

2022

When are we guessing? An investigation of the impact of guessing on the validity of results in the assessment of students in large-scale assessment programs

Christopher Roy Freeman

Follow this and additional works at: <https://ro.uow.edu.au/theses1>

University of Wollongong

Copyright Warning

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site.

You are reminded of the following: This work is copyright. Apart from any use permitted under the Copyright Act 1968, no part of this work may be reproduced by any process, nor may any other exclusive right be exercised, without the permission of the author. Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material.

Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

Unless otherwise indicated, the views expressed in this thesis are those of the author and do not necessarily represent the views of the University of Wollongong.

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

When are we guessing?

**An investigation of the impact of guessing on the validity of results in
the assessment of students in large-scale assessment programs**

Christopher Roy Freeman

Supervisors:

A/Professor Steven Howard

Professor Jim Tognolini

This thesis is presented as part of the requirement for the conferral of the degree
Doctor of Philosophy

This research has been conducted with the support of
an Australian Government Research Training Program Scholarship

University of Wollongong
School of Education

June
2022

Abstract

The 21st century shift towards economic rationalism has included a trend towards measuring the outcomes of education and, equipped with these data, placing a higher priority on accountability in all sectors of the educational community. Corresponding to this shift, policy makers and government officials continue to demand better quality information on which to make data-informed decisions and allocate resources. To measure educational outcomes, for instance, stakeholders routinely look to national and international large-scale testing programs for indicators of success in achieving national goals and to identify areas for systemic attention.

A majority of these large-scale assessments use multiple-choice item types to assess student ability for the reporting of the outcomes of student achievement against curriculums and national or international standards. Students are encouraged to attempt all questions in these assessments. By doing so it is likely that students will guess the answers to those items where the concepts assessed are beyond their capabilities and be rewarded with an incremental increase in their results. This has the potential to threaten the validity of students' overall results, due to the chance of correctly guessed responses inflating assessment results beyond students' actual ability levels.

This PhD research investigated the potential impact of – and proposed a statistical solution for – guessing in large-scale education assessments. Specifically, it examined the outcomes from a novel application of the Rasch analysis technique, which quantified and adjusted for the measurement error associated with guessing and, consequently, increased the validity of the outcomes of the assessments.

The initial study to evaluate this potential extension and application of the Rasch model developed simulated data with specific responses defined as guesses. These data were used to develop a protocol that investigated item parameters that consistently identified the defined guesses. These parameters were termed the Guessing Indication Protocol (GIP). Next, GIP was applied to a small-scale field study and several sets of existing large-scale assessment data to ascertain the efficacy of the procedure. Application of GIP with the large-scale data revealed considerable changes in student ability estimates in the upper and lower regions of achievement, compared to the Rasch calibrations that did not attempt to take account of guessing in the data. Without GIP, there was a consistent trend for higher-ability students to have their ability underestimated and the lower-ability students to have their ability overestimated. The recalibration of item estimates with the GIP-conditioned data provided improved fit of the data to the Rasch model and caused significant changes in the student ability estimates, scaled scores, and achievement levels relative to initial estimates that ignored guessing. The implementation of the GIP process caused a significant proportion of students to have their assigned achievement level reclassified.

These findings suggest that the GIP procedure should be considered as a process for generating results from large-scale assessments, in which multiple-choice items are analysed using the Rasch Model, to improve the accuracy, and integrity of student results and the validity of actions and decisions based on the data.

Acknowledgments

I am greatly indebted to my friend and mentor Professor James (Jim) Tognolini for his unflagging, invaluable guidance and support throughout this entire process. He continually challenged me to expand my horizon and grow as a researcher and consider deeper aspects of the study. The completion of this study would not have been possible without his patience and dedication. My appreciation also goes to Associate Professor Steven Howard for his encouragement, advice, unwavering support and assistance given, particularly in navigating the process of completing this research.

Special thanks goes to Frances Eveleigh for her advice on many aspects of the study including her dedication to revising and editing drafts, and her many and varied ideas and contributions to the methodologies and analyses employed.

I would also like to thank and acknowledge Professor Massoud Al Badri of the Abu Dhabi Education Council who provided support for the study and authorised access to the de-identified data used to examine the impact of the developed protocol to identify probable guessing in a large-scale authentic environment.

I also owe a debt of gratitude to the principals and supporting teachers at the following Australian schools who generously gave of their time and resources in the collection of the field data:

Mary Immaculate Catholic Primary School, Bossley Park,
Reddam House, Woollahra,
St Francis de Sales College, Nowra,
St George Christian School, Hurstville,
St Michael's Catholic Primary School, Daceyville, and
St Patrick's College, Sutherland.

I acknowledge and am very thankful for the editorial assistance of professional editor Dr Terry Fitzgerald, whose academic area is education.

I owe sincere gratitude to Chris Giles who used his considerable resources and networks to liaise with and arrange the sample schools and continually encourage me to complete the dissertation.

Finally, my thanks to my family, colleagues and friends who have supported and maintained me over the course of this study.

Certification

I, Christopher Roy Freeman declare that this thesis submitted in fulfilment of the requirements for the conferral of the degree Doctor of Philosophy from the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Christopher Roy Freeman

24 June 2022

List of Names or Abbreviations

ACARA:	Australian Curriculum, Assessment and Reporting Authority
CR:	Constructed response
CTT:	Classical Test Theory
GIP:	Guessing Indication Protocol
Gonski 2.0:	Review to Achieve Educational Excellence in Australian Schools – ‘Through Growth to Achievement’
IEA:	International Association for the Evaluation of Educational Achievement
IRT:	Item Response Theory
MC:	Multiple Choice
MCEETYA:	Ministerial Council for Education, Employment Training and Youth Affairs
MTT:	Modern Test Theory
NAP:	National Assessment Program
NAPLAN:	National Assessment Program Literacy and Numeracy
OECD:	Organisation for Economic Co-operation and Development
PISA:	Programme for International Student Assessment
PIRLS:	Progress in International Reading Literacy Study
RM:	Rasch Model
RMT:	Rasch Model Theory
TIMSS:	Trends in International Mathematics and Science Study

Reader's Guide

Ability estimate: The location of students on the variable of interest inferred by the observations on a collection of data.

Calibration: The procedure of estimating item difficulty and/or student ability by converting raw scores to logit values on a measurement scale.

Construct: A single latent trait, characteristics, attribute or dimension assumed to be underpinning a set of items.

Deterministic: Characteristic of a model that implies the exact prediction of an outcome.

Dichotomous: Dichotomous data have only two possible outcomes: correct (1) or incorrect (0).

Estimation: The Rasch process of using obtained raw scores to calculate the probable values of student parameters and item parameters.

Fit: The degree of match between the pattern of observed responses and the expectation of the model.

Fit Statistics: Indices that estimate the extent to which responses adhere to the expectations of the model.

Item Difficulty: An estimate of an item's underlying difficulty calculated from the total number of successful responses to the item. Item difficulty is expressed in logits (also known as location or delta).

Item Facility: Observed number of correct responses compared to the number of attempted responses. Facility is expressed as a percentage.

Latent Trait: A characteristic or attribute of a student that can be inferred from the observation of the student's behaviours.

Logit: The unit of measurement that results when the Rasch model is used to transform raw scores obtained from the original student responses to log odds ratios on an interval scale.

Missing data: An item that the student does not answer or is omitted as a result of not reaching the item in the test sequence or an item omitted by some process.

Residual: A value that represents the difference between the Rasch model's theoretical expectation and the actual observed result.

Standard Error of Measure (SEm): A measure of how much measured test scores are spread around a "true" score.

Student ability: An estimate of a student's underlying capability on a trait as measured by the performance on a set of items that measure the trait.

Targeted: The extent to which items on the test match the range of students' abilities on the trait being measured.

Unidimensionality: A single attribute being tested that represents a hierarchical continuum.

Variable: An attribute of the trait of interest that can be assigned a variety of magnitudes. A variable will reflect the estimate of student ability in the latent trait examined by the construct.

Contents

When are we guessing?	1
Abstract.....	i
Acknowledgments.....	i
Certification	2
List of Names or Abbreviations	3
Reader’s Guide.....	4
Chapter 1 Background to the Study	18
1.1 Introduction.....	18
1.1.1 Educational Environment – Curricula Structures and Assessment Constructs	19
1.2 Issues Threatening Validity in Multiple-Choice Items	20
1.3 The Development of Assessment Models and the Evolution of Educational Measurement	21
1.3.1 Classical Test Theory (CTT).....	21
1.3.2 Modern Test Theory and the Concept of an Educational Scale of Proficiency	22
1.4 The Problem: The Impact of Guessing in Multiple Choice Items.....	24
1.5 The Current Study.....	25
1.6 Organisation of the Thesis	25
Chapter 2 Issues in Measuring Student Ability and Achievement.....	27
2.1 Introduction.....	27
2.2 Measurement and Scales.....	27
2.2.1 Measurement in the Physical Sciences	27
2.2.2 Measurement in the Social Sciences	28
2.3 Measurement in Student Achievement Tests	29
2.3.1 Measurement Scales in Education.....	29
2.3.2 Issues in Measurement With Multiple Choice Items	30
2.4 Early Research and Traditional Strategies to Address Guessing in MC Items	31
2.4.1 Strategies to Discourage Guessing.....	31
2.5 Psychometric Approaches to Account for Guessing.....	34
2.5.1 Introduction to Modern Analysis Techniques to Account for Guessing.....	34
2.5.2 Previous Approaches to Account for Guessing in the IRT Paradigm.....	34
2.6 Recent Approaches to Account for Guessing Within the Rasch Model Paradigm	38

2.6.1	The Approach of Andrich et al. (2012, 2015)	38
2.6.2	Comments on the Approach of Andrich et al. (2012, 2015)	39
2.7	A Proposal to Account for Guessing in the Rasch Model	40
2.8	Summary	41
Chapter 3 Background of Modern Test Theory Applied to Student Achievement		42
3.1	Introduction	42
3.1.1	Relationship Between Student Ability and Achievement	42
3.1.2	Early Estimation of Student Ability	42
3.2	Development of Modern Test Theory	44
3.2.1	Estimation of Educational Achievement	44
3.2.2	The Two Parameter Model (2PL Model)	44
3.2.3	The Three Parameter Model (3PL Model)	45
3.3	The Rasch Model	50
3.3.1	Introduction	50
3.3.2	Features of the Rasch Model	51
3.3.3	Measurement Principles Underpinning the Rasch Model	56
3.4	Summary	59
Chapter 4 Methodology Overview		60
4.1	Introduction	60
4.1.1	Overview of Research Design and Phases	60
4.2	GIP Principles	62
4.3	Analysis Methodology Overview	62
4.4	Participants	63
4.4.1	Study 1: Simulated Data	63
4.4.2	Study 2: Small-Scale Field Data	63
4.4.3	Study 3: A Large-Scale, System-Wide Sample	63
4.5	Data Sources	64
4.5.1	Study 1: Simulated Data Generation	64
4.5.2	Study 2: Small-Scale Sample Data Collection	64
4.5.3	Study 3: Large-Scale Cohort Data Collection	65
4.6	Procedures	66
4.6.1	Study 1: Simulated Data	66

4.6.2	Study 2: Small-Scale Field Study Data.....	66
4.6.3	Study 3: Large-Scale Study Data.....	66
4.7	Plan for Analysis.....	67
4.7.1	Study 1: Simulated Data.....	67
4.7.2	Study 2: Field study.....	68
4.7.3	Study 3: Investigations of a Large-Scale Data Set.....	69
4.7.4	Investigation of Student Response Time as a Parameter in the Identification of Guessing	69
4.8	Summary.....	70
	Chapter 5 Study 1: Analysis of the Simulated Data	71
5.1	Introduction to the Chapter	71
5.2	The Guttman Structure as a Starting Point for Investigations	71
5.2.1	Introduction to the Section.....	71
5.2.2	The Guttman Scale.....	72
5.3	Elaboration of the Methods.....	76
5.3.1	Development of Simulated Data Algorithm.....	76
5.3.2	Plan for Analysis (PfA) of the Simulated Data.....	79
5.4	Results for the Simulated Data.....	80
5.4.1	PfA Step 1: Preliminary Results of Observed Item Performance in the Simulations	80
5.5	Rasch Analyses to Inform the Guessing Indication Protocol (GIP).....	85
5.5.1	PfA Step 2: Introduction	85
5.5.2	Rasch Analyses Conducted on Simulated Data	85
5.5.3	Simulation Item/Student Maps.....	86
5.6	Initial Observations and Iteration of a Guessing Indication Protocol (GIP)	91
5.6.1	Initial Determination of the GIP Parameters.....	91
5.6.1	PfA Step 3: Initial Development of the Guessing Indication Protocol (GIP)	93
5.6.2	Identification of Guesses Using Critical Indices.....	93
5.7	Preliminary Analyses and Observations of the Initial Defined Protocol.....	96
5.7.1	Initial Analysis Results	96
5.7.2	Observations Regarding the Initial Analysis Results	100
5.8	A Second Iteration of the Protocol: Adjusting the p Value.....	105
5.8.1	Modification of the Protocol Parameters	105
5.8.2	Result of the Modification of the p Value With the Protocol Parameters.....	106

5.9	Implementation of Phase 3 of the Plan for Analysis	111
5.9.1	The Phase 3, GIP Process	111
5.9.2	PfA Step 3: Comparison of Rasch Analysis Results: INIT vs GIP3A and GIP3B	111
5.10	Results of the Implementation of Phase 3 of the Plan for Analysis	112
5.10.2	Comparison of Rasch Results for the Four Analysis Phases.....	118
5.11	Analysis of Fit Statistics.....	124
5.12	Summary	128
Chapter 6 Study 2: Investigating the Proposed GIP Process Using Small-Scale Field Data		129
6.1	Introduction.....	129
6.2	Plan for Analysis (PfA) for the English Versions of the Tests.....	129
6.3	Study 2, Stage 1, English Versions of the Tests.....	130
6.3.1	PfA Step 1 Analysis – Investigation of Self-Identified Guesses	130
6.3.2	PfA Step 2 and Step 3 Rasch Analysis – English Version Test Data.....	135
6.3.3	PfA Step 3 Investigations of the Field Data	141
6.3.4	GIP Indication Rates by Quartile	146
6.3.5	Analysis of Fit.....	146
6.4	Guttman Analysis of GIP-Indicated Items	147
6.5	Discussion – Implications of the Outcomes of the English Tests	148
6.5.1	The SIG Outcomes.....	148
6.5.2	The GIP outcomes.....	150
6.6	Study 2, Stage 2, Arabic Versions of the Tests	151
6.6.1	Introduction: The Arabic Versions Study	151
6.7	PfA of the Arabic Versions	151
6.7.1	Comparisons of English Version and Arabic Version of Selected Items.....	151
6.8	PfA Step 2, Rasch Analyses of the Arabic Response Data	156
6.8.1	Year 5 Item Statistics	156
6.8.2	Year 5 Student Ability Statistics	157
6.9	Discussion and Implications of the Arabic Test Outcomes.....	162
6.10	Summary	162

Chapter 7 Study 3: Results of the Application of the Proposed GIP Model with Large-Scale Authentic Data.....	163
7.1 Introduction.....	163
7.2 Plan for Analysis (PfA).....	163
7.1.1 Application of the GIP Procedure.....	164
7.3 PfA Step 1, Rasch Analysis of the Large-Scale data.....	165
7.3.1 Summary Item Statistics.....	165
7.3.2 Summary Student Ability Statistics.....	166
7.3.3 Standard Errors.....	167
7.3.4 Reliability Indices.....	168
7.4 Display of Distributions – Student/Item Maps for Each Test.....	168
7.4.1 Graphic Representation of the Rasch Outcomes.....	168
7.5 Shifts in Distributions of the GIP Analyses Compared to the INIT Analysis.....	173
7.5.1 Item Distributions.....	173
7.7 Summary of Indicated Guesses of the GIP Procedure.....	180
7.7.1 Grade 4 Mathematics.....	180
7.8 PfA Step 2, Supplementary Analyses.....	186
7.8.1 Analysis of Fit.....	186
7.8.2 Statistical Significance of the GIP Interventions.....	186
7.8.3 Effect Size.....	188
7.8.4 Cohen's d.....	188
7.9 Discussion.....	190
7.9.1 Significance of Test Targeting and Item Difficulty Distributions.....	190
7.10.2 Significance of Accounting for Guessing in These Data.....	191
7.11 Summary.....	191
Chapter 8 Response Time as a Factor in Indicating Guessing.....	193
8.1 Introduction.....	193
8.2 Plan for Analysis (PfA).....	193
8.3 PfA Step 1 Results: Interaction of Response Time and Raw Scores.....	194
8.4 PfA Step 2 Results: Relationship Between Item Location and Response Time by Ability Group	196
8.5 Investigation of Response Time as a Unique Indicative Guessing Parameter.....	198
8.5.1 Frequency of Rapid Responses by Criterion Investigated.....	198

8.5.2	Consideration of the Relationship Between Item Difficulty and Rapid Responses	199
8.5.3	Consideration of the Relationship Between Rapid Responses and GIP Indicated Items	201
8.6	Results of Analyses of Student Response Times	203
8.7	Discussion of the Outcomes of PFA Step 1 and Step 2	203
8.8	Conclusion Regarding the Student/Item Response Time as an Additional Parameter in the GIP Algorithm	204
	Chapter 9 Reports of Student Achievement	205
	Impact of the Guessing Indication Protocol (GIP) on Reported Student Performance	205
9.1	Introduction.....	205
9.1.1	Achievement Standards in Australia.....	205
9.1.2	Reported Levels and the Link to Achievement Standards	206
9.1.3	Scaled Scores and the Link to Levels	209
9.2	Study 1, Simulated Data – Reporting of Achievement Levels for Each Analysis Phase	211
9.2.1	Reported Achievement Levels for the SIM1 Data.....	211
9.2.2	Discussion – Achievement Levels for SIM1 Simulated Data.....	212
9.2.3	Reported Achievement Levels for the Simulated Data Sets	213
9.2.4	Simulation Analyses – Scaled Scores	215
9.3	Study 2, Field Trial Data.....	218
9.3.1	Year 5.....	219
9.3.2	Year 7.....	219
9.4	Study 3, Large-Scale Data Sets.....	220
9.4.1	Reported Levels for Large-Scale Data.....	220
9.4.2	Scaled Score Comparisons of the Large-Scale Data sets	222
9.5	Scale of Reclassifications.....	226
9.5.1	Degree of Reclassification of Reported Outcomes in Simulated Data.....	226
9.5.2	Degree of Reclassification of Levels in Authentic Data	229
9.6	Review of Findings.....	231
9.6.1	Study 1, Simulated Data.....	231
9.6.2	Study 2, Small-Scale Sample	232
9.6.3	Study 3, Large-Scale Outcomes	233
9.7	Summary	234

