	0.780	.755	1	0.242***	0.780	0.758	1	-	-	-	-
45	0.241***	0.782	0.241***	0.783	.765	1	0.241***	0.782	0.766	1	-	-	-	-
46	0.178***	0.374	0.178***	0.374	.606	1	0.178***	0.373	0.625	1	0.233***	0.488	0.491	1
47	0.245***	0.870	0.245***	0.870	.823	1	0.245***	0.870	0.825	1	-	-	-	-
48	0.240***	0.767	0.240***	0.767	.756	1	0.240***	0.767	0.760	1	0.232***	0.739	0.711	1
49	0.130***	0.261	0.130***	0.261	.626	1	0.129***	0.260	0.645	1	0.186***	0.373	0.525	1
50	0.219***	0.548	0.219***	0.548	.668	1	0.219***	0.546	0.666	1	0.283***	0.708	0.548	1
51	0.041***	0.112	0.041***	0.112	.374	1	0.041***	0.112	0.410	1	-	-	-	-
52	0.161***	0.325	0.161***	0.325	.642	0	-		-	-	-	-	-	-
53	0.235***	0.669	0.235***	0.669	.731	1	0.235***	0.669	0.731	1	-	-	-	-
54	0.046***	0.122	0.046***	0.122	.507	1	0.045***	0.121	0.504	1	-	-	-	-
55	0.222***	0.558	0.222***	0.558	.671	1	0.221***	0.557	0.675	1	0.274***	0.689	0.525	1

Note. *p < 0.05; **p < 0.01; ***p < 0.001, LH = Loevinger's *H*, AISP = Automatic Item Selection Procedure, Stan. *B* = Standarized Beta

Appendix B Lexical Knowledge ITOS Rasch Item Fit Statistics

		51 items			6 items	
Item	SX-2	RMSEA	*p*	SX-2	RMSEA	*p*
4	52.242	0.077	0.000	-	-	-
5	56.012	0.075	0.000	-	-	-
6	68.429	0.084	0.000	-	-	-
7	49.366	0.069	0.000	-	-	-
8	27.510	0.048	0.001	-	-	-
9	15.225	0.033	0.033	-	-	-
10	21.433	0.039	0.006	-	-	-
11	35.521	0.045	0.000	-	-	-
12	30.019	0.040	0.002	-	-	-
13	36.478	0.044	0.000	-	-	-
14	32.387	0.040	0.001	-	-	-
15	34.127	0.041	0.001	-	-	-
16	61.422	0.056	0.000	-	-	-
17	33.838	0.041	0.001	-	-	-
18	33.588	0.041	0.001	-	-	-
19	32.996	0.035	0.003	-	-	-
20	31.727	0.037	0.003	-	-	-
21	27.139	0.030	0.018	-	-	-
22	32.904	0.038	0.002	-	-	-
23	42.758	0.038	0.001	-	-	-
24	26.616	0.025	0.046	-	-	-
25	29.137	0.028	0.023	-	-	-
26	25.521	0.024	0.061	-	-	-
27	27.359	0.026	0.038	-	-	-
28	28.598	0.027	0.027	-	-	-
29	27.807	0.024	0.047	-	-	-
30	33.290	0.032	0.007	-	-	-
31	32.147	0.031	0.010	-	-	-
32	27.528	0.026	0.036	-	-	-
33	26.090	0.024	0.053	-	-	-
34	28.774	0.039	0.002	-	-	-
35	32.772	0.031	0.008	-	-	-
36	35.751	0.034	0.003	-	-	-
37	41.255	0.038	0.001	-	-	-
38	27.837	0.023	0.065	-	-	-

	51 items			6 items		
Item	**SX-2**	**RMSEA**	***p***	**SX-2**	**RMSEA**	***p***
39	18.900	0.023	0.091	5.284	0.027	0.152
40	27.124	0.025	0.040	-	-	-
41	29.672	0.026	0.029	-	-	-
42	58.178	0.063	0.000	-	-	-
43	38.402	0.036	0.001	-	-	-
44	27.558	0.026	0.036	-	-	-
45	28.512	0.027	0.027	-	-	-
46	21.814	0.033	0.016	6.016	0.043	0.049
47	39.735	0.037	0.001	-	-	-
48	19.764	0.015	0.231	11.766	0.052	0.008
49	11.923	0.021	0.155	4.675	0.035	0.097
50	15.752	0.020	0.151	8.699	0.042	0.034
51	97.703	0.131	0.000	-	-	-
53	30.994	0.036	0.003	-	-	-
54	37.165	0.077	0.000	-	-	-
55	17.021	0.023	0.107	4.068	0.018	0.254

Appendix C Induction ITOS CFA Parameters and Mokken Analysis

	33 Items		31 Items				9 items				8 items			
Item	***B***	**Stan. *B***	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**
1	0.028***	0.169	0.027***	0.163	0.211	0	-	-	-	-	-	-	-	-
2	0.015	0.082	-	-	-	-	-	-	-	-	-	-	-	-
3	0.024***	0.124	0.020*	0.106	0.123	0	-	-	-	-	-	-	-	-
4	0.052***	0.194	0.050***	0.188	0.157	0	-	-	-	-	-	-	-	-
5	0.028***	0.199	0.027***	0.192	0.283	1	0.022***	0.155	0.320	1	0.02**	0.143	.327	1
6	0.034***	0.157	0.033***	0.155	0.168	0	-	-	-	-	-	-	-	-
7	0.020***	0.180	0.018***	0.167	0.299	1	0.012*	0.111	0.312	1	-	-	-	-
8	0.027***	0.191	0.026***	0.186	0.283	3	-	-	-	-	-	-	-	-
9	0.018***	0.088	-	-	-	-	-	-	-	-	-	-	-	-
10	0.026	0.213	0.025***	0.205	0.332	1	0.19***	0.156	0.352	1	0.02**	0.135	.339	1
11	0.048***	0.196	0.046***	0.189	0.171	0	-	-	-	-	-	-	-	-
12	0.095***	0.197	0.099***	0.204	0.087	0	-	-	-	-	-	-	-	-
13	0.116***	0.233	0.119***	0.237	0.107	0	-	-	-	-	-	-	-	-
14	0.134***	0.297	0.136***	0.300	0.146	3	-	-	-	-	-	-	-	-
15	0.97***	0.209	0.103***	0.220	0.094	0	-	-	-	-	-	-	-	-
16	0.109***	0.255	0.109***	0.254	0.144	3	-	-	-	-	-	-	-	-
17	0.060**	0.120	0.053*	0.106	0.062	0	-	-	-	-	-	-	-	-
18	0.109***	0.221	0.110***	0.223	0.111	0	-	-	-	-	-	-	-	-
19	0.233***	0.506	0.230***	0.500	0.259	1	0.234***	0.508	0.355	1	0.23***	0.506	.355	1
20	0.107***	0.213	0.115***	0.231	0.087	0	-	-	-	-	-	-	-	-
21	0.194***	0.425	0.192***	0.420	0.210	2	-	-	-	-	-	-	-	-
22	0.118***	0.272	0.120***	0.277	0.142	0	-	-	-	-	-	-	-	-
23	0.197***	0.408	0.201***	0.415	0.197	0	-	-	-	-	-	-	-	-
24	0.047*	0.105	0.048*	0.106	0.051	0	-	-	-	-	-	-	-	-
25	0.149***	0.319	0.153***	0.327	0.141	2	-	-	-	-	-	-	-	-
26	0.139***	0.316	0.143***	0.323	0.160	1	0.145***	0.330	0.338	1	0.15***	0.329	.337	1
27	0.280***	0.605	0.281***	0.608	0.306	1	0.282***	0.609	0.385	1	0.28***	0.613	.388	1
28	0.223***	0.524	0.226***	0.532	0.287	1	0.236***	0.553	0.417	1	0.24***	0.559	.423	1
29	0.118***	0.250	0.117***	0.248	0.106	0	-	-	-	-	-	-	-	-
30	0.180***	0.360	0.184***	0.369	0.149	0	-	-	-	-	-	-	-	-
31	0.138***	0.308	0.143***	0.319	0.140	1	0.143***	0.320	0.321	1	0.14***	0.318	.317	1
32	0.247***	0.496	0.247***	0.497	0.224	1	0.256***	0.513	0.384	1	0.26***	0.516	.385	1
33	0.043*	0.108	0.043*	0.108	0.055	0	-	-	-	-	-	-	-	-