Chapter 10 Discussion: Limitations and Merits of the GIP	235
10.1 Introduction – Limiting Factors	235
10.2 Reflections on the Circular Constraints in Identifying Guessing in Student Response Data	235
10.2.1 Calculation of Item Difficulty	235
10.2.2 Distribution of Item Difficulties	237
10.2.3 Estimation of Student Ability	237
10.2.4 Calculation of Item/Student Probability of a Correct Response.	238
10.2.5 Calculation of Item/Student Residual	238
10.3 Merits of the GIP: Student Scaled Scores and Reported Levels	239
10.3.1 Summary Observation in Simulated Data Sets	239
10.3.2 Large-Scale Data Sets	240
10.4 Overall Comments	246
10.4.1 Degree of Reclassification.....	247
10.5 Summary	248
Chapter 11 Conclusion	249
11.1 Introduction to the Chapter	249
11.2 Summary of the Research	249
11.3 Summary of the Key Results.....	250
11.4 Implications of the study	250
11.5 Limitations of the Study	252
11.6 Recommendations for Further Research	252
11.7 Next Steps	253
11.8 Conclusion to the Chapter	254
List of References	255
Appendices	
Appendix A Detail of the Data Construct of Study 1: the Simulated Data	
Appendix B Example of GIP Calculations for SIM1	
Appendix C Analyses of Small-Scale Field Study Data	
Appendix D Investigation of Alternative p Values	

List of Figures

Figure 1.1 <i>Evidence of Achievement in Australian Schools – NAPLAN</i>	23
Figure 3.1 <i>Example of an Item Characteristic Curve</i>	57
Figure 3.2 <i>Example of Item Characteristic Curve for a 3-Item Test–Equal Discrimination</i>	57
Figure 3.3 <i>Example of Item Characteristic Curve for a 3-Item Test–Unequal Discrimination</i>	58
Figure 5.1 <i>Impact of NOT Accounting for Errors on the Distribution of a Set of Raw Scores</i>	72
Figure 5.2 <i>SIM1 Comparison of Facility With Correct Guesses</i>	80
Figure 5.3 <i>SIM2 Comparison of Facility With Correct Random Guesses</i>	81
Figure 5.4 <i>SIM3 Comparison of Facility With Correct Random Guesses</i>	82
Figure 5.5 <i>SIM4 Comparison of Facility With Correct Random Guesses</i>	83
Figure 5.6 <i>SIM5 Comparison of Facility With Correct Random Guesses</i>	84
Figure 5.7 <i>SIM1 Item/Student Map for INIT Analysis and GS Analysis</i>	88
Figure 5.8 <i>SIM3 Item/Student Map for INIT Analysis and GS Analysis</i>	89
Figure 5.9 <i>SIM1 Comparison of Item Student Maps for INIT Analysis, GS Analysis, and GIP3A Analysis</i>	99
Figure 5.10 <i>SIM1 Comparison of Item Locations INIT Analysis, GS Analysis and GIP3A $p=0.5$ Analysis</i>	101
Figure 5.11 <i>SIM1 Comparison of Count of Defined Guesses Recovered From GS and GIP3A $p=0.5$ Analysis</i>	101
Figure 5.12 <i>SIM3 Comparison of Item Student Maps for INIT Analysis, GS Analysis, and GIP $p=0.5$ Analysis</i>	103
Figure 5.13 <i>SIM3 Comparison of Item Locations – INIT Analysis, GS Analysis, and GIP $p=0.5$ Analysis</i>	104
Figure 5.14 <i>SIM3 Comparison of Count of Defined Guesses Recovered From the GS and GIP3A $p=0.5$ Analyses</i>	105
Figure 5.15 <i>SIM1 Comparison of Item/Student Maps for INIT Analysis, GIP3A Analysis, and GIP3B Analysis</i>	113
Figure 5.16 <i>SIM2 Comparison of Item/Student Maps for INIT Analysis, GIP 3A Analysis, and GIP 3B Analysis</i>	114
Figure 5.17 <i>SIM3 Comparison of Item/Student Maps for INIT Analysis, GIP Analysis, and GIPINIT Analysis</i>	115
Figure 5.18 <i>SIM4 Comparison of Item/Student Maps for INIT Analysis, GIP Analysis, and GIPINIT Analysis</i>	116
Figure 5.19 <i>SIM5 Comparison of Item/Student Maps for INIT Analysis, GIP Analysis, and GIPINIT Analysis</i>	117
Figure 6.1 <i>Year 5 Relationship Between Item Facility and the Percentage of Self-Identified Guesses</i> ...	131
Figure 6.2 <i>Year 5 Relationship Between the Percent of Self-Identified Guesses and the Percent of Correct Guesses</i>	131
Figure 6.3 <i>Year 7 Relationship between item Facility and the percentage of Self-Identified Guesses</i>	132

Figure 6.4 <i>Year 7 Relationship Between the Percentage of Self-Identified Guesses and the Percentage of Correct Guesses</i>	133
Figure 6.5 <i>Year 5 Relationship Between INIT Item Location and the Proportion of Self-Identified Guesses Irrespective of Result</i>	134
Figure 6.6 <i>Year 7 Relationship Between INIT Item Location and the Proportion of Self-Identified Guesses Irrespective of Result</i>	134
Figure 6.7 <i>Field Trial Data Item Student Maps – Year 5 English Version Analyses</i>	144
Figure 6.8 <i>Field Trial Data Item-Student Maps – Year 7 English Version Analyses</i>	145
Figure 6.9 <i>Year 5 Math5Q03</i>	152
Figure 6.10 <i>Year 5 Math5Q08</i>	153
Figure 6.11 <i>Math7Q12</i>	154
Figure 6.12 <i>Math7Q01– Arabic Version</i>	154
Figure 6.13 <i>Item/Student Maps for Year 5 Arabic INIT and SIG Analyses</i>	158
Figure 6.14 <i>Item-Student Maps for Each of the Analysis Phases Performed on the Year 7 Data</i>	161
Figure 7.1 <i>Comparison of Analysis Distributions – Grade 4 Mathematics INIT vs GIP Conditioned Data</i>	170
Figure 7.2 <i>Comparison of Analysis Distributions – Grade 8 Mathematics INIT vs GIP Conditioned Data</i>	171
Figure 7.3 <i>Comparison of Analysis Distributions – Grade 4 Science INIT vs GIP Conditioned Data</i>	172
Figure 7.4 <i>Grade 4 Mathematics Break-Even Mean Ability Estimates by Decile</i>	176
Figure 7.5 <i>Grade 8 Mathematics Break-Even Mean Ability Estimates by Decile</i>	178
Figure 7.6 <i>Grade 4 Science Break-Even Mean Ability Estimates by Decile</i>	179
Figure 9.1	206
Figure 9.2 <i>An Example of a Mathematics Standards Framework (Grade 3 to Grade 6)</i>	207
Figure 9.3 <i>NAPLAN Reported Bands (Reading) and Scaled Score Cut Scores</i>	209
Figure 9.4 <i>Comparison of Grade 4 Mathematics Scaled Scores by Decile</i>	223
Figure 9.5 <i>Comparison of Grade 8 Mathematics Scaled Scores by Decile</i>	224
Figure 9.6 <i>Comparison of Grade 4 Science Scaled Scores by Decile</i>	225

List of Tables

Table 2.1 <i>Vertical Equating Design for the 2013 NAPLAN Reading Scale</i>	38
Table 2.2 <i>Methodology Design for the Overall Research</i>	40
Table 4.1 <i>Summary of the Analysis Phases of This Research</i>	60
Table 4.2 <i>Summary of the Data Investigated and the Analysis Phases Performed on Each Study</i>	61
Table 5.1 <i>Guttman Scale Responses for a Three-Item Test</i>	74
Table 5.2 <i>Non-Guttman-Like Responses for Three-Item Test</i>	74
Table 5.3 <i>An Example of the Guttman Scale: Six items, Seven Test-Takers</i>	77
Table 5.4 <i>An Abridged Example of the Structure of the Simulated Data</i>	78
Table 5.5 <i>Summary of Intended Targeting and Observed Traditional Statistics for Each Data Set</i>	78
Table 5.6 <i>Summary Results of Simulated Data Sets – INIT ANALYSIS – Defined Guesses (1)</i>	90
Table 5.7 <i>Summary Results of Simulated Data Sets – GS ANALYSIS – Guesses Scored Missing (9)</i>	90
Table 5.8 <i>Elaboration of the Composite Parameter Table (Abridged SIM1 - 16 of the 40 Items)</i>	95
Table 5.9 <i>Summary of the GIP Recovery Rates by Simulation $p = 0.5$</i>	97
Table 5.10 <i>Elaboration of the GIP Recovery Rates by Simulation $p=0.5$</i>	98
Table 5.11 <i>Comparison of Guess Recovery Rates $p = 0.5$ and $p = 0.6$</i>	107
Table 5.12 <i>Elaboration of the Comparison Between Defined Guess and $GIP_{p=0.6}$ by Quartile/Decile</i> ...	108
Table 5.13 <i>SIM1 Extract of GIP Indicated Guesses by Item Location $p = 0.6$</i>	110
Table 5.14 <i>SIM3 Extract of GIP Indicated Guesses by Item Location $p = 0.6$</i>	110
Table 5.15 <i>Comparison of SIM1 Item Statistics for the Four Analyses</i>	118
Table 5.16 <i>Comparison of SIM2 Item Statistics for the Four Analyses</i>	119
Table 5.17 <i>Comparison of SIM3 Item Statistics for the Four Analyses</i>	119
Table 5.18 <i>Comparison of SIM4 Item Statistics for the Four Analyses</i>	120
Table 5.19 <i>Comparison of SIM5 Item Statistics for the Four Analyses</i>	120
Table 5.20 <i>Comparison of SIM1 Student Ability Estimate Statistics for the Four Analyses</i>	122
Table 5.21 <i>Comparison of SIM2 Student Statistics for the Four Analyses</i>	122
Table 5.22 <i>Comparison of SIM3 Student Statistics for the Four Analyses</i>	123
Table 5.23 <i>Comparison of SIM4 Student Statistics for the Four Analyses</i>	123
Table 5.24 <i>Comparison of SIM5 Student Statistics for the Four Analyses</i>	124
Table 5.25 <i>Test of Fit for the INIT, GS, and GIP Analyses</i>	125
Table 5.26 <i>Summary of the Results of the INIT Simulated Data Sets</i>	126
Table 5.27 <i>Summary of the Results of the $GIP_{INIT3B_{p=0.6}}$ Simulated Data GIP_{3A} Locations with INIT Raw Scores</i>	126
Table 6.1 <i>Response Pattern for Lower Ability Year 5 Students Ordered by Item Difficulty ($p = 0.6$)</i>	137
Table 6.2 <i>Response Pattern for Mid-Tange Ability Year 5 Students Ordered by Item Difficulty</i>	138
Table 6.3 <i>Response Pattern for Lower Ability Year 7 Students Ordered by Item Difficulty Including Response Mode (G, Y, or P) ($p = 0.6$)</i>	140
Table 6.4 <i>Year 5 Summary Analysis – Student Ability Estimates by Analysis Phase</i>	142
Table 6.5 <i>Year 7 Summary Analysis – Student Ability Estimates by Analysis Phase</i>	142
Table 6.6 <i>Year 5 Item Summary Statistics by Analysis Phase</i>	143

Table 6.7 <i>Year 7 Item Summary Statistics by Analysis Phase</i>	143
Table 6.8 <i>Proportion of Responses Indicated as a Probable Guess by the GIP3A Procedure by Ability Group</i>	146
Table 6.9 <i>Comparison of Mean Square Statistics – INIT and GIP Analyses for Year 5 and Year 7</i>	147
Table 6.10 <i>Proportions of GIP-Identified Items by Year and Ability Quartile</i>	147
Table 6.11 <i>Year 5 Number of GIP-Identified Items by Item Location</i>	149
Table 6.12 <i>Year 7 Number of GIP-Identified Items by Item Location</i>	149
Table 6.13 <i>Year 5 and Year 7 Proportion of GIP identified items by INIT ability quartile</i>	155
Table 6.14 <i>Year 5 Distribution of GIP-Identified Items by Quartile and Item Location</i>	155
Table 6.15 <i>Year 7 Distribution of GIP-Identified Items by Quartile and Item Location</i>	156
Table 6.16 <i>Year 5 Arabic INIT, SIG, GIP, and GIPINIT Summary of Analyses – Items</i>	156
Table 6.17 <i>Year 5 Arabic INIT, SIG, GIP, and GIPINIT Summary of Analyses – Students</i>	157
Table 6.18 <i>Year 7 Arabic Summary Analysis for Items by Analysis Phase</i>	159
Table 6.19 <i>Year 7 Arabic Summary Analysis for Students by Analysis Phase</i>	160
Table 7.1 <i>Comparison of Rasch Analysis G4 Mathematics – Item Parameters INIT, GIP, and GIPINIT</i>	165
Table 7.2 <i>Comparison of Rasch Analysis G4 Science – Item Parameters INIT, GIP, and GIPINIT</i>	165
Table 7.3 <i>Comparison of Rasch Analysis G8 Mathematics – Item Parameters INIT, GIP, and GIPINIT</i>	166
Table 7.4 <i>Comparison of Rasch Analysis G4 Mathematics Ability Estimates – INIT, GIP, and GIPINIT</i>	166
Table 7.5 <i>Comparison of Rasch Analysis G8 Mathematics Ability Estimates – INIT, GIP, and GIPINIT</i>	167
Table 7.6 <i>Comparison of Rasch Analysis G4 Science Ability Estimates – INIT, GIP and GIPINIT</i>	167
Table 7.7 <i>Comparison of Reliability Indices for Each Analysis INIT vs GIP Conditioned Data Phases</i>	168
Table 7.8 <i>Item Distribution Statistics for Each Analysis</i>	173
Table 7.9 <i>Student Ability Estimate Distribution Statistics for Each Analysis</i>	174
Table 7.10 <i>Grade 4 Mathematics Comparison of Mean Ability Estimate by Decile</i>	175
Table 7.11 <i>Grade 4 Mathematics Comparison of Mean Ability Estimates by Ability Groupings</i>	176
Table 7.12 <i>Grade 8 Mathematics Comparison of Ability Estimate by Decile</i>	177
Table 7.13 <i>Grade 8 Mathematics Comparison of Mean Ability Estimates by Ability Groupings</i>	178
Table 7.14 <i>Grade 4 Science Comparison of Mean Ability Estimates by Decile</i>	179
Table 7.15 <i>Grade 4 Science Comparison of Mean Ability Estimates by Ability Groupings</i>	179
Table 7.16 <i>Comparison of Number of Items Re-Coded by GIP Procedure by Ability Group – Grade 4 Mathematics</i>	180
Table 7.17 <i>Grade 4 Mathematics Count of GIP Implementations by Item Difficulty (Logits)</i>	182
Table 7.18 <i>Comparison of Number of Items Re-coded by GIP Procedure by Ability Group – Grade 8 Mathematics</i>	183
Table 7.19 <i>Grade 8 Mathematics Count of GIP Implementations by Item Difficulty</i>	183
Table 7.20 <i>Comparison of Number of Items Re-coded by GIP Procedure by Ability Group – Grade 4 Science</i>	184

Table 7.21 <i>Grade 4 Science Count of GIP Implementations by Item Difficulty</i>	185
Table 7.22 <i>Comparison of Reliability Indices for Each Rasch Analysis INIT, GIP, and GIPINIT</i>	186
Table 7.23 <i>Comparison of Means – Grade 4 Mathematics Student INIT Ability Estimate and GIP and GIPINIT Ability Estimates</i>	187
Table 7.24 <i>Comparison of Means – Grade 8 Mathematics Student INIT Ability Estimate With GIP and GIPINIT Ability Estimates</i>	187
Table 7.25 <i>Comparison of Means – Grade 4 Science Student INIT Ability Estimate With GIP and GIPINIT Ability Estimates</i>	188
Table 7.26 <i>Cohen’s Interpretation of Effect Size Statistic</i>	189
Table 8.1 <i>Year 4 Mathematics Comparison Between Observed Raw Score and Time Taken on the Test</i>	195
Table 8.2 <i>Year 8 Mathematics Comparison Between Observed Raw Score and Time Taken on the Test</i>	195
Table 8.3 <i>Year 4 Science Comparison Between Observed Raw Score and Time Taken on the Test</i>	195
Table 8.4 <i>Average Time Taken per MC Item by Ability Groups for Grades 4 and 8 Mathematics and Grade 4 Science Students</i>	197
Table 8.5 <i>Count of Rapid Responses Potentially Re-Coded as Guesses by the Defined Time Constraint</i>	198
Table 8.6 <i>Grade 4 Mathematics Rapid Response Proportions by Decile</i>	200
Table 8.7 <i>Extract of Response Time (T/2), Response, and GIP-Indicated Items for a Random Selection of Students</i>	202
Table 9.1 <i>Parameters Determined for Calculation of INIT Standardised Scaled Scores</i>	210
Table 9.2 <i>Comparisons of Percentages in Each Level for Simulation 1 by Analysis Phase</i>	212
Table 9.3 <i>Comparisons of Percentages in Each Level by Analysis Phase</i>	213
Table 9.4 <i>Comparisons of Percentages in Each Level for Simulation 3 by Analysis Phase</i>	214
Table 9.5 <i>Comparisons of Percentages in Each Level for Simulation 4 – Mistargeted – Too Easy.</i>	214
Table 9.6 <i>Comparisons of Percentages in Each Level for Simulation 5 – Mistargeted – Too Hard</i>	215
Table 9.7 <i>Comparison of Mean Scaled Scores by Quartile for SIM1 Data</i>	215
Table 9.8 <i>Comparison of Mean Scaled Scores by Quartile for SIM2 Data</i>	217
Table 9.9 <i>Comparison of Mean Scaled Scores by Decile for SIM3 Data</i>	217
Table 9.10 <i>Comparison of Mean Scaled Scores by Quartile for SIM4 Data</i>	218
Table 9.11 <i>Comparison of Mean Scaled Scores by Quartile for SIM5 Data</i>	218
Table 9.12 <i>Comparisons of Percentages in Each Level for Field Trial Data – Y5 English Version Mathematics</i>	219
Table 9.13 <i>Comparisons of Percentages in Each Level for Field Trial Data – Y7 English Version Mathematics</i>	219
Table 9.14 <i>Comparisons of Percentages in Each Level for Large Scale Assessment – G4 Mathematics</i>	220
Table 9.15 <i>Comparisons of Percentages in Each Level for Large Scale Assessment – Grade 8 Mathematics</i>	221
Table 9.16 <i>Comparisons of Percentages in Each Level for Large-Scale Assessment – Grade 4 Science</i>	222
Table 9.17 <i>Differences Between INIT and GIP Standardised Scale Scores for Grade 4 Mathematics</i> .	223

Table 9.18	<i>Differences Between INIT and GIP Standardised Scale Scores for Grade 8 Mathematics ..</i>	224
Table 9.19	<i>Differences Between INIT and GIP Standardised Scale Scores for Grade 4 Science</i>	225
Table 9.20	<i>Discrepancy Between Reported Percentages in Levels – the Degree of Reclassification SIM1</i>	227
Table 9.21	<i>Discrepancy Between Reported Percentages in Levels – the Degree of Reclassification SIM2</i>	228
Table 9.22	<i>Discrepancy Between Reported Percentages in Levels – the Degree of Reclassification SIM3</i>	228
Table 9.23	<i>Discrepancy Between Reported Percentages in Levels – the Degree of Reclassification SIM4</i>	229
Table 9.24	<i>Discrepancy Between Reported Percentages in Levels – the Degree of Reclassification SIM5</i>	229
Table 9.25	<i>Comparison of Analysis Outcomes – Percentage in Levels – Grade 4 Mathematics</i>	230
Table 9.26	<i>Comparison of Analysis Outcomes – Percentage in Levels – Grade 8 Mathematics</i>	230
Table 9.27	<i>Comparison of Analysis Outcomes – Percentage in Levels – Grade 4 Science</i>	231
Table 9.28	<i>Summary of Differences in Percentages in Levels, Study 1, Simulations 1 to 5</i>	232
Table 9.29	<i>Summary of Differences in Percentages in Levels, Study 2, Field Trial Data Year 5 and Year 7.....</i>	233
Table 9.30	<i>Summary of Differences in Percentages in Levels, Study 3, Large-Scale Data</i>	233
Table 10.1	<i>Item Estimation – Conditional Probability of Dichotomously Scored Responses</i>	236
Table 10.2	<i>Proportion of GIP Identified Guesses by Ability Group Compared to Simulated Data INIT Analysis</i>	240
Table 10.3	<i>G4 Mathematics Comparison of Proportions in Levels by Scaled Score and Analysis Phase</i>	241
Table 10.4	<i>G4 Mathematics Comparison of Percentages in Levels</i>	242
Table 10.5	<i>G8 Mathematics Comparison of Percentages in Levels by Scaled Score and Analysis Phase</i>	243
Table 10.6	<i>G8 Mathematics Comparison of Percentages in Levels</i>	244
Table 10.7	<i>Grade 4 Science Comparison of Percentages in Levels by Scaled Score and Analysis Phase</i>	245
Table 10.8	<i>Grade 4 Science Comparison of Percentages in Levels</i>	246

Chapter 1

Background to the Study

1.1 Introduction

Since the 1980s there has been an increasing use of analysis methods grounded in modern measurement models to quantify and report student educational performance. In Australia, the analyses of student performance and the estimations of student ability in the major learning domains have used Rasch (1960, 1980) measurement models and their extensions (Andrich et al., 2010; Wright & Masters, 1982; Wu et al., 1998). Nationally, the aggregated outcomes of large-scale testing programs provide a baseline for measuring the achievement of educational outcomes. Internationally, assessments provide information regarding Australia’s educational standards compared to its overseas counterparts. The information derived from these national and international assessments drive Australian federal-, state-, and territory-specific policy decisions and the distribution of resources under the banner of “data-driven decision making”.

The guiding principles of the goals of the national education agenda in Australia were promulgated by the Ministerial Council for Education, Employment, Training and Youth Affairs (MCEETYA) through the *Adelaide Declaration on National Goals for Schooling in the Twenty-First Century* (1999), which included the aspiration of:

increasing public confidence in school education through explicit and defensible standards that guide improvement in students’ levels of educational achievement and through which the effectiveness, efficiency and equity of schooling can be measured and evaluated (p. 2).

These goals have been reinforced by the *Melbourne Declaration on Educational Goals for Young Australians* (2008), which highlighted the need for “reliable, rich data on the performance of their students because they have the primary accountability for improving student outcomes” (p. 16). A significant outcome of the Melbourne Declaration was the introduction of the National Assessment Program – Literacy and Numeracy (NAPLAN), which has provided a vehicle for the measurement and estimation of student ability in Numeracy and Literacy at Years 3, 5, 7 and 9. Within this program a national taskforce, with representation from all jurisdictions, undertakes statistical monitoring processes to provide data on the overall “health” of the national educational system and enable state and territory performances to be compared.

In 2010, the Australian Curriculum, Assessment and Reporting Authority (ACARA) was established to coordinate these national curriculum, assessment and reporting activities and provide annual feedback on the extent to which the state and territory jurisdictions have achieved the national goals articulated in the Declarations, as measured through the standardised testing and census-collection activities. As evidence of this focus on educational data enduring, the *Alice Springs (Mparntwe) Education Declaration* of 2020 has reinforced a commitment to high-quality assessment, with the declaration stating that the Commonwealth continues to “commit to ensuring that all education sectors deliver world-class curriculum and assessment in Australian schools” (p. 15).

A direct impact of these declarations in the national educational environment is that governments now require measures that indicate the impact of educational initiatives on learning outcomes. These initiatives include better articulation of what students “know, understand and can do” at various stages of their educational journey; quantifiable estimates of how systems are improving educational outcomes as a consequence of programs, initiatives and interventions; and a system of national and international “benchmarks” to empirically support the argument that Australia’s educational outcomes are comparable, or ideally better than, other countries around the world.

In practice, a pragmatic outcome of the declarations has been the increasing pervasiveness of the nationally implemented standardised educational assessments. All states and territories are now required to participate in cyclical implementations of the cohort-wide National Assessment Program (NAP) suite of assessments, as well as in international population sample programs, including the OECD assessment PISA (15-year-old students), the IEA’s PIRLS (Year 4), and TIMSS (at Years 4 and 8).

At state level the information from national and international assessments is used to monitor the progress of groups, regions, schools, and students over time. At the school level, national assessments provide an indicator for monitoring the impact of school-based interventions and improvement programs and an external reporting mechanism to stakeholders.

1.1.1 Educational Environment – Curricula Structures and Assessment Constructs

ACARA developed national curriculum documents that define the sequence of a student’s progress in terms of desired outcomes. The curriculum documents define (with flexibility to accommodate local situations and environments across Australia) content, skills, understandings, and standards in terms of the expected outcomes that should be demonstrated by students at different stages of their educational journey. These curriculum documents, and accompanying guidance and examples, provide foci for item writers involved in large-scale assessment programs to develop items and prepare tests to assess students’ performance. Professional item writers and test constructors engage with curriculum documents to produce a range of items that accord with the desired range of observable behaviours that may be expected from the target cohort of students. These processes mitigate threats to the validity of the test instruments (Messick, 1989b).

However, the legislative requirements that govern national testing programs have imposed several unintended constraints on item writers. For example, the interaction of legislative requirements, budgetary constraints, and the timeline demands for providing results to stakeholders have led to multiple-choice items being the dominant item type in large-scale testing programs. This is due to the relative ease and low marginal cost of scoring multiple-choice items through electronic data-capture and scanning processes. The high accuracy of data-capture processes contributes to the validity of the scoring. However, since the inception of multiple-choice items there has been contention over the validity of the students’ responses to them due to the chance of students obtaining inflated scores through random guessing.

At present both the NAPLAN suite of assessments and the international assessments referenced above rely heavily on multiple-choice items to assess student achievement against defined assessment frameworks. The advantages of multiple-choice items are well documented and include allowing a wide range of curriculum coverage; objective scoring; permitting a variety of cognitive domains and relative content item difficulties; allowing for a range of formats and stimuli; the potential for distractor analysis to inform misconceptions or misunderstandings; quick turnaround time for feedback; and the capacity for items to be auto-scored. A commonly cited disadvantage of multiple-choice items is the potential for students to guess a correct answer with no effort or knowledge of the skill underpinning the item. The dependence on multiple-choice items introduces a potential for uncontrolled guessing that threatens the accuracy and validity of the data derived from these assessments. Hence, given the dominance of multiple-choice items in such assessments, it is imperative that the data extracted from these items, along with the information provided to and the consequent data-driven decisions of all stakeholders, be grounded in accurate, reliable, and non-biased results. The extent to which guessing threatens the intended use of these data is the topic of interest in this study.

1.2 Issues Threatening Validity in Multiple-Choice Items

Moss et al. (2006, p.206) comment on the use of assessments in relation to “the soundness of those interpretations, decisions, or actions” taken in response to the information provided by different sources, and constrained administrative and data availability. In particular, they contend “educational assessment should be able to support professionals in developing interpretations, decisions and actions that enhance students’ learning. Validity refers to the soundness of those interpretations, decisions and actions” (p. 109).

An imperative of effective data-driven decision making is the availability of high-quality, valid information. The concept of validity in assessment has been researched over several decades. There is general agreement that validity can be defined as “how well a test measures what it is purported to measure” (Brown, 1996, p.231). Researchers (e.g., Embretson, 1983; Lord, 1964; Messick, 1989b) have identified multiple characteristics that support the concept of validity, not only in terms of the instruments assessing particular traits, but in the manner in which the information is used following the implementation and reporting of the assessment. According to Messick (1989), a source of bias that may impact validity is uncontrolled guessing in multiple choice items, which may influence construct and content validity by introducing uncertainty into the achievements reported. Uncertainty is also relevant to this study with respect to the concept of “consequential validity”, which relates to “the appraisal of both potential and actual social consequences of applied testing” (p. 20).

The validity of test data and the use of student results is an imperative for effective decision making. Central to this study is the identification and management of guessing in multiple-choice items to reduce the instances of actions based on a potentially flawed sets of results.

1.3 The Development of Assessment Models and the Evolution of Educational Measurement

The next section introduces Test Theory and the analysis techniques that have been derived from those theories. In relation to educational measurement, test theory defines the test construct, the analysis model and the range of outcomes that can be obtained from a test grounded in a theory. Test theories that guide the current Australian national and international assessment context will be discussed in detail in the body of this thesis.

1.3.1 Classical Test Theory (CTT)

Modern test analysis procedures have evolved from statistical procedures termed Classical (or Traditional) Test Theory (CTT). Initially dating back to the early 1900s, this theory stems from the subsequent work of Thurstone (1929, 1959), Fisher (1935), Cronbach (1951), and Guttman (1944, 1950). CTT concentrates on the features of the test: student performance is reported in terms of raw scores, and these are used to produce mean scores for categories (e.g., male and female), classes, schools, and jurisdictions, along with their respective standard deviations as indicators of student achievement.

CTT is constrained by the fact that each test is unique and the parameters that report performance on, or about, the test are unique to that test. Lord (1953) observed that in this model, student test scores and true scores are not synonymous with student ability. That is, ability scores are test independent, whereas observed scores and true scores are test dependent. Hence the measurement of student ability can be obfuscated by specific test interactions.

Traditionally under CTT models, student performance is reported in terms of raw scores, possibly with a standard error in the metric of the test. A test containing items with a high level of difficulty will generally result in a relatively lower score than would have been obtained if the items were easier. However, the underlying student ability in the trait¹ remains constant within a defined period. Parameters reported for test items include item difficulty, expressed as percent-correct or facility, and/or point bi-serial correlations, and the reliability of the test overall is reported using Cronbach's alpha (α) (Cronbach, 1951).

It should be noted that many international testing models are firmly rooted in CTT, with established norms and benchmarks for longitudinal comparisons (Goldstein, 2011). Such models are test specific; that is, if any of the test items are changed then the scores are no longer directly comparable. In CTT, test results are easily computed, explained, and readily understood by the broad community, but comparisons over time are not as well understood and explained. Modern Test Theory (MTT) (Kline, 2005) practices have been developed to address these perceived shortcomings in CTT.

¹ A trait is quality or characteristic that distinguishes a person. In the educational context this relates to the amount or quality of information or ability a person possesses in relation to a particular skill.

The confounding issue regarding public perceptions about the comparability of the nominally “same” test (e.g., NAPLAN) over time is that adjacent tests are expected to have the same overall level of difficulty when in fact they do not. In the Australian context, where it is considered important to provide feedback to teachers and students about performance on the test, new tests are constructed each year. Yet to assess student outcomes, and the results from year to year, the tests must be comparable. To achieve this, Australian jurisdictions have indicated that they want different tests of the same “construct” to be used to allow for comparison of results. This is very difficult to achieve with CTT, and so a modern measurement theory, Rasch Measurement Theory, is preferred because it transforms the performances of students on a particular test from a raw score to an ability estimate on a latent trait scale that is independent of time and/or specific test instruments used to assess the trait (Lord, 1952, 1953).

Hambleton et al. (1977) describe a latent trait in relation to testing situations:

Examinee performance on a test can be predicted (or explained) by defining characteristics of the examinees, referred to as traits, estimating scores for examinees on these traits, and using the scores to predict or explain test performance (Lord and Novick, 1968). Since the traits are not directly measurable and therefore ‘unobservable’ they are often referred to as latent traits or abilities (p.75).

1.3.2 Modern Test Theory and the Concept of an Educational Scale of Proficiency

Rasch Measurement Theory (RMT) is commonly included in the family of models described as Modern Test Theory (MTT). It is a psychometrically based measurement model that defines the performance of items and students in terms of a scale that defines a latent trait (e.g., Reading as a developmental construct over time, or Number as a subdomain of a Numeracy construct that includes learning in Algebra, Geometry, Measurement, etc.). The term *latent* emphasises the concept that discrete item responses are assumed to be observable indicators of ability that contribute to the estimation of traits. The trait being assessed is defined as a variable or scale that is the object of the testing program, wherein items are constructed specifically to provide evidence of mastery of skills and concepts that contribute to the scale.

RMT is underpinned by a relatively simple concept: the probability that a student will respond correctly to an item is a function of the ability of the students compared to the difficulty of the item. If the ability of the student exceeds the relative difficulty of the item, there is a greater than 50% probability that the student will answer the item correctly. As the positive difference between student ability and item difficulty increases so does the probability of a correct response. A negative difference between ability and item difficulty predicts a less than 50% probability of a correct response. The only variables that contribute to the student’s success in a response is ability and item difficulty. There is no consideration of a guessed correct answer, for which ability is irrelevant.

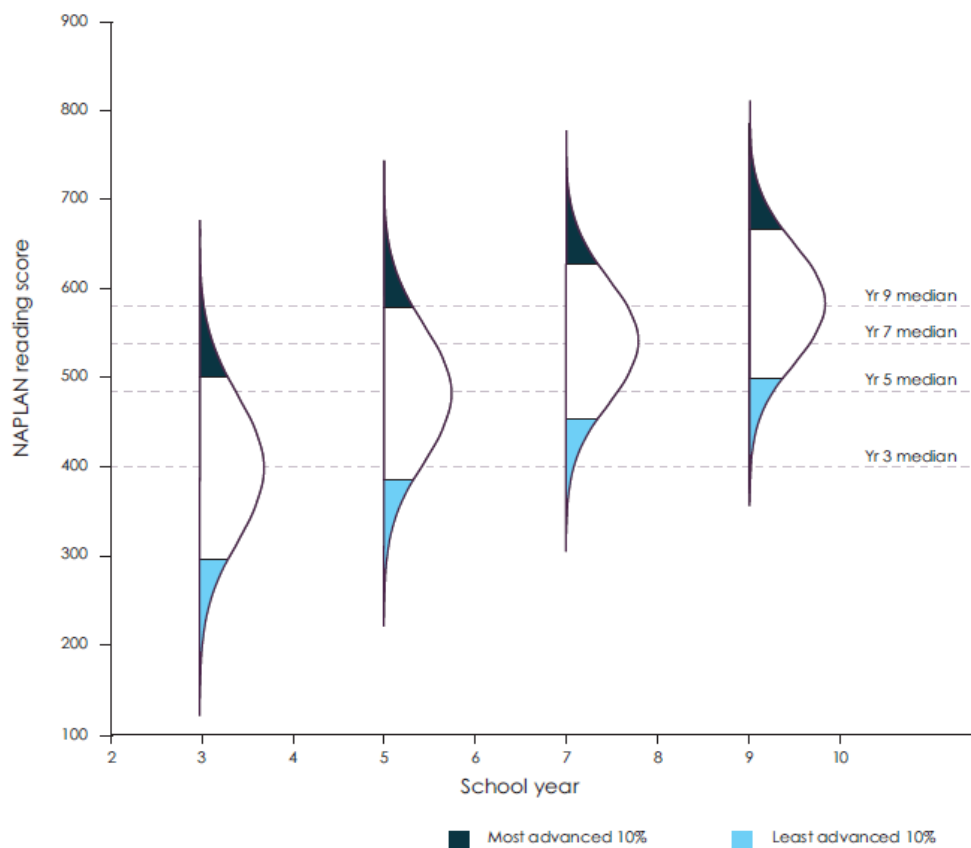
A representation of the NAPLAN Reading scale, which is derived using the Rasch measurement model, is shown in Figure 1.1. Of note is that the scale is vertical in order to cover multiple cohorts within a calendar year, and longitudinal to allow comparisons over time. Figure 1.1 shows the distributions of ability estimates in NAPLAN Reading scores as students’ ages increase.

Reference to the NAPLAN scale been included at this stage of the study due to its relevance to the Australian educational environment and to demonstrate the observed overlap between distributions of abilities in adjacent grades, and in particular a “contraction” of the spread of “achievement” reported in each successive cohort, with a reduction in the median “growth” between successive assessed cohorts. The distributions represented in Figure 1.1 have been generated using Rasch measurement techniques, which do not account for guessing in the analysis. The focus of this research is the degree to which the plateauing and contraction of reported ability may be a function of the reduced calibration of item difficulty due to guessing, which is not controlled for in the analysis applied.

Figure 1.1

Evidence of Achievement in Australian Schools – NAPLAN

Exhibit 11. There is a wide spread of achievement in Years 3, 5, 7 and 9



Source: Based on data from the NAPLAN National Report published by the Australian Curriculum, Assessment and Reporting Authority, <https://www.nap.edu.au/results-and-reports/national-reports>.