Note. **p* < 0.05; ***p* < 0.01; ****p* < 0.001, LH = Loevinger’s *H*, AISP = Automatic Item Selection Procedure, Stan. *B* = Standarized Beta

Appendix D Induction ITOS Rasch Item Fit Statistics

		9-items			8-items	
Item	**SX-2**	**RMSEA**	***p P***	**SX-2**	**RMSEA**	***p***
5	4.80	0.03	0.19	6.12	0.04	0.11
7	-	-	-	-	-	-
10	1.13	0.01	0.29	2.00	0.04	0.16
19	5.91	0.03	0.21	5.03	0.02	0.28
26	8.11	0.05	0.04	8.35	0.06	0.04
27	9.355	0.05	0.05	10.80	0.05	0.03
28	1.64	0.00	0.80	1.38	0.00	0.85
31	5.28	0.04	0.15	5.21	0.03	0.16
32	1.48	0.00	0.83	0.69	0.00	0.88

Appendix E Visualisation ITOS CFA Paramters and Mokken Analysis

	52 items				37 items				17 items				14 items			
Item	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**
1	0.009	0.068	0.193	3	-	-	-	-	-	-	-	-	-	-	-	-
2	0.047***	0.193	0.241	2	-	-	-	-	-	-	-	-	-	-	-	-
3	0.018***	0.175	0.537	1	0.016***	0.164	0.561	1	-	-	-	-	-	-	-	-
4	0.074***	0.190	0.161	2	-	-	-	-	-	-	-	-	-	-	-	-
5	0.026***	0.125	0.182	0	-	-	-	-	-	-	-	-	-	-	-	-
6	0.058***	0.231	0.285	2	-	-	-	-	-	-	-	-	-	-	-	-
7	0.039***	0.127	0.129	0	-	-	-	-	-	-	-	-	-	-	-	-
8	0.088***	0.264	0.237	2	-	-	-	-	-	-	-	-	-	-	-	-
9	0.066***	0.197	0.176	4	-	-	-	-	-	-	-	-	-	-	-	-
10	0.093***	0.295	0.269	0	-	-	-	-	-	-	-	-	-	-	-	-
11	0.117***	0.395	0.374	1	0.112***	0.379	0.404	1	-	-	-	-	-	-	-	-
12	0.121***	0.360	0.323	1	0.116***	0.346	0.346	1	-	-	-	-	-	-	-	-
13	0.133***	0.384	0.305	1	0.130***	0.375	0.335	1	-	-	-	-	-	-	-	-
14	0.149***	0.367	0.28	3	-	-	-	-	-	-	-	-	-	-	-	-
15	0.125***	0.251	0.206	3	-	-	-	-	-	-	-	-	-	-	-	-
16	0.177***	0.437	0.303	1	0.176***	0.434	0.338	1	-	-	-	-	-	-	-	-
17	0.118***	0.236	0.193	4	-	-	-	-	-	-	-	-	-	-	-	-
18	0.225***	0.531	0.377	1	0.222***	0.524	0.409	1	-	-	-	-	-	-	-	-
19	0.219***	0.686	0.581	1	0.216***	0.679	0.675	1	0.209***	0.656	0.707	1	-	-	-	-
20	0.233***	0.613	0.442	1	0.231***	0.608	0.494	1	-	-	-	-	-	-	-	-
21	0.221***	0.481	0.34	1	0.219***	0.477	0.373	1	-	-	-	-	-	-	-	-
22	0.265***	0.710	0.502	1	0.263***	0.707	0.564	1	0.267***	0.718	0.615	1	-	-	-	-
23	0.172***	0.344	0.235	0	-	-	-	-	-	-	-	-	-	-	-	-
24	0.201***	0.401	0.3	1	0.199***	0.399	0.336	1	-	-	-	-	-	-	-	-
25	0.124***	0.258	0.224	0	-	-	-	-	-	-	-	-	-	-	-	-
26	0.272***	0.685	0.464	1	0.271***	0.683	0.517	1	0.278***	0.700	0.558	1	0.271***	0.683	0.547	1
27	0.222***	0.444	0.339	1	0.220***	0.440	0.383	1	-	-	-	-	-	-	-	-
28	0.281***	0.731	0.51	1	0.281***	0.731	0.577	1	0.282***	0.732	0.618	1	0.279***	0.726	0.623	1
29	0.264***	0.583	0.375	1	0.264***	0.582	0.411	1	-	-	-	-	-	-	-	-
30	0.283***	0.695	0.457	1	0.283***	0.696	0.514	1	0.287***	0.704	0.556	1	0.282***	0.693	0.544	1
31	0.187***	0.380	0.338	1	0.186***	0.378	0.394	1	0.180***	0.366	0.377	1	0.184***	0.375	0.387	1
32	0.302***	0.818	0.571	1	0.303***	0.821	0.652	1	-	-	-	-	-	-	-	-
33	0.299***	0.728	0.482	1	0.300***	0.731	0.543	1	0.303***	0.738	0.588	1	0.296***	0.721	0.567	1
34	0.308***	0.834	0.583	1	0.310***	0.841	0.676	1	-	-	-	-	-	-	-	-
35	0.219***	0.438	0.314	1	0.218***	0.437	0.351	1	0.215***	0.431	0.354	1	0.221***	0.443	0.352	1