Note: Exhibit prepared by Geoff Masters, Australian Council for Educational Research.

Source (Gonski, 2018, p. 29)

In the NAP suite of assessments, the presence of correct guesses in student responses has been largely ignored by contemporary analysis techniques. A likely outcome of continuing with the current analysis practices will be to negate the ability of the assessments to fully represent the distributions of student achievements and thus reduce the capacity of new assessment strategies to recognise improvements in student achievement, as intended by the national agenda.

It is appreciated that in recent iterations of the NAPLAN assessments an adaptive model has been developed that takes account of student success to assigned items more appropriate to the observed success in common items presented in the early stages of the test. This strategy may have the potential to reduce the impact of guessing which is discussed in more detail in Chapter 11. However other national (e.g., NAP ICTL, NAP SL, NAP CC) and international assessments still administer non-adaptive assessment instruments (PISA and TIMSS with multiple linked forms) that utilize m/c as a dominant item type.

1.4 The Problem: The Impact of Guessing in Multiple Choice Items

Guessing (in particular random guessing) and its impact on individual student and cohort ability have been subjects of psychometric inquiry since the inception in the 1920s of multiple-choice items as a vehicle for “objective” assessment (Wilson & Engelhard, 2000, p. 744). They have been investigated in relation to test-taking strategies (Crocker & Benson, 1976), correction for guessing (Diamond & Evans, 1973; Fray 1969, 1980, 1988; Harper, 2003) and, to a lesser degree, Modern Test Theory (Hambleton, 1982; Andrich et al., 2012, 2015).

For the purpose of this research, a correct guess is a notion, judgement, or conclusion gathered from mere probability or imperfect information that is not associated with knowledge or ability in the trait of interest. In relation to using data generated from student assessments, a correct guess does not contribute to the measurement of the trait or the ability of the student as it will have been achieved by test-taking strategies unrelated to the student’s skill and/or knowledge of the trait of interest. Yet given the importance of the currently reported outcomes of national and international assessment, students are encouraged and rewarded by guessing in the absence of knowledge (Mehrens & Lehmann, 1973). Indeed, an unintended outcome of the raising of the stakes associated with national testing programs such as NAPLAN in the Australian context, is that students are being encouraged, and in some cases taught, to guess when they do not know the correct answer. This violates the requirements of not only measurement (Sadeghi, 2000), but also the measurement models (e.g., the Rasch Measurement Model) used to analyse the results of the NAPLAN tests, with consequent impact on the calibrations, estimation of ability, and reporting of student achievements.

A correct guess thus contaminates the estimate of the difficulty of the item and, consequently and directly, the estimate of student ability. Higher-ability students can typically engage with the relatively easy questions in a test and will have a high probability of answering correctly. However, as the items become increasingly more difficult, fewer students will have the requisite skills and understandings to answer the questions correctly, and a greater proportion will resort to some form of elimination processes or, in the extreme, random guessing to maximise their scores.

The systematic consequences of guessing on the analysis and reporting of achievement of a cohort, and of the individuals within it, cannot be uniformly adjusted. Lower-ability students will be advantaged by guessing, relative to their true score, and higher-ability students will be disadvantaged in relation to the estimate associated with their raw score. In addition, the consequent inflation of the estimation of the item difficulty will directly impact the subsequent estimation of student abilities across the whole cohort.

1.5 The Current Study

The problem of guessing in multiple-choice items is neither new nor unique, although no universally accepted solution has been found for it. Multiple approaches have been adopted to identify and correct for guessing in Classical Test Theory (CTT) and in Rasch applications. The limitations of these approaches are addressed in detail in Chapter 3 in relation to the properties of measurement. The current study advocates for an alternative solution from a validity and misinformation perspective by including a review and a critical evaluation of the past and current mechanisms applied to large-scale data.

A benefit of the Rasch Model is that it is a mathematical model that predicts the interaction between students of a given ability and items of a derived difficulty. Guessed responses are manifest by misfit to the predicted model. Leveraging the indicators of misfit, the current study has proposed and evaluated a possible way of indicating and accounting for guessing in students' responses that is grounded in an analysis of the degree to which misfit identifies a guessed item. The viability of this solution was evaluated through a series of studies involving simulated and other data collected for the express purpose of this research, as well as through its application to an existing large-scale educational dataset.

To improve the estimations of item difficulty and hence the fidelity of the measure of the trait, this study considered individual item–student interactions and associated parameters to identify and suppress individual interactions that exhibit a high probability of guessing. This differs from previous work in that rather than using current approaches of analysis of aggregated data, which do not consider the individual item/student interactions, it drew on the work of Guttman (1950), Waller (1989), Andrich et al. (2004, 2012, 2015), and others to identify guessing by investigating individual student response patterns.

This research has quantified the impact of guessing on the calibration of item difficulty and, as a consequence, student ability estimates. It has articulated and provided evidence in support of an indication procedure – the Guessing Indication Protocol (GIP) – that can account for guessing in multiple choice items analysed using the Rasch (1960,1980) measurement models that feature in the Australian context. The implementation of the revised model may mitigate the misinformation that is a consequence of uncontrolled guessing. In particular, it addresses that the degree to which the ability estimates of the more able students are underestimated and those of the less able students are overestimated. It is hoped that a consequence of the successful development of this algorithm to better account for guessing will be a higher quality of information provided to stakeholders and hence more appropriate data-driven policy decisions and resource allocation.

1.6 Organisation of the Thesis

Chapter 2 describes the theoretical underpinning of the mathematical models used to produce “measurement scales”. These scales provide the basis for reporting and generating the inferences upon which decisions are made at local and federal levels about student progress and the health of the Australian education system. It is the potential misinformation in these scales due to student guessing that is the subject of this research. Chapter 2 also reviews recent investigations and research into the issue of guessing with multiple-choice items.

Chapter 3 reviews the available analysis models that can be applied to the study of guessing and compares the strengths and weaknesses of each in relation to accommodating guessing. The chapter also elaborates the properties of the Rasch (1960) model, which is the focus of this study.

Chapter 4 presents the rationale for the research methodology implemented in this study.

Chapter 5 describes the application of the proposed algorithm to the simulated data and the GIP model that was thus developed. The chapter also reports on the procedures and analyses that were designed to refine the development of the GIP based on its application to multiple student-response distribution profiles.

Chapter 6 describes the application of the proposed GIP to a sample of locally sourced small-scale field data. These convenience-sampled “live” data were used to test the efficiency of the GIP model. Students were required to indicate which items had been “guessed” as a component of their individual responses. The data were first analysed by ignoring the self-identified guessing in each response pattern, and then by taking account of guessing as indicated by the students. These data were intended to provide a platform to inform the efficiency of the GIP model developed from the simulations described in Chapter 5.

Chapter 7 describes the application of the GIP on an existing set of data from a large-scale test and reports on the analyses conducted. The use of large-scale authentic data provided a platform to test the efficiency and efficacy of the GIP. The analyses report the degree of reclassification of achievements uncovered by accounting for guessing compared to the current practice that does not account for guessing.

Chapter 8 introduces time as a possible variable in indicating or confirming guessed items in a student’s response pattern. The consideration of response time as a potential indicator of probable guessing emerged as a result of the analyses with the large-scale test data. “Item-time” has become an available variable due to the evolution of online testing and its consequent capacity to record student response times by item. These data were factored into the model to confirm its application to the large-scale sample, and to potentially provide a further parameter in any future revisions of the model.

Chapter 9 discusses the findings of this research. It compares the outcomes achieved by the application of the GIP, highlighting the observed differences between those of the developed strategy and those that would otherwise be reported to stakeholders when probable guessing was not accounted for. It also comments on the degree to which improvement in student achievement estimates and test parameters might be assessed using the GIP.

Chapters 10 and 11 consider the implications of the study’s findings with respect to the potential for correcting misrepresentation in assessments that do not take account of guessing in student response patterns. Chapter 10 is concerned with the degree to which student achievement might be reclassified when using the GIP procedure to account for guessing, and Chapter 11 notes further potential topics for research in the area. The study contributes to the assessment literature through the investigation of the consequences of not accounting for guessing on the distribution of student ability estimates in a Rasch analysis environment.

Chapter 2

Issues in Measuring Student Ability and Achievement

2.1 Introduction

Access to accurate and reliable data is a fundamental requirement of sound decision making. In the physical sciences, there are defined scales that provide vehicles for knowledge and comparison of factors that influence decisions in spheres of interest. However, in the social sciences, particularly education, such scales do not exist a priori, and the measurement of knowledge, ability, and cognitive skills of humans is an ongoing and developing field.

The purpose of this chapter is to articulate the principles and properties of measurement as they relate to the social sciences and how they were used to evaluate the various analytical and assessment procedures undertaken in this study. The chapter briefly traces the history of the methodologies developed to account for guessing in multiple-choice (MC) items and to evaluate issues implicit in these approaches that mitigate the properties required by “measurement”. It then elaborates on the measurement approaches based on the Rasch Model (RM) theory that have informed this study’s proposed approach to address guessing.

2.2 Measurement and Scales

Since ancient times, units of measure have allowed for commonly understood quantification of lengths or amounts of produce (<https://www.britannica.com/science/measurement-system>). According to Tal (2020), “Measurement is an integral part of modern science as well as of engineering, commerce, and daily life ... Despite its ubiquity and importance, there is little consensus among philosophers as to how to define measurement, what sorts of things are measurable, or which conditions make measurement possible. Most (but not all) contemporary authors agree that measurement is an activity that involves interaction with a concrete system with the aim of representing aspects of that system in abstract terms (para. 1)”.

This quotation highlights the importance of measurement, but not what it entails or requires. The Science Learning Hub of New Zealand defines measuring “the process of obtaining the magnitude of a quantity relative to an agreed standard. Measurement of any quantity involves comparison with some precisely defined unit value of the quantity” (Measurement Standards Laboratory of New Zealand, 2019). The term implies assigning some numeric value that is relevant to the variable of interest and has a defined and commonly agreed meaning.

2.2.1 *Measurement in the Physical Sciences*

In the physical sciences, a measurement is a collection of quantitative or numerical data that describes a property of an object or event by comparing a quantity with standards that have been established by the use of guidelines such as the International Bureau of Weights and Measures, the Metric System, and the more recent International System of Units (SI) determined at the 26th meeting of the General Conference on Weights and Measures in 2018.

Such units and scales can be defined by reference to some directly observable property, and they are commonly developed over time. For example, the metre was originally defined as one ten-millionth of the distance between the equator and the pole measured along a meridian. To make this tangible, the metre was construed as the distance between two lines on a platinum-iridium bar maintained under a controlled environment in the International Bureau of Weights and Measures in Paris. The most recent, more precise, definition of the metre is $1/299,792,458$ of the distance travelled by light in a vacuum in one second. Scales are also used with other phenomena such as temperature, humidity, electrical currents, and luminosity, and their refinement over time has allowed for advances in science and technology.

Measurement in the physical sciences is thus defined by the characteristics of scalability, standardisation, a common understanding, acceptability, consistency, invariance, reproducibility, independence of the physical object of interest, and stability over time. It is also notable that a measurement scale relates to a single attribute (e.g., metre to length) and that the instrument used to measure the variable of interest is independent of the defined variable. Hence temperature (variable) is measured by a thermometer (instrument). This independence is a fundamental requirement of measurement. Rasch (1966) argued that the comparison of any two objects should be independent of the instrument used. He described this property of measurement as “specific objectivity”.

2.2.2 Measurement in the Social Sciences

When considering the factors that contribute to the measurement of social phenomena, physical attributes such as age, height, hair colour, weight, income, marital status, home ownership, etc. can be counted or classified in a consistent manner. Some of these attributes may also contribute to other scales that measure social conditions such as socio-economic status (SES). However, where the attribute of interest is student ability in a trait, as in this study in the mathematics domain, its lack of tangibility precludes the presence of directly measurable characteristics and therefore of instruments to directly measure them. For instance, there are no commonly understood and recognised instruments that can measure student ability; measuring it is constrained by the same challenges that affect the measuring of social phenomena such as attitudes, values, beliefs, culture, and well-being.

The principles of measurement in the social sciences should have the same requirements as in the physical sciences, but since it is not possible to physically measure attributes such as knowledge, understanding, or skills, it is necessary to collect artefacts and observe behaviours that demonstrate them. For example, surveys are used to generate data that inform the development of social scales. In educational attainment specifically, the instruments used include informal observations of student behaviour (e.g., conversations and observations); evaluating assignments and tasks; and formal assessments such as classroom and large-scale standardised tests.

2.3 Measurement in Student Achievement Tests

In the social sciences the prevailing measurement theories have evolved from psychology and psychometrics. They are grounded in the observation of cases and mathematical models that can predict and to some extent explain various phenomena. Current practices also draw on specifically constructed standardised tests that include MC items to measure student ability and achievement in traits of interest.

In his review of the literature on psychological scaling, Torgerson (1958) conceptualised the concepts of rigour and validity that underpin the measurement ideal:

Measurement of a property involves the assignment of numbers to systems to represent the property. In order to represent the property, an isomorphism, i.e. a one-to-one relationship must be obtained between certain characteristics and the number of the system involved and the relations between various quantities (instances) of the property to be measured. The essence of this procedure is the assignment of numbers in such a way as to reflect this one-to-one correspondence between these characteristics of the numbers and the corresponding relations between the quantities (p. 14).

Student ability, for instance, may be depicted by a series of numbers designed to represent increasing amounts of the property of interest. The intent of the measurement model, as defined by Torgerson, is to represent a system characterised by a one-to-one relationship between mathematical ability and the number assigned to represent that ability. The development of the one-to-one relationship is affected by students' responses to items that represent increasing skill in the attribute.

2.3.1 Measurement Scales in Education

The purpose of establishing a scale is therefore to have a fixed external reference against which to compare objects or observations that are stable over time. In the physical sciences, precise measurement includes allowing for an acceptable margin of error that is a function of the quality of the measuring instrument, the object, and its use, and how critical the margin of error is in relation to its use.

When measuring student ability in education through testing, precision is influenced by the quality of the items included in the test as indicators of both the trait and increasing knowledge and ability in that trait. Models of modern measurement theory will be discussed in detail in Chapter 3. Typically, they calculate two parameters to provide information regarding the precision of the instrument. These are Fit Statistics, which denote the difference between an observed outcome and the expected outcome of a student–item interaction; and the Measurement Error associated with the individual items and the overall test. The concept of measurement error is a feature of any quantification of a variable. In modern assessments, MC items are the dominant structure in test constructs. In MC items, a source of measurement error is random guessing, and as such, its removal would improve the precision of measurement estimates.

As indicated in Chapter 1, in random guessing the response is independent of any knowledge or understanding of the trait being measured. Hence a correct answer, generated randomly by chance, imparts no information to the external observer about the amount of knowledge the student has displayed in the variable of interest. This situation is a direct contravention of Torgerson's definition of a one-to-one relationship between the number representing the property and the amount of the property demonstrated by the student. Although correct responses to items increase the measure of the overall ability of a student in the attribute, correct guesses are a threat to the veracity of that measurement and contribute to uncontrolled measurement error.

2.3.2 Issues in Measurement With Multiple Choice Items

Traditionally, tests comprised of MC items have been scored using the "sum of correct answers" scoring method (Bereby-Meyer et al., 2002; Kurz, 1999). In the context of MC, correct answers are typically scored with a value of one (1), incorrect answers and absent or omitted answers with a value of zero (0). The sum of the scores for correct responses is the student score, which is traditionally a "sufficient statistic" to rank the student in the test and provide a "measure" of the student's ability in the trait of interest. Students who lack the ability to solve a particular item gain marks by guessing (Budescu & Bar-Hillel, 1993; Choppin, 1988; Frary, 1988; Kubinger et al., 2010), which introduces a random factor into test scores that lowers their reliability and validity (Bereby-Meyer et al., 2002; Burton, 2001; Kubinger et al., 2010; Prihoda, 2006). In such cases, analysts cannot distinguish between correct answers based on knowledge mastery from those based on a guess (Bar-Hillel et al., 2005).

Clearly the aim of any test is to create an instrument that results in accurate data, is fit for purpose, and permits the generation of reliable and valid reports. For instance, in large-scale national educational testing programs such as NAPLAN, the measurement of student ability involves various processes and highly specialised experts. These are mobilised toward maximising the probability that the instrument used to determine a student's true score on a test (reflecting their ability) minimises avoidable errors in the data and consequent reports. Typically, professional test developers construct an assessment framework that aligns the types of test items with an accurate indication of the students' ability on the trait of interest. The framework is then further refined as a test specification that defines the combination of the items and their relative difficulty so that the full range of student abilities in the content area can be assessed. A series of reviews are then conducted as the experts prepare individual items to match the framework and specifications, thus ensuring that the final instrument has good fit to the intended scale that reports ability in the variable of interest. Supporting guides and instructions to standardise the administration of the instrument are also prepared to minimise any inconsistencies that can add to errors in the measures. These strategies and processes contribute to the reliability and validity of the test and the results generated from their implementation.

2.4 Early Research and Traditional Strategies to Address Guessing in MC Items

2.4.1 Strategies to Discourage Guessing

Despite extensive efforts such as those just mentioned, the inclusion of MC items in testing instruments has introduced an unintended source of uncontrolled measurement error. There is statistically a $1/n$ probability (in cases of one correct answer among n response options in the item) of a correctly guessed response independent of knowledge of the variable. By definition, a correctly guessed response contributes nothing to the measurement of the ability of a student in the variable and, by association with the calibration of the item, introduces error and bias into an item's true facility/difficulty.

Since the inception of MC items, there has been acknowledgement of the potential issue of guessed responses and various strategies and techniques have been implemented to minimise their impact (Yerkes, 1917). Researchers have concluded that guessing reduces reliability by increasing the error variance without a corresponding increase in true score variation, and that since higher-ability students have a lower probability of obtaining a correct answer by guessing, guessing disproportionately advantages lower-ability students.

To discourage random guessing in tests that include MC items, there have been attempts to reduce it by imposing penalties for incorrect answers. These have taken two forms:

1. Implementing very explicit administration rules to discourage guessing (Traub et al., 1969); and
2. Penalising students for incorrect responses (underpinned by the assumption that all incorrect responses are random guesses) (Kurz, 1999).

In relation to the first strategy, in its simplest form students are simply advised not to guess (Davis, 1967; Frary, 1988). The issue of managing guessing behaviour through instructions about responding to items may also be related to "test-wiseness". By contrast, and particularly in large-scale and high-stakes assessments, students are instructed to answer all the questions they are sure of and omit any items that they are unable to answer with certainty (Prieto & Delgado, 1999). In an alternative strategy, students are instructed to guess whenever they can eliminate one or more alternative choices (Betts et al., 2009; Davis, 1967; Frary, 1988; Hammond et al., 1998).

In relation to the second strategy, various scoring formulae have been proposed to correct for guessing (Kurz, 1999). Perhaps the most prevalent is the "rights minus wrongs" correcting model (Kurz, 1999), which penalises the student for incorrect responses. The fundamental idea behind this scoring method is that students acknowledge they will lose marks for incorrect answers and, as a consequence, they will be discouraged from guessing (Betts et al., 2009). This is expected to increase reliability and validity because the test score is a truer reflection of a student's ability (Kurz, 1999). The expected total score should be zero if a student guesses all answers at random. The term "negative marking" has been used to describe this scoring method. For this to happen, the penalty for an incorrect answer should be $1/(n - 1)$, where n stands for the number of response options (Karandikar, 2010).

The logic that underpins negative marking is that if the student does not have a correct response, then any response is a function of guessing or misconceptions. Accordingly, some studies have reported an increase in validity or reliability when negative marking is implemented for incorrect responses (Burton, 2002; Muijtjens et al., 1999). Other authors argue that by implementing negative marking, MC tests measure students' answering strategies and risk-taking behaviours instead of their mastery of domain knowledge (Budescu & Bar-Hillel, 1993; Choppin, 1988; Fowell & Jolly, 2000; Hammond et al., 1998; Kurz, 1999; Moss, 2001; Prihoda, 2006). However, even in cases where negative marking has been implemented, it has not solved the guessing problem (Bar-Hillel et al., 2005; Betts et al., 2009) and has been observed to introduce new problems. This is because students differ in their guessing behaviours, with some daring to take more risks than others (Chan, 2019; Choppin, 1975; Sharma et al., 2020). This introduces a concern about students' risk attitudes adding to uncontrolled sources of variance, thus reducing the test's reliability and validity (Bar-Hillel et al., 2005). It is also apparent that the process of negative marking disadvantages lower-ability students disproportionately, as students of higher ability have less need to guess. The contention over these issues highlights the challenge in giving recommendations that are fair and beneficial to all students. Students react in inconsistent ways, and thus Budescu and Bar-Hillel (1993) contend that the question of whether students should be instructed to guess or not is far more difficult to answer than it seems. It is also difficult for students to ascertain the optimal decision strategy under negative marking.

An alternative model proposed by Traub et al. (1969) rewards a student for *not* guessing by awarding points for omitted items rather than penalising for incorrect responses. They argue that students do not feel threatened by receiving a reward for skipping items compared to receiving a penalty for incorrect responses. This presents a psychological advantage since it rewards desired behaviours instead of penalising undesirable behaviours, and it thus leads to more valid and reliable estimates of their achievement (Crocker & Algina, 1986; Prieto & Delgado, 1999).

Another approach is "correction for guessing". Prior to being able to access the power that computers now provide for investigating and analysing data, examiners and researchers implemented correction-for-guessing algorithms that adjusted student scores based on the number of incorrect responses using the generalised formula:

$$S = R - W(k-1)$$

where S is the corrected score;

R is the observed number of correct responses;

W is the number of incorrect responses; and

k is the number of options in each item.

This correction-for-guessing formula was based on the assumption that all incorrect responses were a result of random guessing and that therefore a proportion of the correct responses would have been achieved by random guessing at a rate of $1/k$ where k is the number of options in the MC item. Hence the "true score" was adjusted down to reflect the assumed guessed component (Hendrickson, 1971; Sabers & Wright, 1969; Stanley & Wang, 1968). Research into the reliability and validity of tests for which the correction-for-guessing strategies were implemented has provided a range of results (Davis, 1964, 1967; Davis & Fifer, 1959; Frary, 1969; Thomas et al., 2005). Lord (1963) researched the use of correction-for-guessing

procedures and concluded that the increased validity in the results as a consequence of this procedure was only marginal, and he supported the implementation of the procedures only in cases where

1. the tendency to guess varies considerably among the students;
2. there are fewer than five options in each item; and
3. the tests overall were very difficult.

There is general agreement among the researchers just mentioned that strategies involving correction for guessing, which physically adjust a student's score, actively discourage risk taking by students who have partial or incomplete knowledge of a particular skill as tested by an item. The imposition of a penalty for an incorrect response discourages any "not for certain" responses, and there has been some contention about whether the assumptions underpinning the application of scoring formulae are realistic (Cliff, 1958). Moreover, guessing might benefit students who guess frequently compared to students with equal ability levels who are more reticent to guess (Choppin, 1988; Muijtjens et al., 1999).

Given that the probability of a random guess for an unknown answer in a traditional MC item is $\frac{1}{k}$, where k is the number of distractors, there have been other MC item structures that attempt to reduce the probability of a correct guess. These include structures such as two correct answers in a five-option item that reduces the probability of the composite correct response to $\frac{n}{k} \times (n-1)/(k-1)$. In the two-from-five structure the probability of a correct response is $\frac{1}{10}$, or 10%, compared to the traditional proportion of 25% (Kubinger et al., 2010).

It is apparent, therefore, that the problem of guessing in MC test administrations is both longstanding and persistent, and there is still no common approach to addressing it. In addition, traditional approaches such as those discussed above may introduce new factors into response patterns that impact on measurement, for example, the face validity of negative marking. However, in each of the strategies so far described, an overarching issue is that they are not in accord with the concept of measurement. Assessment theory presumes that a correct answer is an indication of ability in the trait and contributes to the measurement of ability. An incorrect answer does not detract from the measurement of the student's ability or the measurement of the trait.

It is worth noting that the implementation of these traditional approaches to account for guessing contravenes the requirements of "measurement" because they do not meet the demands of an external, independent, stable scale that relates the process to a measure. The assumption underpinning the processes of correcting for guessing – that all incorrect responses involve random guesses – is questionable as well. It ignores the argument that a student's knowledge of a particular trait can be evolving and that a correct answer can be determined by the application of partially incomplete knowledge.

2.5 Psychometric Approaches to Account for Guessing

In modern administrations of large-scale MC assessments that use Modern Test Theory, such as NAPLAN, there tend not to be any a priori penalties applied to incorrect responses of MC items. Paradoxically, there are anecdotal reports that because students cannot be penalised, they have been encouraged to use both elimination strategies and, as a last resort, random guessing to improve their overall total scores in these tests. This is despite the advances in computing capacity to interrogate and manipulate data in multiple ways.

2.5.1 Introduction to Modern Analysis Techniques to Account for Guessing

Advances in the development of a mathematical theory that underpins the relationship between student performance and a measurement scale have been grouped in MTT under the general term Item Response Theory (IRT). As well, in the modern large-scale testing environment, two analytical applications within the RM (Rasch, 1960) dominate the field: the 1PL Model and the 3PL Model (Birnbaum, 1968). As mentioned in Chapter 1, the RM, is particularly relevant to the Australian context due to its dominance in the NAP suite of assessments and the context of those assessments in the national goals. These theories are discussed in detail in Chapter 3.

Although the use of the RM has become commonplace over several decades, contention remains about its relative appropriateness and the information it provides to various stakeholders in relation its treatment of guessing. This issue is important for researchers who have investigated mechanisms to account for guessing in the RM paradigm (Andrich et al., 2012, 2015; Marias, 2015; San Martin et al., 2006; Waller, 1987). In recognition of their success in these efforts, Andrich et al. (2015) noted, “Guessing in the responses will affect the fit to the Rasch Model, thus eliminating responses that are likely to have been guessed should improve fit” (p. 430).

The following sub-sections describe the previous approaches that have informed the current study.

2.5.2 Previous Approaches to Account for Guessing in the IRT Paradigm

As just mentioned, two general approaches have typically been adopted to address guessing using Rasch Model. The 1PL Model tests item fit and other parameters to assess whether the item contributes to or confounds the measurement of the trait. In cases where the fit statistics indicate that the item does not contribute to the measure, the item is excluded from the final analysis. This strategy is possible when the instruments are secure and not in the public domain. It can be problematic to omit items in cases where other stakeholders are aware of the test structure.

The second approach attempts to build on the RM in the analysis phase of the assessment procedures to account for inconsistencies in the data, including possible guessing. This strategy has been imbedded in the 3PL analytical procedures (Birnbaum, 1968; Hambleton, 1983; Hambleton et al., 1991). However, as will be discussed in Chapter 3, this technique violates fundamental measurement principles. Nevertheless, the 3PL Model attempts to account for guessing, and possibly other noise in item calibration, from student response patterns by creating a parameter to address the issue in the item difficulty estimate (Birnbaum, 1968).

The next sub-section traces previous research and highlights an issue common among these approaches, which is that they attempt to address the issue of guessing by analysis and modification of the calibration of items attempted by a *group* of students involved in the test. Guessing is not a group activity; it is assumed to increase in frequency as items increase in difficulty for the group and as individual abilities within the group decrease. But guessing is an individual item/student interaction that impacts estimates of an individual student's ability. As such, the typical approach of aggregation of the potential changes in individual student ability estimates consequently impacts estimates of achievement of the group. The point of differentiation between previous approaches and the one proposed in this study is that in the latter the guessing is addressed at the individual item/student level, not at a more global level. The next sub-section will review approaches that have previously been attempted and from which the proposed approach diverges.

In recent times there has been an increase in the development of methodologies that elaborate the RM and attempt to account for guessing through analysis algorithms (e.g. 1PL-AG procedures described below). These methodologies recognise that the RM does not account for guessing in student responses. As Park et al. (2015) put it: "In terms of IRT modelling, multiple-choice (MC) items impose an interesting problem that cannot be handled with the conventional Rasch model" (p. 449).

Lin (2018) implemented a methodology to correct for likely guessing in the CAL English Proficiency Test (EPT) for Students: Listening and Reading. This strategy suppresses any item that, in the initial analysis, had a difficulty estimate greater than 2.0 logits higher than the ability estimate of the student. A discrepancy -2.0 logits between student ability and item difficulty means the probability of a correct response is less than 12%. Lin (2018) found that "likely guessing ... [has] a greater impact on difficult items than on easy items" (p. 414), and

It is evident that correcting for the effects of likely guessing results in increased person ability measures for high-performing examinees, while such effects have a smaller impact on low performing examinees ... [and the] increase in person ability measures for high-performing examinees are a directly due to the relatively difficult items becoming more difficult after removing responses with likely guessing (p. 416).

Lin's approach seems a viable concept, but the threshold value is arbitrary and does not appear to have a stochastic underpinning. And it remains limited because it is biased towards impacting the estimates of higher-ability students, not those less able.

In other approaches, researchers (e.g., Barnes, 1988; Waterbury & Mars, 2019) have investigated various approaches to accounting for guessing within a RM paradigm. These researchers generally contend that methods that adjust for guessing tend to increase standard errors in the measures but decrease the bias. Barnes (1988) undertook research in which simulated and actual small-sample test data were used to compare Rasch ability and item estimates with estimates obtained from the three-parameter 3PL model, and from two modified Rasch models that incorporated a constant guessing parameter. She found that "ability estimation errors tended to be systematic in that the Rasch model overestimated ability, whereas the guessing models underestimated ability" (p.1).

Jia et al. (2019) also investigated the effect of aberrant responses as indicators of guessing or cheating and attempted to derive a generalised formula of bias with aberrant responses. Barnard (2013) proposed and investigated a methodology called Option Probability Theory (OPT) and concluded that although the challenge of identifying guessed item/student responses was not significantly determined, there was a significant increase in the certainty of responses that would be useful to inform teachers.

San Martin et al. (2006) investigated the ability-based guessing model 1PL-AG, including an ability function to the guessing parameter that addressed the shortcomings of the 3PL model. The 1PL-AG method breaks the estimation of parameters and person ability estimates into two components, in recognition of the contention of Hutchinson (1991). It also attempts to resolve the interaction between the probability to guess and student ability with respect to each item. San Martin et al. (2006) identify a possible trade-off between the estimates of guessing and the discrimination parameter (a_i) caused by the simultaneous estimation of both, and the interaction effect of lower discrimination for lower-ability groups that are the object of the estimate of the ‘guessing’ parameter. According to the 1PL-AG, the probability of correct response to item i given ability of person j is β_j and item parameters is:

$$\Pr\{x = 1; \beta_j\} = 1/[1 + \exp(-(\beta_j - \delta_i))] + [(1 - 1/(1 + \exp(-(\beta_j - \delta_i)))) \cdot 1/[1 + \exp(-(\alpha_i \beta_j - \lambda_i))] \quad \text{Eqn 2.1}$$

Source: San Martin et al, 2006. <http://apm.sagepub.com/cgi/content/abstract/30/3/183>

where: β_j is the latent ability of the student (j) in the trait of interest

δ_i is the relative difficulty of any particular item (i) that is an indicator of understanding in the trait of interest

α_i is the slope parameter – the discrimination - of item I , and

λ_i is a guessing probability for an examinee with an average ability ($\beta_{j=0}$) for a normal distribution that fits the model.

Although this study concentrates on the RM as the analysis technique applied to large-scale dichotomous data, the RM is not the only model derived from the one-parameter Simple Logistic Models (SLM). Other models that explain student responses include:

- Rating Scale Model (Andrich, 1978b; Masters, 1980)
- Partial Credit Model (Masters, 1982; Master & Wright, 1981)
- Extended Logistic Model (Andrich, 1988; Tognolini, 1989)
- Linear Logistics Model (Fisher, 1973)
- Binomial Trial Model (Andrich 1978a, b); and
- Poission Counts Model (Rasch, 1960/1980).

Other alternative IRT models have also been proposed to explain guessing behaviour, including:

- (a) a fixed value of $1=L$, with L being the number of options in MC items. However, this model, to account for guessing, reflects the concept of random guessing in its parameterization, which studies have shown may be under-estimating or over-estimating the guessing factor depending on the attractiveness of incorrect options (Hambleton et al., 1991);
- (b) an average guessing parameter across items in a MC testing domain;
- (c) a guessing parameter specific to each item, as in the three-parameter logistic (3PL) model (Birnbaum, 1968); and,
- (d) models for guessing that are dependent on person ability (Hutchinson, 1991).

It is beyond the scope of this study to consider all these models; they are described in detail in the cited publications.

In relation to the approach proposed by this study, the works of Anderson (1995, 2005), Waller (1989), and Andrich et al. (2012, 2015), summarised below, have had the most influence due to their focus on the relationship between ability and item difficulty and the degree of fit divergence from the expected outcomes that may provide indicators of guessing.

Andersen (1995, 2002) proved a theorem that relates variances of parameter estimates from samples and subsamples and showed an application where the theorem was central to the hypothesis tested, namely, whether random guessing to multiple choice items affects their estimates in the RM. The theorem provided a context for the work of Andrich et al., (2012), which this study extends.

Waller (1989) and Choppin (1985) independently hypothesised that guessing was more present in student responses for more difficult items. In the same way that the current study attempts to identify guessing in student response patterns, Choppin (1985) proposed a two-parameter variation of the RM based on the contention that it is possible to identify those responses to relatively difficult items that may have resulted from guessing. Once those responses have been identified they should be suppressed when constructing the variable. Yet the problem encountered by Choppin was in the confidence around identification of guesses. These approaches highlighted a need to develop indicators in which one can have a degree of confidence regarding the identification of highly probable guessed responses at the individual item/student interaction level. This need is elaborated on in Chapter 4.

The studies outlined above have contributed to the strategy to account for guessing proposed in this research. In most cases, these prior investigations focused on the items rather than on the interaction between items and students in developing strategies and recommendations. However, the approach that has most influenced the method developed in this study in relation to accounting for guessing is that of Andrich et al. (2012, 2015). This is because it concentrates on the misfit in item/student interactions as an indicator of error due to guessing. The next section elaborates on Andrich et al.'s work.

2.6 Recent Approaches to Account for Guessing Within the Rasch Model Paradigm

2.6.1 *The Approach of Andrich et al. (2012, 2015)*

The psychometric team led by David Andrich developed a logic to account for guessing that built on the work of Waller (1989) and Andersen (1995, 2002). They “formalised Waller’s (1989) procedure for removing the effects of guessing and demonstrated how statistical bias in (item) difficulty estimates can be quantified, tested for statistical significance and, most importantly, removed” (2012, p. 413). The Andrich et al. (2012) procedure is based on a parsimonious identification of guessing that involves the inspection of the item–student residuals based on item difficulty and student ability. Specifically, Andrich et al. (2015) acknowledge the issue in MC items in relation to attempts to remove the effects of the problem:

The major hypothesis is implied ... random guessing is a matter of degree, increasing as a function of the difference between the difficulty of the item and the proficiency of the student – that is, the more difficult the item relative to the proficiency of the student the more likely the student will guess (p. 413).

Andrich et al. (2012, 2015) interrogated student response patterns and the interaction between ability and item difficulty to suppress items that have a high likelihood of being guessed in the item calibration phase. They then recalibrated ability estimates based on the outcomes of a multi-phase process. This is termed the Tailored Analysis. Table 2.1 shows the Tailored Analysis step in the Vertical Equating Design process implemented by Andrich et al. (2015) to better estimate student ability by accounting for guessing in a NAPLAN assessment.

Table 2.1

Vertical Equating Design for the 2013 NAPLAN Reading Scale

Step	Analysis	Description
1	Initial	An analysis of initial data (no missing responses and equal sample sizes of 9,382 for all year groups)
2	Tailored Analysis	Analysis of anchored data (responses likely to have been guessed changed to missing responses)
3	Origin-equated	Reanalysis of the initial data (the origin equated to the mean of the four easiest items of the Year 3 group in the tailored analysis)
4	All-anchored	Reanalysis of the initial data (all difficulty estimates anchored to those from the tailored analysis)

Source: Andrich et al. (2015, p. 419).

The logic that underpins Andrich et al.’s (2015) Tailored Analysis to account for guessing is that the easiest items are the least likely to have been impacted (biased) by guessing. In response, their method involves treating *all* items beyond a defined item–student interaction level, for which the probability of a correct response is less than (an arbitrary value) of 0.3, as “not administered”.

The impact of this process is to recode items as missing data (p. 430). Treating these items as missing has a dual effect by reducing

1. the number of student responses and hence the data set being analysed; and, as a consequence,
2. the number of interactions available for the conditional difficulty estimations that are a feature of the implementation of the model for item difficulty estimation with student data.