	52 items				**37 items**				**17 items**				**14 items**			
Item	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**
36	0.261***	0.526	0.376	1	0.262***	0.528	0.430	1	0.252***	0.509	0.423	1	0.263***	0.530	0.422	1
37	0.285***	0.643	0.414	1	0.288***	0.648	0.464	1	0.289***	0.651	0.509	1	0.292***	0.659	0.494	1
38	0.237***	0.480	0.328	1	0.240***	0.487	0.380	1	0.234***	0.474	0.382	1	0.238***	0.482	0.376	1
39	0.302***	0.713	0.473	1	0.306***	0.720	0.540	1	-	-	-	-	-	-	-	-
40	0.231***	0.463	0.327	1	0.231***	0.463	0.367	1	0.237***	0.475	0.39	1	0.241***	0.483	0.376	1
41	0.250***	0.501	0.355	1	0.251***	0.503	0.405	1	0.253***	0.509	0.42	1	0.262***	0.527	0.409	1
42	0.211***	0.425	0.287	1	0.213***	0.428	0.326	1	-	-	-	-	-	-	-	-
43	0.120***	0.262	0.261	1	0.122***	0.266	0.314	1	-	-	-	-	-	-	-	-
44	0.097***	0.216	0.197	1	0.099***	0.220	0.237	1	-	-	-	-	-	-	-	-
45	0.282***	0.607	0.407	1	0.283***	0.610	0.455	1	0.281***	0.606	0.486	1	0.288***	0.620	0.476	1
46	0.298***	0.669	0.45	1	0.300***	0.672	0.503	1	0.288***	0.646	0.516	1	0.291***	0.653	0.511	1
47	0.295***	0.614	0.431	1	0.295***	0.615	0.480	1	-	-	-	-	-	-	-	-
48	0.173***	0.353	0.302	1	0.173***	0.352	0.345	1	0.180***	0.366	0.362	1	-	-	-	-
49	0.252***	0.507	0.354	1	0.253***	0.509	0.402	1	0.249***	0.501	0.402	1	0.252***	0.508	0.384	1
50	0.302***	0.642	0.446	1	0.303***	0.645	0.500	1	-	-	-	-	-	-	-	-
51	0.289***	0.610	0.425	1	0.290***	0.613	0.477	1	-	-	-	-	-	-	-	-
52	0.135***	0.282	0.232	2	-	-	-	-	-	-	-	-	-	-	-	-

Note. *p < 0.05; **p < 0.01; ***p < 0.001, LH = Loevinger's *H*, AISP = Automatic Item Selection Procedure, Stan. *B* = Standarized Beta

Appendix F Visualisation ITOS Rasch Item Fit Statistics

	37 items			17 items			14 items		
Item	**SX-2**	**RMSEA**	***p***	**SX-2**	**RMSEA**	***p***	**SX-2**	**RMSEA**	***p***
3	10.867	0.112	0.001	-	-	-	-	-	-
11	95.445	0.065	0.000	-	-	-	-	-	-
12	155.772	0.082	0.000	-	-	-	-	-	-
13	133.000	0.074	0.000	-	-	-	-	-	-
16	118.387	0.064	0.000	-	-	-	-	-	-
18	99.160	0.063	0.000	-	-	-	-	-	-
19	42.574	0.030	0.016	66.419	0.080	0.000	-	-	-
20	52.449	0.035	0.002	-	-	-	-	-	-
21	90.141	0.061	0.000	-	-	-	-	-	-
22	33.068	0.019	0.160	28.669	0.042	0.004	-	-	-
24	39.740	0.035	0.005	-	-	-	-	-	-
26	29.384	0.008	0.393	24.397	0.036	0.018	26.680	0.050	0.002
27	43.511	0.039	0.002	-	-	-	-	-	-
28	28.030	0.007	0.409	19.376	0.028	0.080	21.955	0.043	0.009
29	53.124	0.039	0.001	-	-	-	-	-	-
30	31.668	0.015	0.245	18.326	0.026	0.106	12.138	0.021	0.206
31	31.516	0.031	0.025	11.336	0.018	0.253	13.392	0.034	0.063
32	53.376	0.037	0.001	-	-	-	-	-	-
33	38.565	0.026	0.041	25.197	0.037	0.014	12.979	0.024	0.164
34	63.757	0.043	0.000	-	-	-	-	-	-
35	31.951	0.028	0.044	15.470	0.026	0.116	12.107	0.026	0.147
36	34.115	0.030	0.025	12.191	0.017	0.272	14.405	0.032	0.072
37	30.616	0.019	0.165	11.360	0.006	0.414	11.639	0.019	0.234
38	18.574	0.000	0.612	16.094	0.028	0.097	12.595	0.027	0.127
39	45.079	0.033	0.006	-	-	-	-	-	-
40	24.524	0.017	0.220	11.031	0.011	0.355	11.490	0.024	0.175
41	24.540	0.017	0.220	21.853	0.039	0.016	12.420	0.027	0.133
42	54.779	0.047	0.000	-	-	-	-	-	-
43	47.108	0.047	0.000	-	-	-	-	-	-
44	138.189	0.095	0.000	-	-	-	-	-	-
45	30.264	0.020	0.142	15.878	0.024	0.146	8.203	0.000	0.514
46	35.985	0.025	0.055	15.437	0.023	0.163	11.924	0.020	0.218

47	42.462	0.036	0.004	-	-	-	-	-	-
48	31.183	0.031	0.027	24.202	0.046	0.004	-	-	-
49	25.418	0.019	0.186	10.854	0.010	0.369	7.962	0.000	0.437
50	54.717	0.044	0.000	-	-	-	-	-	-
51	42.529	0.036	0.004	-	-	-	-	-	-

Appendix G Working Memory ITOS CFA Parameters and Mokken Analysis

	38 Items		37 items				23 items				19 items			
Item	***B***	**Stan. *B***	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**	***B***	**Stan. *B***	**LH**	**AISP**
1	0.047***	0.218	0.047***	0.218	0.195	0	-	-	-	-	-	-	-	-
2	0.052***	0.354	0.051***	0.353	0.380	1	0.044***	0.299	0.350	1	0.031***	0.211	0.333	1
3	0.046***	0.224	0.046***	0.223	0.166	0	-	-	-	-	-	-	-	-
4	0.043***	0.313	0.043***	0.312	0.321	0	-	-	-	-	-	-	-	-
5	0.044***	0.355	0.043***	0.355	0.356	2	-	-	-	-	-	-	-	-
6	0.047***	0.247	0.046***	0.246	0.175	2	-	-	-	-	-	-	-	-
7	0.039***	0.318	0.039***	0.317	0.302	1	0.035***	0.286	.307	1	-	-	-	-
8	0.050*	0.107	-	-	-	-	-	-	-	-	-	-	-	-
9	0.112***	0.374	0.112***	0.375	0.240	2	-	-	-	-	-	-	-	-
10	0.075***	0.572	0.075***	0.573	0.615	1	0.072***	0.551	0.670	1	-	-	-	-
11	0.111***	0.605	0.111***	0.607	0.512	1	0.116***	0.631	0.598	1	0.094***	0.513	0.609	1
12	0.145***	0.728	0.145***	0.730	0.559	1	0.148***	0.745	0.648	1	-	-	-	-
13	0.151***	0.723	0.151***	0.723	0.551	1	0.156***	0.748	0.645	1	-	-	-	-
14	0.147***	0.638	0.148***	0.639	0.533	1	0.154***	0.669	0.624	1	0.148***	0.642	0.644	1
15	0.160***	0.751	0.160***	0.751	0.606	1	0.166***	0.778	0.702	1	0.146***	0.686	0.721	1
16	0.171***	0.570	0.171***	0.570	0.451	1	0.174***	0.580	0.518	1	0.175***	0.584	0.507	1
17	0.177***	0.464	0.177***	0.464	0.362	1	0.177***	0.465	0.382	1	0.196***	0.513	0.369	1
18	0.204***	0.479	0.204***	0.478	0.342	1	0.202***	0.475	0.367	1	0.221***	0.517	0.346	1
19	0.176***	0.398	0.176***	0.397	0.314	1	0.173***	0.391	0.337	1	0.209***	0.472	0.327	1
20	0.166***	0.388	0.166***	0.388	0.297	1	0.164***	0.384	0.318	1	0.188***	0.438	0.304	1
21	0.183***	0.395	0.182***	0.393	0.308	1	0.178***	0.383	0.343	1	0.207***	0.447	0.328	1
22	0.167***	0.354	0.167***	0.353	0.293	1	0.159***	0.337	0.304	1	0.192***	0.407	0.293	0
23	0.176***	0.358	0.175***	0.357	0.311	1	0.163***	0.333	0.347	1	0.208***	0.423	0.337	1
24	0.182***	0.431	0.182***	0.430	0.410	1	0.176***	0.417	0.425	1	0.224***	0.531	0.421	1
25	0.145***	0.292	0.145***	0.291	0.282	1	0.134***	0.268	0.359	1	0.181***	0.362	0.352	1
26	0.138***	0.275	0.137***	0.274	0.290	3	-	-	-	-	-	-	-	-
27	0.128***	0.260	0.127***	0.259	0.262	0	-	-	-	-	-	-	-	-
28	0.113***	0.237	0.113***	0.236	0.253	0	-	-	-	-	-	-	-	-
29	0.139***	0.287	0.138***	0.286	0.307	1	0.122***	0.252	0.357	1	0.170***	0.351	0.350	1
30	0.117***	0.239	0.118***	0.240	0.271	0	-	-	-	-	-	-	-	-
31	0.106***	0.238	0.105***	0.237	0.278	1	0.090***	0.202	0.313	1	0.117***	0.262	0.303	1
32	0.054**	0.132	0.054***	0.131	0.203	0	-	-	-	-	-	-	-	-