This process makes the response patterns more “Guttman-like”, with any noise beyond the estimated person ability or probability of a correct response interaction being eliminated from the analysis. The process has the effect of refining the variable and improving the fit of the Tailored data set to the model. However, the recoding of items to “missing” reduces the number of interactions that are available for estimating item difficulty, and the relative difficulty of harder items. The process thus reduces the precision of the estimates and increases the measurement errors associated with estimation.

It should be noted that in any test design there are typically a few items targeted to the lower-ability students. It is these easy items that were the focus of Andrich et al.’s (2015) Tailored Analysis; they selected the easiest five or six items in the anchoring stage of the method. It should also be noted that the facility rates of these items (as informed by the NAPLAN trialing process) are in the range 80% to 95%; in other words, the items are specifically designed to engage the lower-ability students. This approach remains problematic because it means that, on average, these items are beyond the ability level of between 5% and 20% of the student population. Hence it is likely that these items too will be impacted to some degree by the students *most likely* to guess in the NAPLAN assessment.

Andrich et al. (2015) evaluated the efficacy of accounting for guessing in their Tailored Analysis by investigating the relative fit of the analyses to the RM, using the chi-square statistic as an indicator of how well the results represented the expected outcomes. They note: “This statistic is a test of deviation from perfection of the data from the model” (p. 427). Andrich et al. (2015) concluded:

The fit in the tailored analysis is substantially better than in the origin-equated analysis. ... This confirms that fit can be substantially improved by tailored analysis, further confirming the hypothesis that guessing is a function of the difficulties of the items relative to the proficiencies of the students (p.428).

2.6.2 Comments on the Approach of Andrich et al. (2012, 2015)

Andrich et al. (2015) found that “not controlling the guessing bias underestimates the progress of students over seven years of schooling with important educational implications” (p. 416). Specifically, their research concluded that their methodology showed that the abilities of higher-ability students are under-estimated when guessing is ignored. However, since the method anchors a number of easy items, the interaction between item difficulty and raw score determination defined by the model will limit the impact on students who are only able to engage with the easiest items. Hence this model is limited in its ability to adjust for lower-ability students.

Commenting on the work of Andrich et al. (2012), Humphry (2015) noted that “although this approach may account for the effects of guessing on item estimates, it entails a loss of information and provides no

solution for obtaining person estimates based on all of the response data” (p. 193). The approach proposed in this research addresses this limitation.

2.7 A Proposal to Account for Guessing in the Rasch Model

As will be detailed in Chapter 4, the approach to account for guessing proposed that is evaluated in this study follows a similar strategy to that of Andrich et al. (2015) outlined in Table 2.1. The major differentiation between the proposed approach, as summarised in Table 2.2, and the work of Andrich et al. (2015) is that in the proposed methodology only items identified as probable guessing are modified in the development of the revised measurement scale. All other responses are maintained. This means that only items that fail the defined guessing identification parameters are recoded. The advantage of this method, rather than recoding all items beyond a specified threshold, is to minimise the consequent loss of data and the impact on the facility rates of items that have had significant numbers of cases in which the response is recoded as missing.

Table 2.2

Methodology Design for the Overall Research

Step	Analysis	Description
1	Initial (INIT) - simulated - field trial - large scale	An analysis of initial data (no missing responses all data analysed using Rasch measurement applications (RUMM and Conquest)
2	Suppressed (GS) - simulated (defined guess) - field trial (self-identified)	Recoding of the data and re-analysis to develop a revised Raw Score to ability table (responses self-nominated, or defined as, guessed recoded as missing data)
3	Guessing Indication Protocol GIP3A – item recalibration GIP3B – ability estimate recalibration - all data sets	a. (GIP3A) recalibration of item locations b. (GIPINIT3B) reanalysis of the initial data (the original responses) with the locations of items identified as likely guessed by the GIP3A procedure recalibrated.

To evaluate this approach, the proposed solution employed simulated data to determine an algorithm that is feasible in identifying guessing, the Guessing Indication Protocol (GIP). The GIP was then applied to condition data from a field trial and a large-scale assessment in an attempt to remove/reduce the measurement error imputed through guessing and to resolve the impact of guessing on the ability estimates of the students.

Since this study is concerned with the measurement and reporting of individual students, as well as groups and/or cohorts, with a particular focus on NAPLAN, the RM is its focus. This is due to its measurement properties and the fact that it addresses the face validity issue of reporting to every participant in the test.

The proposed approach attempts to both build upon and improve the research described earlier in order to provide better estimates of student ability and item difficulty, as well as an indication of the presence of guessing by the individual at item-person interaction.

The ultimate intended outcome of this research is to estimate the impact of not accounting for guessing in the reporting of student ability to stakeholders in the variable of interest. Apart from Marais (2015), this impact evaluation is the area least investigated in previous research.

2.8 Summary

This chapter has articulated the principles and properties of measurement as they relate to the social sciences and how they were used to evaluate the various analytical and assessment procedures undertaken in this study. It also explained how the issue of guessing is a threat to accurate measurement, and briefly described some of the strategies that have traditionally accounted for guessing in student achievement tests.

Modern assessment theory commonly uses IRT to analyse student assessment and report achievement. In particular, in the Australian context of NAPLAN, the RM is commonly utilised. However, this model does not take account of guessing. The chapter concludes by introducing a model that uses the RM as the base measurement instrument and proposes supplementary analyses based on individual item–student interactions to account for probable guessing in the student response patterns.

Chapter 3

Background of Modern Test Theory Applied to Student Achievement

3.1 Introduction

The purpose of this chapter is to review the development of modern data analysis techniques to estimate student ability and achievement and to evaluate each in relation to the aims of the research. The chapter provides a detailed review of requirements of the Rasch Model (RM), which was identified as the most appropriate measurement model to estimate individual student ability and as having specific relevance to the Australian context and the analysis of large-scale population assessments.

3.1.1 Relationship Between Student Ability and Achievement

For the purposes of this study, student ability is defined as “the possession of the means or skill to do something” (Messick, 1984, p. 159). Messick (1984) contends that student ability evolves as a result of an interaction with environment, culture, and prior or current instruction. Its value can be estimated as the amount of the trait that the student has exhibited in a particular test. To determine student ability, consideration of the responses of the cohort that has completed a test is used to calibrate the relative difficulty of the items of the test and subsequently the ability of each student based on his/her result on the test.

The considerable research and commentary on the relationship between student ability and achievement (Abel, 1984; Nizoloman, 2013; Rohde & Thompson, 2007) relate to the impact of external factors in addition to student ability that contribute of measures of achievement. In the education environment, the term achievement is often used interchangeably with “performance”, although it is contended that the two are not synonymous. For Messick (1984), for example, “Educational achievement refers to what an individual knows and can do in a specified subject area” (p. 154), and the capacity to demonstrate ability (or performance) is a function of “comprehension, memory retention and retrieval, reasoning, analysis and restructuring, evaluation or judgement and fluency” (p. 156). In this distinction, student achievement is the product of several factors, including, but not limited to, ability. Accordingly, and throughout this thesis, student ability is estimated using specific psychometric analysis techniques. In contrast, achievement is reported in relation to the comparison of the outcomes of groups of students in the tests evaluated.

3.1.2 Early Estimation of Student Ability

The evolution of assessment instruments and analysis techniques to assess student achievement has its roots in the work of Fischer (1935), Guttman (1944, 1950a, 1950b, 1954) and Thurstone (1927, 1929, 1959) in the first half of the 20th century, and its later development with Andersen (1977), Birnbaum (1968), Hambleton et al. (1991), Rasch (1960/1980), Wright (1968, 1977), and Wright and Masters (1982). Most recently, Andrich (2012, 2015, 2016) and other noted psychometricians and statisticians have added to the

body of knowledge and the development of applications that fall under the generic terms Modern Test Theory (MTT) and Item Response Theory (IRT).

Modern Test Theory and its component Item Response Theory have their origins in the traditional statistical treatments of data and, in particular, the relationships between test participants and the relative difficulty of test items for the target group. Indicators of the quality of the instrument in achieving its aim of measuring student ability were grounded in statistics such as item discrimination as a measure of the correlation of the items with the total score and reliability statistics. These approaches were coordinated in Gulliksen's (1950) work, *Theory of Mental Tests*, and were reviewed by Guttman (1953) as he developed theories on measurement that challenged the traditional test theory approaches.

Guttman's (1950a,b, 1954) work is a significant development in the theory of student ability measurement as applied to educational testing. For instance, Guttman (1950a) proposed:

A person's score tells what his responses were to each question ... [the deterministic paradigm] ... A person with a higher score is "favourable" on all questions that a person on a lower score answers "favourably", and on at least one more in addition. This rank order, furthermore, exists not only for the given series of questions, but is the same as the rank order that would be obtained with any series of questions in the same area (p. 20).

Our definition of a single continuum as a series of items each of which is a simple function of the scale scores permits a clear-cut statement of what is meant by a rank order based on a single variable (p. 154).

Taken together, these statements define a unidimensional variable that can be constructed from a series of related test items.

Although not explicitly stated in the literature, Guttman's (1950) proposition that student achievement in a test item is a function of student ability and item difficulty was a precursor to the development of IRT. Guttman proposed a deterministic theory that "explains" the pattern of student responses to items. In essence, if items are strictly ordered according to relative difficulty, then as a student interacts with the set of items, they will correctly answer those for which they have knowledge and/or ability and incorrectly answer those for which the inherent difficulty of the items is in excess of their knowledge and/or ability. In this construct, a student's measure of achievement and ability is derived in a deterministic manner. For Andrich (1988), the similarity of Guttman's and Rasch's aims when constructing measures align with the theoretical construct of what is described later in this chapter as a "model of intent" (p.39). The particular importance of Guttman's work, the rigour that underpins requirements of the Guttman scales, and their relevance to this research are discussed in detail in Chapter 5.

3.2 Development of Modern Test Theory

3.2.1 *Estimation of Educational Achievement*

MTT is premised on the assumption that a student possesses an amount of skill and knowledge with respect to a construct or variable (termed the “latent variable” (Andrich, 1988, p14). Whereas in the physical sciences a variable is observable and can be measured, a latent variable is not directly observable and is inferred from other variables typically derived from mathematical modelling. The amount of the latent variable “possessed” by a student can be measured by the interaction of the student with items that represent increasing knowledge or ability in the construct of interest (e.g., mathematical ability). The amount of the latent variable is used to “locate” a student along a developmental continuum according to the amount of the construct possessed by the student.

In both MTT and IRT, the concept that defines the estimation of student educational achievement and progression of learning is quite simple. When students are confronted with a problem in which their skills and understanding of the latent variable are at a higher level than the cognitive demands of the problem then the students will most likely be able to complete the problem successfully. Conversely, if the cognitive demands of the problem exceed the ability of the student, then it is likely that the student will not be able to successfully complete the problem. In the case where the student’s level of skills and understanding are of the same relative level as the cognitive demands of the problem, the student has a 50:50 chance of solving the problem successfully.

IRT currently includes three basic models: the Rasch Model (RM), also known as the Simple Logistic Model (SLM)); the Two Parameter Logistic Model (2PL Model); and the Three Parameter Logistic Model (3PL Model). Each model attempts to estimate item difficulty and student ability in relation to the latent variable. The term “logistic” relates to the scales typically generated in logarithmic units (logits) that are used to change the response intervals from natural numbers to an ordinal scale in which the distance between units is a constant unit of measure, not an arbitrary observation of raw scores.

It is noted that criticism of the RM has led to the other two models (2PL, 3PL) in the family of IRT models. For reasons expanded upon below, it is appropriate to discuss the 2PL Model and 3PL Model in advance of the discussion of the RM, which is the ultimate focus of this research study.

3.2.2 *The Two Parameter Model (2PL Model)*

Birnbaum (1968) contended that in addition to the basic parameters of latent ability (β_j) and item difficulty (δ_i), items tend *not* to be of equal discrimination. To account for this, the 2PL Model introduces the discrimination parameter (α), as shown in Eqn 3.1. In the 2PL Model, Birnbaum (1968) modifies Rasch’s one-parameter logistic model (or RM) by allowing item parameters to vary, not only in terms of their difficulty (δ), but also in terms of their ability to discriminate (a) among individuals of varying ability.

$$\Pr\{x = 1; \beta, \delta, \alpha_i\} = \exp[\alpha_i(\beta - \delta)] / (1 + \exp[-D\alpha_i(\beta - \delta)]) \quad \text{Eqn 3.1}$$

where β is the latent ability of the student in the trait of interest;
 δ is the relative difficulty of any item that is an indicator of understanding in the trait of interest;
 α_i is the slope parameter – the discrimination – of item i ;
 D is an arbitrary scaling constant typically set to 1.7 to approximate results from the normal ogive model; and,
 $\Pr\{x = 1; \beta, \delta, \alpha_i\}$ defines the probability of a correct response given the parameters β, δ, α_i .

The model defined in Eqn 3.1 generates an item discrimination index, which is a point biserial correlation coefficient. Item discrimination indicates the extent to which success on an item corresponds to success on the whole test. Since all items in a test are intended to cooperate to generate an overall test score, any item with negative or zero discrimination undermines the test. Eqn 3.2 shows the formula for point biserial discrimination.

$$PB_c = (M_c - M) / S \sqrt{(P_c / 1 - P_c)} \quad \text{Eqn 3.2}$$

where M_c is the mean score of the students who answered the question correctly;
 M and S are the mean and standard deviation, respectively, of all the students; and,
 P_c is the proportion of students who answered the question correctly.

It is worth noting that although the 2PL Model does not explicitly account for guessing in its fundamental equation, its impact is in the way guessing biases the values of both M_c and P_c . In the 2PL Model, guessing inflates the value of M_c and by inflating the proportion of P_c it results in an overestimation of the value of the point biserial (PB_c). Hence, the 2PL Model suffers the same deficiency as the RM with respect to guessing.

The 2PL Model also suffers from a face-validity issue for large-scale assessment in which individual student achievement is reported to stakeholders. The inclusion of the discrimination parameter results for students who achieve the same raw score may lead to a different ability estimate because the ability estimate is a function of which items a student has answered correctly. Hence, the 2PL Model has been most commonly used in research programs that do not report individual student performances, but instead are concerned with aggregated results of sample populations.

Given the 2PL Model's inability to explicitly account for guessing and its challenges to implementation in large-scale testing programs, no further investigation of it is included in the research and analysis phases of this study.

3.2.3 The Three Parameter Model (3PL Model)

Birnbaum (1968) further developed the 2PL model by incorporating an item discrimination parameter, modelling the slope of the item characteristic curve, and introducing a lower asymptote parameter often interpreted as modelling “guessing” or “item guessability”. This model is called the Three-Parameter-Logistic Item Response Theory model (3PL Model), and it is defined in Eqn 3.3.

$$\Pr\{x = 1; \beta, \delta, \alpha_i, c\} = c + (1-c)\{\exp[\alpha_i(\beta - \delta)]\} / (1 + \exp(\beta - \delta)) \quad \text{Eqn 3.3}$$

where: β is the latent ability of the student in the trait of interest;
 δ is the relative difficulty of any item that is an indicator of understanding in the trait of interest;
 α_i is the slope parameter – the discrimination – of item i ;
 c is the quantification of a lower asymptote that reflects the probability of low ability students answering the item correctly: and,
 $\Pr\{x = 1; \beta, \delta, \alpha_i, c\}$ defines the probability of a correct response given the parameters $\beta, \delta, \alpha_i, c$.

The 3PL Model reflects the development of a specific set of parameters as a “best fit” of the data to derive a model to analyse the data. Interpretations of the alpha (α) parameter as item discrimination and delta (δ) parameter as difficulty have found general agreement in the field, but interpretation of the c parameter as guessing has generated considerable debate. Including a ‘ c ’ parameter in the model was Birnbaum’s (1968) approach to allow for statistical adjustment of item responses for the non-zero performance of low-proficiency students on multiple choice (MC) items. This approach involves the computation of a non-zero asymptote that reflects the impact of non-ability-based, random guessing by students who exhibit low ability in the trait of interest. Typically, the constant generated to reflect guessing in the 3PL Model is in the range $0 < g < \frac{1}{k}$, where k is the number of distractor options of the item.

In this respect, practitioners have generally termed the c parameter the “guessing” or the “pseudo-chance” parameter (Chiu & Camilli, 2012; DeMars, 2007; Harris, 1989; Socha & DeMars, 2013). The c -parameter estimates, however, typically tend to be smaller than the value that would result if examinees answered an item correctly by random chance (Lord, 1974), and consequently Hambleton et al. (1991) proposed the term “pseudo guessing parameter” as more appropriate for the c parameter.

The 3PL Model violates a fundamental aspect of measurement, namely, that there is a one-to-one relationship between the trait and the items. With the 3PL Model, the resulting estimate of ability in the trait is not independent of the items used to generate the ability. A measurement scale exists as an inviolate reference of the characteristics of a trait, and not a function of observations of specific data sets. In the physical sciences, it is not the practice to recalibrate the variable and associated scales with every measure.

In relation to the 3PL model, Smith (1993) comments

The measurement task is to differentiate between random guessing, which contains no information, and informed guessing, which contains some information. This cannot be accomplished by modelling guessing as an item or even as a person parameter. The pseudo-guessing item parameter in the three-parameter model is useless. It mistakes guessing as a function solely of the item, when, in fact, guessing is an interaction between item propensity to provoke guessing and person proclivity to guess. In addition, the parameterization of guessing obliterates differentiation between random and informed guessing. (p.262).

At the face-validity level for the reporting of individual student ability, the 3PL Model is constrained in the same manner as the 2PL Model because it generates results for which the same raw score may produce a different ability estimate expressed in the logistic parameter. Hence, public reporting is inherently problematic at the individual student level, and consequently it is generally recommended that individual results are not reported.

Psychometrically, there are three additional issues with the 3PL Model:

1. The guessing parameter (c_i) represents the expected ability of lower performing students and therefore typically there are fewer data points on which to accurately estimate its value. Consequently, this lack of precision leads to relatively large standard errors about each estimate (Thissen & Wainer, 1982). Research shows that the c parameter is poorly recovered in simulation studies. This is particularly the case in the easier items where lower performing students are more likely to answer correctly, and with poorly discriminating items, where guessing is more likely to be an influence (Lord, 1975; McKinley & Reckase, 1980). These factors detract from the 3PL Model's capacity to provide precision in the measurement of the trait.
2. The introduction of a guessing parameter has an unintended impact on the model because of its interaction with the discrimination parameter. By introducing a lower, non-zero asymptote, the slope of the discrimination is depressed. As the c parameter approaches $\frac{1}{k}$, the discrimination factor for the item is reduced, which introduces a double interaction on the item difficulty and student ability estimates. The consequence in relation to the estimate of student ability is to reduce the numerator in the model equation, which in turn reduces the estimates of the higher-ability students to a degree more evident than in the estimates of the lower-ability students.
3. The third issue is the simplification of the interpretation of the c parameter given its determination. The 3PL Model assumes random guessing, in which the response is independent of ability. Studies indicate that in real test situations this assumption does not hold, with students using various strategies to eliminate less plausible options or implement partial knowledge in the trait of interest (Emberson & Riese, 2000; Yen & Fitzpatrick, 2006), especially in cases where the c parameter estimates are at the lower extreme of the range $0 < c < \frac{1}{k}$ (Kinston, 1985; Waller, 1989). Alternatively, De Ayala (2013) showed that in cases where the value the c parameter exceeded $\frac{1}{k}$, it was common that the distractors functioned poorly, thus impacting significantly on discrimination. Obinne (2012) observed values of c in the MC items ranged from 0.09 to 0.50 using the 3PL Model. The mean value was 0.26, which approximates the $\frac{1}{k}$ logical value for random guessing, but in 28 of the 56 MC items the c factor exceeded 0.25. This is problematic, as it implies that in half the items there was successful guessing at a rate higher than the expected value of 25%.

These issues highlight the fallacy of the internal assumption implied by the 3PL Model that guessing is a variable independent of ability, and that the guessing parameter is applied universally, independent of student ability.

An overarching issue with the implementation of both the 2PL and 3PL Models in the analysis of data is that they are not constrained by a measurement theory, insofar as both attempts to derive a model from data. The central approach of each is to determine a combination of parameters that describe the data; that is, to create a model that is a best fit to the data. In this respect these models are not grounded in measurement theory, but rather are solutions that reflect a particular data collection. By extension, this concern draws into question the stability of the developed scales, as they suffer from the same issues as Traditional Test

Theory (TTT) analyses, with the scale being to some degree test bound. The differences in underlying assumptions and approaches to analyses between the 2PL and 3PL models and the RM approach are articulated in Table 3.1 (Wright, 1992). The RM rejects guessing, its presence being noted by the misfit of person ability estimates. However, the application of the Rasch model in authentic assessments is a function of mean misfit statistics for each person, not individual item/student interactions and, typically apart from noting the statistic abnormality, no action is taken to account for guessing.

Table 3.1
Comparison of the Properties of the 3PL and Rasch Models

Birnbaum	Model:	3PL	Rasch Model
For 2PL, For 1PL, set $a_i=1.7, c_i=0$	set	$c_i=0$	
Allan Birnbaum 1957/1968		Georg Rasch 1952/1960	
imitates data		defines measures	
contrived to fit observed MCQ ICC's		derived to construct scientific measurement	
$\log \left[\frac{P_{\theta i} - c_i}{1 - P_{\theta i}} \right] = a_i (\theta - b_i)$ <p>θ is the assumed, not actual, person sample distribution</p>		$\log \left[\frac{P_{ni}}{1 - P_{ni}} \right] = (B_n - D_i)$ <p>n is the actual individual person ability</p>	
$\sum_i a_i X_{\theta i} = \sum_i a_i P_{\theta i} \rightarrow \theta$ $\sum_{\theta} \theta X_{\theta i} = \sum_{\theta} \theta P_{\theta i} \rightarrow a_i$ <p>Shared a_i and θ causes $\theta \leftrightarrow a_i$ feedback: divergence unless constrained</p>		$\sum_i X_{ni} = \sum_i P_{ni} \rightarrow B_n$ $\sum_n X_{ni} = \sum_n P_{ni} \rightarrow D_i$ <p>B and D estimable separately: inevitable convergence</p>	
MCQ [1992: Eiji Muraki's Generalized Partial Credit Model]	dichotomies	only	any ordered observation dichotomy, rating, ranking, counting
guessing reliable item asset	accepted	c_i	guessing unreliable person liability
discrimination as a useful item scoring weight	variation	welcomed	a_i
discrimination as a misleading item-bias interaction	variation	rejected	
crossed natural item-difficulty-ordering is different for different persons	ICCs and is different for different	accepted unavoidable	crossed prevents item-difficulty-ordering is the same for everyone
			ICCs construct is the same for
			rejected validity

Source: www.rasch.org/rmt/rmt61a.htm

3.3 The Rasch Model

3.3.1 Introduction

As noted earlier, the RM is relevant in the Australian context because it is used in the national assessment suite of programs (NAPLAN, NAP_SL, NAP_ICTL, and NAP_CC). The RM is sometimes termed the Simple Logistic Model (SLM) due to its requirement that the interaction of item difficulty and student ability are the only factors that are relevant in determining the nature of the latent variable and the estimation of student ability in the latent variable.

In theory, the RM is fundamentally different from the 2PL and 3PL Models in that it embodies the property of specific objectivity (which will be defined later in this chapter). The RM is a mathematical extension of the following requirements that define the measurement properties of the variable:

- Item statistics are independent of the sample from which they were estimated.
- Student scores are independent of item difficulty.
- Item analysis accommodates matching test items to student knowledge level.
- Test analysis does not require strictly parallel tests for assessing reliability.
- Item statistics and student ability are both reported on the same scale.

These theoretical requirements of the RM generate psychometric and measurement advantages compared to classical models. For example, in contrast to the 2PL and 3PL Models, the RM is a mathematical derivation from a set of requirements grounded in defined measurement principles. The RM is applied to the data and the psychometric question is whether the data fit the model, that is, do the data comply with the measurement principles that dictate the appropriate use of this procedure.

This “simple” construct can be converted into the mathematical function defined in Eqn 3.4.

$$\Pr\{x = 1; \beta, \delta\} = \exp(\beta - \delta) / (1 + \exp(\beta - \delta)) \quad \text{Eqn 3.4}$$

where; β is the latent ability of the student in the trait of interest;

δ is the relative difficulty of any item that is an indicator of understanding in the trait of interest;

and

$\Pr\{x = 1; \beta, \delta\}$ defines the probability of a correct response given the parameters β, δ .

The denominator is a normalising function that ensures the resultant probability calculated is in the range $0.0 < \Pr(1) < 1.0$;

As proposed by Rasch (1960), this mathematical relationship defines the SLM, and it represents how data that conform to the model can be analysed and reported using the SLM. The unit of the RM is the Logarithmic Unit, which is called the “logit”. The mathematical implementation of the RM develops a logarithmic scale with equidistant units in the integer scale. If the data do not accord with or fit the RM, then the use of the RM is inappropriate as a reliable estimator of student achievement on the trait. This approach contrasts with those of the 2PL and 3PL Models, which “adjust” to maximise the fit of the data, and whose every analysis is defined by the set of data, not by a theoretical model that is consistently applied to a data set.

The RM is a mathematical model defined by Eqn3.4, and it therefore underpinned by a number of assumptions. In relation to measurement, these assumptions have been deemed by Andrich (1988) to be requirements of measurement. These requirements of the RM are discussed in detail below, but it should be noted that, in theory, once the item difficulty (when ordered from easy to hard) exceeds the ability of the student then there should be no credit achieved through guessing. The inclusion of guessed items corrupts the fundamental relationship between students' abilities, as demonstrated by responses to items of varying difficulty, and the relationship of the items to the trait of the investigation. Correct guesses impute unintended bias into the estimations of both student ability and item difficulty. In regard to student/item interaction, this problem is central to the current study. When the constraints of measurement theory are overlaid onto the RM, it is necessary to unpack the contributing factors and environmental conditions with which the RM must comply to allow these estimations to be meaningful, valid, and invariant over time. The following sub-sections discuss the features of the RM.

3.3.2 Features of the Rasch Model

3.3.2.1 A Model of Intent

The term Model of Intent (MOI) was proposed by Andrich (1986a) to make explicit the construct validity of an instrument. An MOI underpins the fundamental assessment construct and addresses the issue of content validity in the analyses of a set of data (Messick, 1987). The overarching concept is that the items included in the instrument to assess subject ability are manifestations of the trait of interest, and that the sum of the collection of items provides valid information regarding the interaction of the subject with the trait of interest. The characteristics that are evidence of increasing knowledge or ability in a particular trait can be used to construct a collection of items that capture the range of abilities that the students display/possess regarding the trait of interest. The Rasch Measurement Model is interested in the development of scales that reflect a measurement variable rather than modelling an existing data set. Rasch (1960/1980, 1966, 1968a) and Wright (1968) have strongly supported the relevance and advantages of objectivity in measurement. In describing the MOI, Andrich (1986a) made explicit the difference between modelling data and constructing a measurement variable:

In the Rasch approach, a theoretical position is articulated first ... the theoretical position is then transformed into variables ... and ... the transformation reflects the intention to construct a variable. The variable is constructed through a series of items or micro-replications that are intended, and expected, to hang together ... The model chosen is an expression of this intention, and therefore comes before any data are collected. The data are then collected and compared to the model, which is a mathematical rendition of the intention. If the data do not accord with the model then this is evidence that the data do not reflect the intention. (p. 46)

3.3.2.2 Measurement Objectivity

The RM requirement of measurement objectivity relates to the independence of the instruments used to measure the amount of the trait exhibited by the object of the measurement. This feature has previously been referenced as a fundamental requirement of measurement. For example, the latent trait of mathematical ability cannot be measured without some evidence of students' knowledge, understanding, and capacity to apply the fundamental knowledge in some tangible form. To achieve this, assessments are

developed with test items of varying difficulty and a range of scenarios that attempt to enable students to display manifestations of their ability in the trait. According to Wright (1968), there are two necessary conditions of objective measurement: “First, the calibration of measuring instruments must be independent of those objects that happen to be used for calibration. Second, the measurement of objects must be independent of the instrument that happens to be used for measuring” (p. 87).

3.3.2.3 *Separation of Parameters*

The principle that the parameters of student ability and item difficulty are independent is significant, both in relation to the measurement properties and to the estimation of ability in the trait. Rasch (1961) referred to this feature as “specific objectivity” and was explicit in articulating this feature as a necessary condition of measurement:

The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared. Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or some other occasion. (p. 331)

It is important to appreciate that although the item difficulty and student ability estimates are sample free, this does not suggest that the characteristics of the sample are totally irrelevant. Rasch (1961) relates this to the “class”, which pertains to the frame of reference of the test and, by association, the MOI. Andrich (1985) makes this relationship explicit: “Relationships need to be established with some specific frame of reference, the frame of reference included a definition of a class of persons, the class of items, and any other relevant conditions that would ensure objective relationships were maintained” (p. 44).

When considering the presence and impact of guessed responses in a test analysed using the RM, Tognolini’s (1989) comments on this characteristic of the model are salient:

Persons’ characteristics govern their responses to the items and their responses are used in conjunction with the model to determine the item difficulties. Since the locations of the items on the variable define the variable, it can be deduced that the definition of the variable results from the reaction of the sample of persons to the set of items. If the characteristics of the persons that interact with the items are changed, so is the definition of the variable (p. 53).

Andrich (1988) comprehensively covers the derivation of the mathematical logic demonstrating the separation of the parameters, so it is not elaborated upon here. Instead, the following commentary explains the relationship between item difficulty and student ability as derived in the RM analyses. This relationship is critical in determining which responses may reflect a guess.

A first estimate of relative difficulty of an item is the ratio of the number of correct responses compared to the total number of responses to the item. From this, an initial estimate of relative item difficulty is developed. In the estimation of item parameters, the matrix of the possible outcome space in two adjacent items provides two combinations that impart information regarding the relative difficulty of the two items, as follows:

Correct _{Item(i,)} Incorrect _{Item(i+1)}	$\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$	Incorrect _{Item(i,)} Correct _{Item(i+1)}
---------------------------------------------------------------	----------------------------------------------	---------------------------------------------------------------

In observing a student's response patterns, these outcomes are indicators of the transition in the ability-item difficulty interaction from within the ability range to beyond the ability range (the correct/incorrect-1, 0-pattern), or an indication of a zone of approximate equivalence of the student ability and the item difficulty in which there is a relatively equal chance of a correct response.

In reducing the conditional probability of the 0,1 or 1,0 events occurring, Andrich (1989) showed that the total score of an item (its facility) "is a key statistic containing all of the information about the difficulty of an item" (p. 75). The reduction of the probability of the 1, 0 event is shown in Eqn 3.5.

$\text{Prob}\{x_{v1} = 1, x_{v1} = 0 \mid r_v = 1\} = e^{-\delta_1} / (e^{-\delta_1} + e^{-\delta_2})$
 The reduction of the probability of the 0, 1 event is;
 $\text{Prob}\{x_{v1} = 0, x_{v1} = 1 \mid r_v = 1\} = e^{-\delta_2} / (e^{-\delta_1} + e^{-\delta_2})$

Eqn 3.5

where r_v is the total raw score of person v on the two items.

These conditional probabilities can then be expanded over multiple items and the estimates of the relative ability of each item maximised by an iterative process.

The significant issue in these observations is that the estimation of the relative item difficulties is independent of person ability. They do not contain any reference to β_v . Hence, if the data fit the RM, the relative difficulty of the items can be estimated independently of the distribution of the abilities of the students who have participated in the collection of the data. Andrich (1989) elaborated on the significance of this feature of the RM: "It means that the total score (of and individual) contains all of the information about the person, and if the data fit the Rasch LTT [latent trait theory], then all of the information about the ability β_v is contained in the total score" (p. 76).

This mathematically derived condition of the RM leads to the raw score being a sufficient statistic for student ability (which is consistent with Guttman's contentions, which will be elaborated on in Chapter 4).

3.3.2.4 *Raw Score as a Sufficient Statistic*

The fundamental requirement for the RM is that the interaction between a student's ability in the trait of interest and the relative difficulty of the item that represents a manifestation of the trait are the only factors that interact to determine the probability of a correct response to an item. This condition leads to the conclusion that all the information regarding a student's ability is contained in the response pattern and the summation of the positive responses to the individual items within the collection that are contributing to the variable. Hence, the achieved score on the set of items represents a "sufficient statistic" to estimate the ability of the subjects that are targets of the analysis in the trait of interest. The term "estimate" is intentional and reflects the measurement property of "measurement error", which is the degree of uncertainty about the precision of the measure due to the uncontrolled factors or natural influences.

Flowing from this requirement is an assumption of "no guessing", as guessing provides no information regarding the ability of the subject with regard to the trait of interest and thus corrupts the data as a whole by overestimating the observed raw score. The presence of guessing would therefore make the information regarding the student's ability less reliable.

3.3.2.5 *Determining Student Ability Estimates*

The raw score is determined as the sum of the probabilities of a given ability interacting with the set of items with the determined difficulty. Table 3.2 provides an example of the way an iterative process develops the relationship between the observed raw score and the ability estimate derived for each student in the test. Specifically, the relationship between raw scores and ability is a function of the interaction of the item difficulties with potential abilities. The allocation of an ability estimate for a student with a particular raw score, which can be achieved by any distribution of correct answers in the response pattern, is independent of the response pattern. It is a function of the item difficulties that are calculated for the items that comprise the test. Table 3.2 shows that a student's raw score of 3 is achieved by an ability estimate somewhere between -1.50 and -1.00. An iterative process applied within this range determines that a score of 1 is achieved when the ability estimates approximate -1.48 logits. The impact of a raw score being inflated due to guessing is evident in the table.

3.3.3 Measurement Principles Underpinning the Rasch Model

Section 3.2 explained the features of the RM. This sub-section elaborates on several underlying requirements that are fundamental to implementation of the RM for analysis and reporting of student data.

3.3.3.1 Unidimensionality

Associated with the concept of an MOI is the requirement of “unidimensionality”. The collection of items which make up the assessment instrument are required to provide evidence of increasing “possession” of the trait of interest and to contribute to the measurement of the trait. In addition, the items should function uniformly as manifestations of the characteristics that are observed in the trait. For example, if the trait of interest is Algebra, the collection of items that comprises the construct validity of the instrument must not include items that assess Calculus ability. The concept is relatively simple in theory; however, in practice it is challenging to achieve when, for example, there are factors such as the reading and language comprehension requirements inherent in Mathematics or Science items and differences in cognitive load across sub-domains. Yet this requirement is essential given that all of the items in a test contribute to the definition of the variable and to the subsequently determined item difficulties and student ability estimates.

3.3.3.2 Local Independence

The concept of “local independence” refers to the correlation of the items that comprise the instrument designed to measure the latent ability of students in the trait of interest. The local independence requirement demands that the items are only correlated through the latent variable. This implies there are no instances of the success of any item being dependent upon the success in a previous item, and that the parameters generated explain all the systematic difference among the data. Tognolini (1989) noted that a violation of this requirement would result in a variation of the definition of the construct: “Altering the sequence of items that constitute the test may alter the ordering of items and persons on the continuum” (p. 62). The mathematical derivation that defines this RM requirement is presented in Eqn 3.5.

$$\Pr\{X_{ni} = 1: X_{nj} = 1 | \beta\} = P(X_{ni}|\beta) \cdot P(X_{nj}|\beta) \quad \text{Eqn 3.5}$$

where β is the latent ability of the student in the trait of interest; and

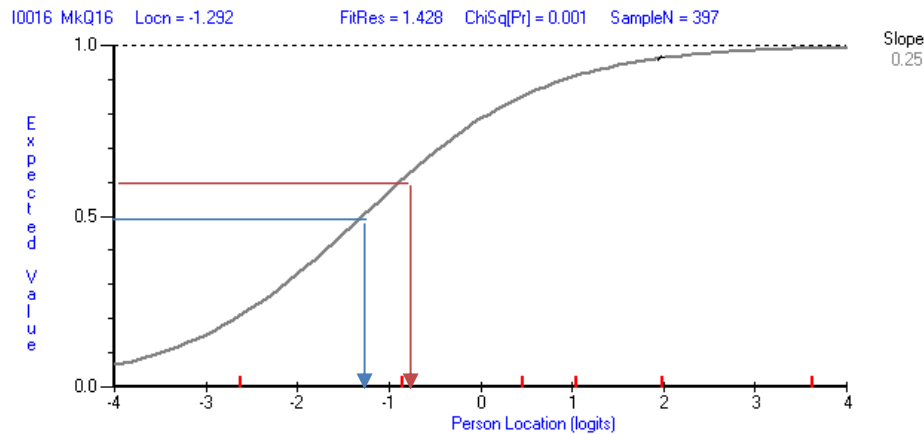
δ_n is the relative difficulty of any item that is an indicator of understanding in the trait of interest and the subsequent item in the test series.