33	0.124***	0.256	0.123***	0.256	0.288	3	-	-	-	-	-	-	-	-
34	0.125***	0.251	0.125***	0.251	0.272	0	-	-	-	-	-	-	-	-
35	0.133***	0.273	0.132***	0.271	0.284	1	0.116***	0.240	0.329	1	0.148***	0.304	0.318	1
36	0.060***	0.172	0.060***	0.171	0.345	1	0.047**	0.136	0.358	1	0.069***	0.199	0.353	1
37	0.078***	0.195	0.078***	0.195	0.279	1	0.068***	0.169	0.324	1	0.088***	0.219	0.313	1
38	0.082***	0.185	0.081***	0.185	0.192	0	-	-	-	-	-	-	-	-

Note. *p < 0.05; **p < 0.01; ***p < 0.001, LH = Loevinger's *H*, AISP = Automatic Item Selection Procedure, Stan. *B* = Standarized Beta

Appendix H Working Memory ITOS Rasch Item Fit Statistics

		23 items			19 items	
Item	**SX-2**	**RMSEA**	***p***	**SX-2**	**RMSEA**	***p***
2	9.554	0.055	0.049	9.360	0.068	0.025
7	5.403	0.098	0.020	-	-	-
10	9.269	0.089	0.010	-	-	-
11	9.157	0.043	0.103	9.794	0.037	0.134
12	6.525	0.000	0.480	-	-	-
13	8.022	0.002	0.431	-	-	-
14	10.990	0.015	0.358	9.003	0.001	0.437
15	11.944	0.033	0.154	10.513	0.026	0.231
16	9.906	0.000	0.449	16.481	0.038	0.087
17	19.089	0.044	0.039	15.811	0.036	0.105
18	10.954	0.014	0.361	10.975	0.022	0.277
19	21.264	0.050	0.019	20.483	0.053	0.015
20	17.110	0.039	0.072	17.288	0.045	0.044
21	14.912	0.033	0.135	10.177	0.017	0.336
22	14.381	0.036	0.109	13.611	0.033	0.137
23	9.327	0.000	0.501	14.113	0.030	0.168
24	17.412	0.040	0.066	14.460	0.036	0.107
25	13.952	0.035	0.124	10.374	0.025	0.240
29	12.212	0.034	0.142	13.034	0.037	0.111
31	5.648	0.000	0.775	6.558	0.000	0.585
35	9.008	0.017	0.342	7.484	0.000	0.485
36	3.640	0.000	0.820	3.602	0.000	0.824
37	4.763	0.000	0.783	6.088	0.000	0.637

Appendix I Lexical Knowledge ICS Rasch, CFA and Mokken Analysis

	72 Item Rasch		66 Item Rasch		54 Item Rasch		54 Item Mokken	48 Item Mokken	47 Item CFA	47 Item Rasch	
Item	**SX-2**	**RMSEA**	**SX-2**	**RMSEA**	**SX-2**	**RMSEA**	**LH**	**LH**	**Stan. *B***	**SX-2**	**RMSEA**
4	NA	NA	-	-	-	-	-	-	-	-	-
5	NA	NA	-	-	-	-	-	-	-	-	-
6	NA	NA	NA	NA	NA	NA	0.514	0.546	0.700***	NA	NA
7	NA	NA	-	-	-	-	-	-	-	-	-
8	NA	NA	-	-	-	-	-	-	-	-	-
9	NA	NA	-	-	-	-	-	-	-	-	-
11	NA	NA	-	-	-	-	-	-	-	-	-
13	4.717	0.000	4.732	0.000	2.333	0.000	0.304	0.323	0.541***	2.777	0.000
14	4.190	0.011	4.206	0.012	13.032*	0.066	0.111	-	-	-	-
36	26.720	0.039	24.481	0.031	21.034	0.017	0.346	0.372	0.700***	17.170	0.014
37	36.623**	0.056	35.036**	0.053	21.850	0.028	0.407	0.427	-	-	-
38	41.617**	0.044	41.408*	0.044	36.921*	0.040	0.441	0.459	0.871***	26.042	0.029
39	38.320	0.025	37.469	0.024	43.187*	0.036	0.298	0.307	0.604***	35.041	0.031
40	30.212*	0.049	31.579**	0.055	23.736*	0.047	0.411	0.436	0.767***	27.944**	0.052
41	13.435	0.000	16.360	0.000	19.847	0.011	0.343	0.360	0.713***	17.740	0.011
42	46.710*	0.041	45.920*	0.040	-	-	-	-	-	-	-
43	13.680	0.019	14.139	0.022	7.143	0.000	0.350	0.364	0.673***	8.691	0.000
44	37.046	0.034	36.243	0.033	35.603	0.029	0.394	0.418	0.807***	34.638	0.037
45	9.883	0.000	7.869	0.000	10.035	0.000	0.257	0.271	0.534***	16.950	0.019
46	36.756	0.025	39.785	0.030	20.329	0.000	0.257	0.275	0.501***	31.138	0.036
48	24.853	0.035	23.535	0.036	20.302	0.027	0.145	-	-	-	-
49	28.270	0.015	29.633	0.019	38.395**	0.052	0.276	0.295	0.408***	32.918	0.053
50	42.731	0.030	37.134	0.025	35.274	0.029	0.360	0.383	0.729***	27.493	0.016
53	47.889*	0.038	48.021*	0.038	41.373	0.034	0.411	0.431	0.828***	32.485	0.026
54	143.419***	0.122	139.166***	0.120	-	-	-	-	-	-	-
55	41.630	0.028	42.348	0.031	-	-	-	-	-	-	-
56	64.118***	0.052	66.531***	0.054	63.430***	0.058	0.246	0.263	0.539***	73.523***	0.075
57	17.604*	0.057	17.023*	0.055	12.584	0.039	0.308	0.317	0.550***	11.370	0.034
58	NA	NA	NA	NA	NA	NA	0.128	0.140	0.260***	NA	NA
59	28.737	0.000	29.481	0.000	26.574	0.013	0.330	0.352	0.677***	31.425	0.031
60	27.851	0.009	31.159	0.017	38.239	0.033	0.328	0.337	0.678***	22.327	0.000
62	42.528***	0.064	44.360***	0.063	29.562	0.039	0.442	0.464	0.845***	24.789	0.038
63	36.040	0.037	40.829*	0.043	35.775	0.036	0.205	-	-	-	-
64	30.911	0.033	30.943	0.031	25.729	0.018	0.447	0.469	0.852***	29.525	0.036
65	71.111***	0.068	71.644***	0.069	-	-	-	-	-	-	-

	72 Item Rasch		66 Item Rasch		54 Item Rasch		54 Item Mokken	48 Item Mokken	47 Item CFA	47 Item Rasch	
Item	**SX-2**	**RMSEA**	**SX-2**	**RMSEA**	**SX-2**	**RMSEA**	**LH**	**LH**	**Stan. *B***	**SX-2**	**RMSEA**
66	37.689	0.035	37.465	0.034	27.878	0.033	0.368	0.394	0.579***	32.627	0.047
67	28.545	0.000	25.309	0.000	31.499	0.024	0.327	0.348	0.660***	25.054	0.002
69	59.566**	0.048	58.389**	0.047	50.103**	0.048	0.445	0.465	0.891***	42.570**	0.044
70	34.665	0.028	31.719	0.022	34.765	0.028	0.236	0.247	0.523***	34.877	0.030
71	40.700	0.027	44.125	0.032	58.419***	0.054	0.179	-	-	-	-
72	30.047	0.000	28.587	0.000	42.302	0.043	0.246	0.265	0.512***	26.967	0.022
73	13.675	0.000	13.838	0.000	23.173	0.000	0.294	0.303	0.599***	22.900	0.016
74	23.876	0.000	22.660	0.000	27.419	0.006	0.265	0.277	0.582***	29.302	0.022
75	52.551**	0.045	54.605**	0.047	-	-	-	-	-	-	-
76	53.273**	0.049	55.647***	0.052	68.611***	0.076	0.156	-	-	-	-
77	42.320**	0.048	41.309*	0.046	27.529	0.041	0.257	0.276	0.340***	42.351***	0.067
78	43.232	0.031	42.353	0.030	33.916	0.026	0.362	0.377	0.758***	24.113	0.000
79	107.709***	0.084	107.538***	0.083	-	-	-	-	-	-	-
80	45.577*	0.037	46.674*	0.039	64.721***	0.070	0.183	-	-	-	-
81	37.626	0.022	37.095	0.021	32.000	0.020	0.355	0.375	0.723***	26.025	0.011
82	88.829***	0.088	90.583***	0.089	-	-	-	-	-	-	-
83	100.963***	0.084	102.833***	0.083	-	-	-	-	-	-	-
84	111.317***	0.105	112.834***	0.105	-	-	-	-	-	-	-
85	48.854**	0.047	49.677**	0.046	40.182*	0.036	0.297	0.312	0.649***	30.141	0.021
86	27.572	0.029	26.501	0.023	25.546	0.021	0.396	0.415	0.769***	23.208	0.021
87	30.508	0.019	28.438	0.006	24.678	0.000	0.302	0.315	0.657***	22.954	0.000
88	52.895***	0.079	54.275***	0.080	-	-	-	-	-	-	-
89	43.645	0.031	43.240	0.031	31.101	0.020	0.301	0.308	0.636***	31.654	0.027
90	101.022***	0.088	99.333***	0.087	-	-	-	-	-	-	-
91	28.644	0.013	30.171	0.018	23.591	0.000	0.361	0.378	0.742***	26.851	0.014
92	45.659**	0.043	46.749**	0.042	29.705	0.016	0.363	0.382	0.756***	29.416	0.019
93	25.470	0.037	29.314**	0.044	30.051*	0.045	0.422	0.450	0.812***	25.407**	0.043
94	32.324	0.005	33.731	0.012	23.746	0.000	0.269	0.284	0.590***	36.324	0.035
95	41.914*	0.037	41.093	0.034	36.336	0.031	0.241	0.253	0.544***	42.730	0.042
96	100.165***	0.088	101.090***	0.088	128.808***	0.121	0.056	0.056	0.086	156.833***	0.144
97	55.230**	0.044	56.383**	0.045	42.015	0.037	0.362	0.379	0.738***	45.004**	0.046
98	87.846***	0.078	90.466***	0.078	-	-	-	-	-	-	-
99	57.402***	0.055	57.993***	0.054	37.032	0.032	0.451	0.465	0.889***	45.635**	0.045
100	68.640***	0.057	66.777***	0.056	72.227***	0.065	0.141	0.146	0.312***	86.906***	0.079
101	46.327**	0.038	45.326**	0.037	49.571***	0.058	0.277	0.286	0.491***	35.744*	0.044
102	42.228	0.031	47.019*	0.039	49.274**	0.051	0.176	0.180	0.359***	56.362***	0.063
103	42.700*	0.038	41.767*	0.036	39.882*	0.047	0.262	0.281	0.476***	43.865**	0.054

Note. Results retrieved from Imputation 1 analysis. See Github repository for full results. **p* < 0.05; ***p* < 0.01; ****p* < 0.001, LH = Loevinger's *H*, Stan. *B* = Standarized Beta

Appendix J Lexical Knowledge Item Performance for Poor Fitting Items

Item	Score	6-17	18-29	30-39	40-49	50-59	60-69	70-79	80-90
38	0	39	0	0	1	0	1	1	1
	1	43	142	34	10	36	43	19	2
40	0	24	1	0	0	0	0	0	0
	1	58	141	34	11	36	44	20	3
42	0	54	49	14	4	21	21	11	3
	1	28	93	20	7	15	23	9	0
56	0	48	65	5	1	1	1	1	0
	1	34	77	29	10	35	43	19	3
62	0	29	2	0	0	0	1	1	0
	1	53	140	34	11	36	43	19	3
65	0	56	88	22	6	28	29	14	3
	1	26	54	12	5	8	15	6	0
75	0	47	61	15	4	13	17	7	1
	1	35	81	19	7	23	27	13	2
79	0	41	76	15	3	16	16	8	1
	1	41	66	19	8	20	28	12	2
82	0	57	110	19	8	25	38	16	2
	1	25	32	15	3	11	6	4	1
83	0	44	100	14	6	15	13	4	1
	1	38	42	20	5	21	31	16	2
84	0	62	121	28	9	31	38	11	1
	1	20	21	6	2	5	6	9	2
88	0	72	134	34	9	34	37	16	3
	1	10	8	0	2	2	7	4	0
90	0	53	92	20	6	19	28	10	2
	1	29	50	14	5	17	16	10	1
96	0	49	81	22	8	27	25	10	3
	1	33	61	12	3	9	19	10	0
98	0	45	68	20	7	23	25	13	2
	1	37	74	14	4	13	19	7	1
99	0	55	13	0	0	0	0	0	0
	1	27	129	34	11	36	44	20	3
100	0	41	39	6	1	7	12	6	1
	1	41	103	28	10	29	32	14	2
102	0	50	30	7	8	17	22	14	1
	1	32	112	27	3	19	22	6	2

Appendix K Lexical Knowledge Pooled Item Parameters

Item	a	*d*	*g*	*u*
6	1	5.615	0	1
13	1	4.331	0	1
36	1	2.951	0	1
38	1	2.601	0	1
39	1	1.161	0	1
40	1	3.298	0	1
41	1	2.913	0	1
43	1	3.565	0	1
44	1	2.253	0	1
45	1	3.073	0	1
46	1	0.314	0	1
49	1	-0.828	0	1
50	1	0.672	0	1
53	1	1.214	0	1
56	1	0.909	0	1
57	1	4.058	0	1
58	1	5.615	0	1
59	1	0.520	0	1
60	1	1.933	0	1
62	1	2.951	0	1
64	1	2.633	0	1
66	1	-0.566	0	1
67	1	0.639	0	1
69	1	1.582	0	1
70	1	2.061	0	1
72	1	0.473	0	1
73	1	2.455	0	1
74	1	1.484	0	1
77	1	-1.325	0	1
78	1	1.523	0	1
81	1	0.781	0	1
85	1	1.980	0	1
86	1	2.602	0	1
87	1	1.933	0	1
89	1	1.542	0	1
91	1	2.037	0	1

Item	a	*d*	*g*	*u*
92	1	1.856	0	1
93	1	3.073	0	1
94	1	1.385	0	1
95	1	1.874	0	1
96	1	-0.638	0	1
97	1	0.992	0	1
99	1	1.915	0	1
100	1	1.042	0	1
101	1	0.245	0	1
102	1	0.484	0	1
103	1	-0.063	0	1

Note. These parameters are generated by MIRT (i.e. are the equivalent slope-intercept translation of traditional/classical IRT parameters)

Appendix LInduction ICS IRT, CFA and Mokken Analysis

	37 Item Rasch			34 Item Rasch			34 Item Mokken	23 Item Mokken	23 Item CFA		23 Item Rasch		
Item	**SX-2**	**RMSEA**	***p***	**SX-2**	**RMSEA**	***p***	**LH**	**LH**	**Std. *B***	***p***	**SX-2**	**RMSEA**	***p***
11	19.472	0.053	0.148	15.356	0.036	0.286	0.323	0.349	0.696	NA	7.686	0.000	0.741
12	16.128	0.008	0.444	14.277	0.012	0.429	0.154	0.176	0.472	0.000	17.315	0.064	0.099
13	24.565	0.062	0.078	20.915	0.059	0.104	0.070	NA	NA	NA	NA	NA	NA
14	12.820	0.000	0.541	12.308	0.000	0.503	0.152	0.195	0.445	0.001	10.258	0.000	0.507
15	11.055	0.000	0.524	9.829	0.000	0.708	0.103	NA	NA	NA	NA	NA	NA
16	14.511	0.000	0.561	23.638	0.070	0.051	0.103	0.125	0.293	0.008	27.769	0.097	0.006
17	38.984	0.101	0.001	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
18	19.194	0.038	0.259	13.238	0.000	0.508	0.086	NA	NA	NA	NA	NA	NA
19	14.206	0.000	0.583	12.889	0.000	0.535	0.245	0.271	0.660	0.000	13.738	0.032	0.318
20	30.857	0.081	0.014	30.618	0.086	0.010	0.034	0.061	0.160	0.177	38.605	0.125	0.000
21	27.322	0.071	0.038	21.561	0.062	0.088	0.197	0.235	0.566	0.000	13.010	0.024	0.368
23	21.461	0.055	0.123	13.897	0.000	0.457	0.139	0.149	0.361	0.006	20.056	0.076	0.045
26	23.508	0.069	0.052	13.782	0.021	0.389	0.116	0.149	0.271	0.040	15.782	0.056	0.149
27	17.058	0.022	0.382	14.703	0.000	0.473	0.177	0.231	0.559	0.000	10.953	0.000	0.447
28	25.198	0.064	0.066	29.287	0.088	0.010	0.264	0.304	0.737	0.000	14.684	0.040	0.259
29	30.001	0.084	0.012	15.279	0.025	0.359	0.086	NA	NA	NA	NA	NA	NA
32	12.620	0.000	0.632	15.017	0.023	0.377	0.182	0.223	0.571	0.000	12.386	0.030	0.335
37	26.571	0.080	0.022	21.953	0.077	0.038	0.423	0.463	0.855	0.000	16.114	0.066	0.096
40	10.439	0.000	0.577	14.019	0.035	0.300	0.089	NA	NA	NA	NA	NA	NA
41	17.321	0.033	0.300	15.715	0.029	0.331	0.088	NA	NA	NA	NA	NA	NA
42	13.197	0.000	0.587	16.422	0.026	0.355	0.109	NA	NA	NA	NA	NA	NA
45	21.951	0.051	0.145	14.442	0.015	0.417	0.215	0.245	0.568	0.000	5.213	0.000	0.950
46	12.763	0.000	0.545	11.173	0.000	0.596	0.299	0.330	0.662	0.000	12.075	0.026	0.358
47	7.194	0.100	0.066	4.720	0.036	0.317	0.482	0.473	0.656	0.000	3.516	0.000	0.475
48	10.277	0.105	0.036	10.862	0.110	0.028	0.041	NA	NA	NA	NA	NA	NA
49	19.063	0.044	0.211	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
52	30.057	0.103	0.003	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
53	19.986	0.042	0.221	14.488	0.000	0.489	0.267	0.285	0.730	0.000	14.609	0.039	0.264
54	8.713	0.000	0.727	12.850	0.022	0.380	0.081	NA	NA	NA	NA	NA	NA
57	21.199	0.060	0.097	14.222	0.026	0.358	0.104	0.129	0.289	0.026	15.539	0.054	0.159
58	11.417	0.000	0.494	14.723	0.040	0.257	0.098	NA	NA	NA	NA	NA	NA

	37 Item Rasch			34 Item Rasch			34 Item Mokken	23 Item Mokken	23 Item CFA		23 Item Rasch		
Item	**SX-2**	**RMSEA**	***p***	**SX-2**	**RMSEA**	***p***	**LH**	**LH**	**Std. *B***	***p***	**SX-2**	**RMSEA**	***p***
59	14.858	0.021	0.388	13.251	0.012	0.429	0.099	0.118	0.272	0.018	13.316	0.039	0.273
62	8.978	0.000	0.879	12.384	0.000	0.576	0.164	0.198	0.489	0.000	15.012	0.042	0.241
63	13.570	0.000	0.558	17.036	0.039	0.254	0.060	NA	NA	NA	NA	NA	NA
64	15.743	0.000	0.471	18.459	0.040	0.239	0.200	0.234	0.604	0.000	16.049	0.049	0.189
67	23.868	0.071	0.048	14.132	0.025	0.365	0.123	0.143	0.331	0.005	6.205	0.000	0.859
68	16.535	0.015	0.416	19.127	0.051	0.160	0.153	0.175	0.428	0.000	15.480	0.045	0.216

Note. Results retrieved from Imputation 1 analysis. See Github repository for full results. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$, LH = Loevinger's *H*, Stan. *B* = Standarized Beta

Appendix M Induction Pooled Item Parameters

Item	**a**	***d***	***g***	***u***
11	1	0.992	0	1
12	1	-0.587	0	1
14	1	-1.415	0	1
37	1	1.359	0	1
45	1	0.152	0	1
46	1	1.039	0	1
47	1	3.052	0	1
53	1	0.268	0	1
57	1	-1.318	0	1
59	1	-1.415	0	1
62	1	-0.949	0	1
64	1	-0.370	0	1
67	1	-1.291	0	1
68	1	-0.233	0	1
32	1	-0.731	0	1
16	1	-0.037	0	1
19	1	-0.064	0	1
20	1	-0.517	0	1
21	1	-0.030	0	1
23	1	-1.019	0	1
26	1	-1.327	0	1
27	1	-0.447	0	1
28	1	0.064	0	1

Note. These parameters are generated by MIRT (i.e. are the equivalent slope-intercept translation of traditional/classical IRT parameters)

Appendix N Visualisation ICS IRT, CFA and Mokken Analysis

	52 Item Rasch			42 Item Rasch			42 Item Mokken	31 Item Mokken	31 Item CFA		30 Item Rasch		
Item	**SX2**	**RMSEA**	***p***	**SX2**	**RMSEA**	***p***	**LH**	**LH**	**Std. *B***	***p***	**SX2**	**RMSEA**	***p***
1	10.978	0.027	0.140	6.135	0.005	0.408	0.081	NA	NA	NA	NA	NA	NA
2	30.662	0.028	0.044	16.751	0.000	0.540	0.127	0.138	0.486	0.000	17.129	0.010	0.377
3	37.946	0.047	0.001	36.814	0.049	0.000	0.222	0.244	0.675	0.000	20.599	0.025	0.112
4	29.358	0.025	0.081	23.847	0.018	0.202	0.088	NA	NA	NA	NA	NA	NA
5	50.505	0.045	0.000	39.187	0.039	0.003	0.193	0.225	0.645	0.000	25.693	0.028	0.058
6	15.632	0.000	0.550	12.364	0.000	0.651	0.134	0.122	0.339	0.000	14.421	0.006	0.419
7	26.694	0.021	0.144	25.291	0.023	0.117	0.075	NA	NA	NA	NA	NA	NA
8	27.555	0.022	0.120	22.557	0.018	0.208	0.100	NA	NA	NA	NA	NA	NA
9	29.873	0.025	0.072	33.017	0.031	0.024	0.106	0.126	0.382	0.000	45.094	0.049	0.000
10	55.318	0.046	0.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	7.419	0.000	0.986	28.303	0.027	0.058	0.070	NA	NA	NA	NA	NA	NA
12	63.691	0.054	0.000	50.062	0.046	0.000	0.212	0.259	0.736	0.000	41.176	0.045	0.001
13	34.155	0.034	0.012	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	30.704	0.025	0.079	28.549	0.026	0.073	0.177	0.215	0.634	0.000	32.642	0.035	0.013
15	27.480	0.020	0.156	28.430	0.030	0.040	0.146	0.166	0.360	0.000	31.491	0.040	0.005
16	26.172	0.020	0.160	42.406	0.040	0.002	0.034	NA	NA	NA	NA	NA	NA
17	64.375	0.052	0.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
18	26.326	0.018	0.194	26.729	0.023	0.111	0.155	0.179	0.473	0.000	12.360	0.000	0.778
19	69.624	0.057	0.000	62.626	0.057	0.000	0.228	0.286	0.815	0.000	46.104	0.050	0.000
20	18.973	0.000	0.524	13.658	0.000	0.751	0.135	0.172	0.504	0.000	12.633	0.000	0.699
21	23.064	0.011	0.341	12.663	0.000	0.855	0.147	0.180	0.495	0.000	12.985	0.000	0.674
22	26.843	0.023	0.108	21.488	0.016	0.256	0.168	0.195	0.516	NA	23.599	0.025	0.099
23	124.252	0.078	0.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
24	24.385	0.012	0.327	44.630	0.042	0.001	0.134	0.153	0.366	0.000	56.154	0.060	0.000
25	85.087	0.063	0.000	158.761	0.105	0.000	-0.001	-0.006	-0.046	0.374	NA	NA	NA
26	21.982	0.011	0.342	18.205	0.004	0.442	0.111	0.131	0.361	0.000	9.793	0.000	0.877
27	68.946	0.053	0.000	46.584	0.044	0.000	0.270	0.312	0.748	0.000	37.381	0.044	0.001
28	18.341	0.000	0.500	17.119	0.000	0.515	0.147	0.179	0.490	0.000	11.740	0.000	0.762
29	17.587	0.000	0.675	17.300	0.000	0.570	0.126	0.150	0.409	0.000	20.074	0.018	0.217
30	20.103	0.003	0.452	26.396	0.023	0.120	0.068	NA	NA	NA	NA	NA	NA
31	28.902	0.022	0.116	27.356	0.028	0.053	0.208	0.230	0.497	0.000	19.498	0.023	0.147

	52 Item Rasch			**42 Item Rasch**			**42 Item Mokken**	**31 Item Mokken**	**31 Item CFA**		**30 Item Rasch**		
Item	**SX2**	**RMSEA**	***p***	**SX2**	**RMSEA**	***p***	**LH**	**LH**	**Std. *B***	***p***	**SX2**	**RMSEA**	***p***
32	17.748	0.012	0.339	27.290	0.038	0.011	0.071	NA	NA	NA	NA	NA	NA
33	18.471	0.000	0.556	19.533	0.006	0.423	0.109	0.136	0.380	0.000	10.102	0.000	0.861
34	12.962	0.000	0.739	16.728	0.012	0.335	0.133	0.162	0.401	0.000	16.808	0.016	0.267
35	48.002	0.039	0.001	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
36	59.984	0.048	0.000	33.556	0.032	0.021	0.232	0.270	0.689	0.000	25.739	0.031	0.041
37	36.611	0.033	0.013	36.962	0.035	0.008	0.058	NA	NA	NA	NA	NA	NA
38	30.664	0.025	0.079	27.135	0.026	0.076	0.097	NA	NA	NA	NA	NA	NA
39	23.369	0.015	0.271	17.835	0.000	0.533	0.137	0.164	0.467	0.000	33.623	0.038	0.006
40	32.279	0.025	0.073	41.548	0.039	0.002	0.116	0.127	0.301	0.000	44.571	0.051	0.000
41	32.845	0.025	0.064	23.968	0.019	0.197	0.159	0.176	0.403	0.000	27.518	0.033	0.025
42	73.279	0.057	0.000	81.590	0.068	0.000	0.027	NA	NA	NA	NA	NA	NA
43	74.703	0.058	0.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
44	157.991	0.093	0.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
45	23.651	0.013	0.310	19.239	0.004	0.442	0.110	0.127	0.361	0.000	27.516	0.028	0.051
46	24.291	0.014	0.279	21.877	0.014	0.290	0.153	0.186	0.540	0.000	15.357	0.000	0.499
47	62.774	0.051	0.000	41.759	0.040	0.002	0.231	0.267	0.686	0.000	30.016	0.034	0.018
48	39.652	0.034	0.008	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
49	26.076	0.016	0.248	12.262	0.000	0.874	0.157	0.182	0.440	0.000	26.812	0.032	0.030
50	76.432	0.059	0.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
51	45.615	0.039	0.001	34.161	0.032	0.018	0.193	0.225	0.602	0.000	25.468	0.028	0.062
52	104.951	0.072	0.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Note. Results retrieved from Imputation 1 analysis. See Github repository for full results. **p* < 0.05; ***p* < 0.01; ****p* < 0.001, LH = Loevinger’s *H*, Stan. *B* = Standarized Beta

Appendix O Visualisation Pooled Item Parameters

Item	a	*d*	*g*	*u*
2	1	2.946	0	1
3	1	3.888	0	1
5	1	2.869	0	1
6	1	3.523	0	1
9	1	2.118	0	1
12	1	2.233	0	1
14	1	1.779	0	1
15	1	-0.020	0	1
18	1	1.869	0	1
19	1	2.822	0	1
20	1	2.528	0	1
21	1	1.452	0	1
22	1	3.031	0	1
24	1	0.357	0	1
26	1	2.614	0	1
27	1	0.331	0	1
28	1	2.966	0	1
29	1	1.511	0	1
31	1	-0.093	0	1
33	1	2.458	0	1
34	1	3.721	0	1
36	1	0.691	0	1
39	1	2.415	0	1
40	1	0.535	0	1
41	1	0.629	0	1
45	1	1.735	0	1
46	1	2.020	0	1
47	1	1.307	0	1
49	1	0.725	0	1
51	1	1.433	0	1

Note. These parameters are generated by MIRT (i.e. are the equivalent slope-intercept translation of traditional/classical IRT parameters)

Appendix P Working Memory ICS IRT, CFA and Mokken Analysis

	34 Item Rasch			34 Item Mokken	26 Item Mokken	26 Item CFA		25 Item Rasch		
Item	**SX2**	**RMSEA**	***p***	**LH**	**LH**	**Std. *B***	***p***	**SX2**	**RMSEA**	***p***
11	14.960	0.049	0.184	0.337	0.389	0.679	NA	15.832	0.071	0.070
12	16.083	0.064	0.097	0.420	0.487	0.851	0.000	7.901	0.029	0.341
13	12.277	0.039	0.267	0.136	0.196	0.373	0.014	10.044	0.054	0.186
14	13.671	0.019	0.397	0.255	0.329	0.715	0.000	19.306	0.064	0.081
15	17.209	0.062	0.102	0.418	0.503	0.861	0.000	13.185	0.056	0.154
16	9.110	0.000	0.693	0.249	0.307	0.562	0.000	8.881	0.000	0.633
17	14.913	0.040	0.246	0.223	0.295	0.506	0.000	9.090	0.000	0.695
18	17.948	0.051	0.160	0.294	0.360	0.717	0.000	8.738	0.000	0.792
19	14.997	0.032	0.308	0.305	0.380	0.818	0.000	16.244	0.041	0.236
20	17.688	0.049	0.170	0.295	0.354	0.735	0.000	12.652	0.000	0.475
21	16.572	0.043	0.220	0.274	0.324	0.724	0.000	17.413	0.040	0.235
22	14.176	0.009	0.437	0.225	0.281	0.613	0.000	14.274	0.011	0.430
23	11.527	0.000	0.644	0.236	0.288	0.590	0.000	7.708	0.000	0.862
24	8.813	0.000	0.843	0.271	0.335	0.619	0.000	12.397	0.015	0.414
25	9.502	0.000	0.734	0.206	0.271	0.575	0.000	14.683	0.039	0.259
26	12.456	0.000	0.491	0.191	0.244	0.528	0.000	10.877	0.000	0.540
27	12.358	0.000	0.499	0.243	0.313	0.688	0.000	13.827	0.032	0.312
28	19.885	0.060	0.098	0.083	0.094	0.183	0.121	50.194	0.146	0.000
29	12.675	0.000	0.473	0.203	0.258	0.519	0.000	11.550	0.000	0.482
30	8.994	0.000	0.773	0.126	0.179	0.380	0.000	17.462	0.055	0.133
31	5.813	0.000	0.758	0.195	0.242	0.435	0.000	7.538	0.000	0.581
32	38.251	0.129	0.000	-0.042	-0.062	-0.123	0.320	NA	NA	NA
33	10.958	0.000	0.447	0.185	0.262	0.510	0.000	7.109	0.000	0.790
34	23.017	0.078	0.028	0.186	0.243	0.416	0.000	21.232	0.079	0.031
35	23.474	0.087	0.015	0.024	0.053	0.095	0.466	37.157	0.126	0.000
36	25.686	0.134	0.001	0.056	0.103	0.126	0.446	24.789	0.131	0.001
37	8.683	0.024	0.370	0.016	NA	NA	NA	NA	NA	NA
38	27.349	0.117	0.001	-0.037	NA	NA	NA	NA	NA	NA
39	110.877	0.316	0.000	-0.232	NA	NA	NA	NA	NA	NA
40	43.743	0.173	0.000	-0.111	NA	NA	NA	NA	NA	NA
41	13.784	0.060	0.130	0.013	NA	NA	NA	NA	NA	NA
42	31.476	0.153	0.000	-0.097	NA	NA	NA	NA	NA	NA
43	39.903	0.164	0.000	-0.077	NA	NA	NA	NA	NA	NA
44	31.360	0.153	0.000	-0.047	NA	NA	NA	NA	NA	NA

Note. Results retrieved from Imputation 1 analysis. See Github repository for full results. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$, LH = Loevinger's *H*, Stan. *B* = Standarized Beta

Appendix Q Working Memory Pooled Item Parameters

Item	*a*	*d*	*g*	*u*
11	1	1.989013	0	1
12	1	2.434985	0	1
13	1	2.434985	0	1
14	1	0.684649	0	1
15	1	2.167124	0	1
16	1	1.673237	0	1
17	1	1.530314	0	1
18	1	0.577008	0	1
19	1	0.541552	0	1
20	1	0.471194	0	1
21	1	0.229484	0	1
22	1	0.127372	0	1
23	1	-0.14339	0	1
24	1	0.98374	0	1
25	1	-0.62541	0	1
26	1	-0.6966	0	1
27	1	-0.59012	0	1
28	1	-0.73252	0	1
29	1	-0.91612	0	1
30	1	-0.76869	0	1
31	1	-1.92624	0	1
33	1	-1.52973	0	1
34	1	-1.10817	0	1
35	1	-1.72026	0	1
36	1	-2.42309	0	1

Note. These parameters are generated by MIRT (i.e. are the equivalent slope-intercept translation of traditional/classical IRT parameters)