3.3.3.3 Equal Item Discrimination

The RM requires that all items within a test set have similar item discrimination and differ only in relation to relative difficulty. Item discrimination in the RM is represented diagrammatically by item characteristic curves (ICCs). Figure 3.1 provides a sample of an ICC derived from the RUMM 2030 Rasch Measurement Model program (Andrich et al. 2013).

Figure 3.1

Example of an Item Characteristic Curve.



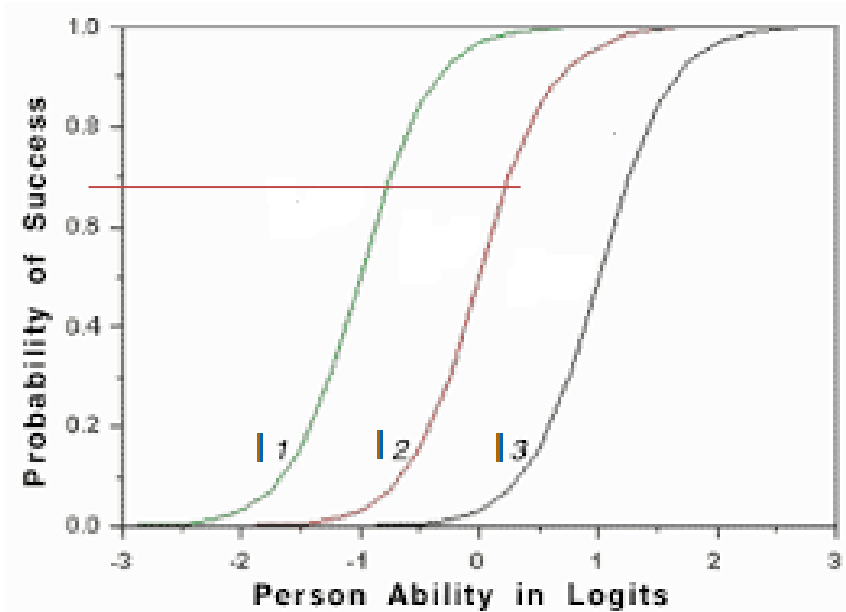
Source: Current Research Study 1, $SIM1.item16_GIPp=0.5$

Figure 3.1 shows that the difficulty of that item is -1.292. This value is determined by the intersection of the probability of a correct response being 0.5 as the Expected Value (50%) with the item ogive. As person ability increases, the probability of success on the item increases. A student with an ability of approximately -0.9 on this item/scale would have an approximate probability of 60% for a successful response. The slope of the ogive represents the defined discrimination of the item in the RUMM program.

The concept of equal discrimination is problematic in the relation to authentic data. The term “equal” becomes a relative concept with the emphasis on “sameness” within acceptable bounds. Ideally the family of ICCs of a test should comply with the example from Sick (2010) shown in Figure 3.2.

Figure 3.2

Example of Item Characteristic Curve for a 3-Item Test–Equal Discrimination

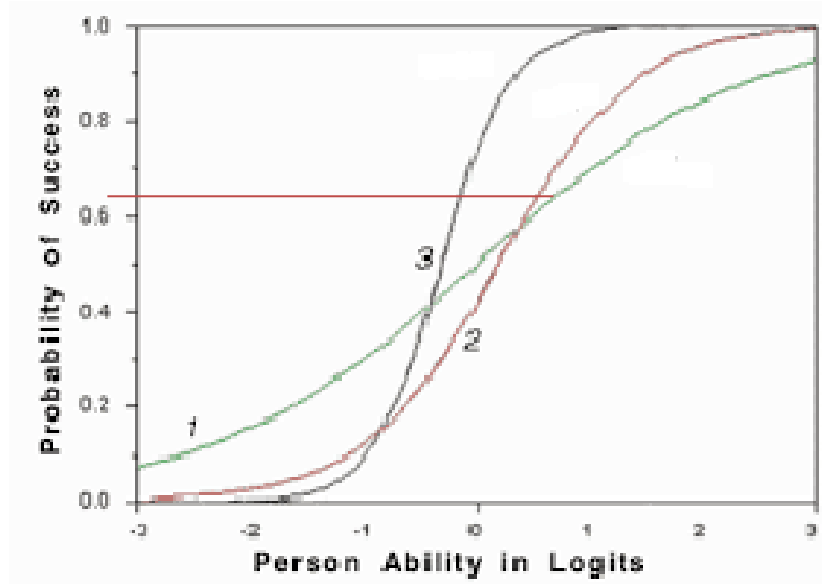


Source: Sick (2010) https://hosted.jalt.org/test/sic_5.htm

Figure 3.2 shows that the probability of a successful response on each of the successively more difficult items is a function of the student ability only. Violation of the requirement that items have equal discrimination may result in item curves such as those displayed in Figure 3.3 (Sick, 2010).

Figure 3.3

Example of Item Characteristic Curve for a 3-Item Test—Unequal Discrimination



Source: Sick (2010) https://hosted.jalt.org/test/sic_5.htm

Figure 3.3 shows the issue that arises from a test that comprises items of unequal discrimination. In tests that exhibit items of this type, success on higher discriminating items has a greater impact on the calculation of student ability than success on those of lower discrimination. As Masters (1988) notes: “When two persons have the same number of right answers the higher ability estimate goes to the student who succeeds on the more discriminating items” (p. 16.) This issue, and its distracting feature, is similar to that addressed in the earlier discussion of the 2PL Model.

3.3.3.4 No Guessing

A fundamental tenet of the RM is that the only factors influencing the probability of a student’s successful response to an item are the difficulty of the item and the ability of the student in relation to the trait of interest. In cases in which a correct response does not relate to ability in the trait of interest, a correctly guessed response not only provides no information about the true ability of the student, it also biases the data and confounds the estimations of items and hence the ability estimates of all participants in the test. As Zimmerman and Williams (2003) note: “Chance success due to guessing limits the reliability of multiple-choice tests ... for many multiple-choice tests, error variance resulting from guessing can be a larger than the error variance from other sources” (p. 367).

Yet this requirement is fraught. Given the symmetrical relationship between the estimate of student ability derived from the raw score and the item difficulties of the instrument designed to measure student ability, as shown in Table 3.2, there is direct impact of a correct guess on the estimate of student ability. For instance, Table 3.2 shows that two correct guesses cause a “true” score of 7 being reported as score of 9, which results in an increase in the ability estimate in excess of one logit in the scale of this test. In a large-scale test such as NAPLAN, this equates to approximately two years of learning in the subject.

Given the current state of the research into the resolution of this “guessing problem”, Humphry (2015) asserts: “There is currently no avenue for practitioners to apply Rasch models to standard multiple-choice items in a manner that accounts for guessing **and** maintains a focus on core measurement criteria” (p. 194, emphasis added). In response, this study has sought to develop a model that identifies and measures the impact of guessing to the calibration of the variable, and the ability estimates of groups of subjects. Specifically, the Guessing Indication Protocol developed and evaluated through this research attempts to address the consequence of this requirement of no guessing, which, when it is routinely violated, awards credit for correct guessing beyond the ability of a student and directly influences the estimate of ability attributed to the student.

3.4 Summary

This chapter has investigated the common analysis techniques that derive from the family of models that comprise Item Response Theory, and it has highlighted current alternative analysis procedures that purport to address the issue of guessing in student response data. In particular, it discussed the 2PL and 3PL Models and provided a rationale for refuting the use of these methods to address this issue.

The Rasch Model was presented in detail as the dominant model in cases in which the intention is to share results with participants, teachers, and the wider field of educational stakeholders by means of reports and diagnostic feedback. This is the case in Australia’s NAP suite of national assessments. The discussion highlighted the requirements of the RM and introduced a tension between the theory and the application of the model by using authentic data that may contain guessing.

This and the preceding chapter thus provide a framework and segue into the primary aims of this study, which are to

1. develop a model that takes account of the interaction between students and items at the item/student interaction level, but maintains the fundamental principles of measurement that underpin the RM; and
2. investigate the extent to which stakeholders are guessing by using inaccurate data in their decision making as a result of information that does not take account of guessing in a Rasch modelling environment.

The next three chapters investigate the ways in which guessed responses may be identified in student response patterns. The thesis then provides a methodology to account for guessing – the Guessing Indication Protocol – which aims to reduce the bias that guessing introduces to student data analysed using the RM.

Chapter 4

Methodology Overview

4.1 Introduction

The purpose of this chapter is to describe the data sources and the analysis methodology applied during this series of studies, and to outline the principles that underpin the Guessing Indication Protocol (GIP) algorithm that has been developed to indicate probable guessing in student responses. As outlined in Chapters 2 and 3, each phase of this research was evaluated in terms of conformity with the properties required of measurement and development of a scale.

4.1.1 Overview of Research Design and Phases

The overall study is quantitative in design and grounded in a combination of simulated and authentic data that were initially used to identify, through the parameterisation of item/student interactions, the characteristics that typify a response with a high probability of being a guess. Those parameters were then applied to two sets of authentic data by conditioning student responses to account for identified guessing. The outcomes were evaluated using tests of fit and reliability to measure the impact of conditioning the data to account for identified guesses. Table 4.1 summarises the analysis phases, purposes, and means of evaluation implemented at in phase, the specific analyses for which are elaborated in this chapter.

Table 4.1

Summary of the Analysis Phases of This Research

Phase	Purpose
1. Initial (INIT)	Determine initial Rasch parameters of student ability and item location to determine the probability of a correct student/item interaction and evaluate misfit.
2. Guessing Suppressed (GS)	Re-code defined (Simulations)/self-identified (Field data) guesses to determine impact on item difficulty (and student ability) of re-coding indicated guessed student/item interactions.
3A. Guessing Indication Protocol (GIP3A) – item re-calibration	Re-calibrate item locations using developed GIP and evaluate impact of re-coding items indicated as guessed in the calibration of item difficulty compared to the INIT values.
3B. Guessing Indication Protocol (GIPINIT3B) – student ability estimate re-calibration	Re-calculate GIP student ability estimates using GIP3A item locations with the INIT student data (INIT_RS) and evaluate the differences in student ability estimates compared to the INIT values.

Each phase contributed to the development of a conditioned data set in which the bias imputed by guessing was reduced. Specifically, Phase 1, the INIT phase produced initial Rasch statistics with item location and student ability estimate as the statistics of interest. Phase 2, the GS phase produced a conditioned set of parameters, with the difference in the item locations and student ability estimates indicating the extent to which guessing had biased the results.

Phase 3, the development of the GIP was completed in two stages. The first stage (GIP3A) determined the interface between the probability of a correct response of an individual student/item interaction and the actual guesses in the simulation data, and thus a threshold that was consistent in indicating a known guess. Having resolved the threshold parameters, these were applied to individual student/item responses, and those indicated as probable guesses were suppressed. Revised item locations were generated with this conditioned data set. The second stage (GIP3B) used the revised item locations determined in GIP3A with the original data set to assess the impact of suppressing probable guesses from the analysis. The research methodology involved three related studies to develop, refine, and evaluate the GIP. These are summarised in Table 4.2.

Table 4.2

Summary of the Data Investigated and the Analysis Phases Performed on Each Study

Simulations			Analyses Phases			
Study 1: Simulated data	Target data design	Name	INIT	GS	GIP3A	GIP3B
Simulation data developed to determine optimal parameters to identify defined guessing in data. Parameters then applied to small-scale field data and large-scale cohort assessments	Normal distribution n(S):400; n(I) 40 Small data set n(S):200; n(I) 20 Large data set n(S):1000; n(I) 40 Test too easy n(S):400; n(I) 40 Test too hard n(S):400; n(I) 40	SIM1 SIM2 SIM3 SIM4 SIM5	√ √ √ √ √	√ √ √ √ √	√ √ √ √ √	√ √ √ √ √
Authentic data sets			Analyses Phases			
Study 2: Small-Scale sample	Target data design	Name	INIT	SIG	GIP3A	GIP3B
Small-Scale sample; Mathematics	Convenience sample n(S):303; n(I) 40	Year 5	√	√	√	√
Small-Scale sample; Mathematics	Convenience Sample n(S):187; n(I) 45	Year 7	√	√	√	√
Study 3: Large-scale data	Target data design	Name	INIT	GS	GIP3A	GIP3B
Large-Scale data; Mathematics	Cohort Population n(S): 26000; n(I) 18	Grade 4	√	x	√	√
Large-Scale data; Mathematics	Cohort Population n(S): 21000; n(I) 23	Grade 8	√	x	√	√
Large-Scale data; Science	Cohort Population n(S): 26000; n(I) 22	Grade 4	√	x	√	√

Note: x indicates not implemented

4.2 GIP Principles

The aim of the GIP procedure was to determine the threshold limits of student/item misfit that can be effective in identifying a highly probable guess, so that data could be conditioned to remove this source of bias and recover more accurate item difficulty and student ability estimates. The basic “building blocks” that underpin the development of the GIP procedure are outlined below:

- The probability of student success on an item is a function of the relative difficulty of the item and the ability of the student.
- The value of the probability of success for an item for a particular student can be calculated using the Rasch model with known estimates of item location and student ability.
- The Rasch model predicts the probability of success on an item for a range of ability levels.
- When a student whose ability is less than the difficulty of an item correctly responds to the item, the result is misfit of the individual student/item interaction.
- The quantum of misfit is the standardised square of residuals between the observed response and the “recovered” response according to the model, after taking account of the student ability and the item difficulty estimated from the RM.
- The degree of the misfit can be used to identify students with an abnormal response to a particular item, with these parameters used to indicate probably guessing.

The GIP was developed from the analysis of sets of parameters observed to be strong indicators of the defined guessed responses in a student response pattern. Simulated data were first generated with the intent of providing a known source of a student/item guessed responses and, by interrogation of the parameters of these data, to develop the GIP. To verify the efficacy of the parameters, and have confidence in their generalisability, a series of structured simulated data was used to cover a wide range of distributions of item difficulties, a range of student abilities, and the kurtosis and skewness of those abilities in different item sets. These variations were intended to cover the range of the typical outcomes that would ensure representativeness across most of the large-scale assessment response patterns that occur in practice.

Approximately 0.5% of “noise” was included and randomly assigned to student responses in each dataset to represent the carelessness or misunderstanding that is common in large-scale assessments. The noise reflected incorrect responses in the “expected correct” regions of student responses. The inclusion of “noise” in the data was to overcome the purely deterministic structures of Guttman, which are typically problematic using a probabilistic analysis model. In addition, about 1% of unexpected correct responses were randomly included from lower-ability students to reflect some experiences outside the formal teaching and curriculum experiences of those students.

4.3 Analysis Methodology Overview

The analysis methodology was hierarchical and implemented using Bayesian methods. The data from each component of the overall study were imported into SPSSv25 (2019) to enable initial coding of the responses. For the simulated data, the analyses applied respectively to each set were as shown in Table 4.1: “Initial (INIT)”, “Defined Guessing Suppressed (GS)” and data conditioned by “the Guessing Indication

Protocol (GIP3A and GIP3B)". In the school-derived, small-scale field responses, data streams were analysed in the "Initial (INIT)", "Student Self-Identified Guessing" (SIG) and conditioned by the "Guessing Indication Protocol (GIP3A and GIP3B)" phases. The large-scale cohort data involved only an INIT collection, and the data conditioned by the GIP3A and GIP3B, as there were no defined guesses indicated by the participants.

SPSSv25 generated descriptive and psychometric statistics (e.g., mean, standard deviation, reliability), and conduct significance testing (t-test) for group differences of means. It was also used to recode student responses and prepare data in the formats used by the Rasch analysis programs for each treatment. RUMM 2030 (Andrich et al., 2013) and Conquest 4.14.1 (Wu et al., 1988) were used to prepare Rasch statistics, including a confirmation of the traditional statistics extracted by SPSS. Item fit statistics, student statistics, and item/student residuals were extracted using RUMM 2030.

4.4 Participants

4.4.1 Study 1: Simulated Data

Five sets of simulated data were generated to cover a range of differing statistical properties: sample size, test length, item difficulty, central tendency, distributions, skewness, and kurtosis. These data five sets provided a variety of structures to represent the generalisability of the protocol developed. No students were involved in the generation of the simulated data. In each simulated dataset the "student" response patterns conformed to a large extent with the Guttman scale.

4.4.2 Study 2: Small-Scale Field Data

A convenience sample of 490 school students from NSW was used for the field study. The total sample comprised 303 students from Year 5 and 187 students from Year 7. The Year 5 students comprised 157 females (52%) and 146 males (48%) from intact and unstreamed classes at four suburban and one regional primary school. The Year 7 students comprised 68 females (36%) and 119 males (64%) from intact, mixed ability classes at two suburban high schools.

4.4.3 Study 3: A Large-Scale, System-Wide Sample

The large-scale data for Study 3 were collected from a standardised, mandated Mathematics and Science assessments implemented in the Abu Dhabi jurisdiction of the United Arab Emirates. These assessments are administered to the full cohort of students in Grades (Years) 4 and 8. Each administration of the assessments comprised a population sample in excess of 20,000 students, which meant no sampling bias was present in the data collected. Throughout this thesis these data are referred to as "large-scale" and "authentic" data, as they represent data collected from system-wide achievement assessments conducted under consistent and system-defined administration processes.

4.5 Data Sources

4.5.1 Study 1: Simulated Data Generation

The simulated data of Study 1 were generated in Microsoft Excel to the specifications outlined in Table 4.1. These specifications provide a variety of distributions from which to determine the initial parameters used to identify the pre-defined guessing in student responses. The use of simulated data in which the guesses are defined gives a certainty to the development of parameters in item/student interaction that indicates a true guess. This diverges from other approaches (e.g. the 3PL model) that impute an overall guessing parameter to the item.

Simulated data sets were constructed with varying item/student interactions, ranging from 1,000 students with 40 items to 250 students with 20 items, and with targeting varying from a test which was too easy through to one which was well targeted and too hard. The response structures were designed to provide sufficient variation in item difficulties and simulated student abilities. This sample size also allowed some disaggregation of student ability estimates into groups to enable comparison of differences in the quantum of variation in parameter estimates for different ability groups (Andrich, 2012; Marias, 2015).

All simulated data reflected a multiple-choice item set with four distractors and a single correct key. The construct of these data was consistent with the items ranging from relatively easy to harder, with targeted facilities ranging from 20% to 90%. In the easy items there were relatively few items defined as guesses (on average about 2.5%). In the harder items, approximately 20% of the responses were defined guesses (25% guessed of 75% incorrect responses). These figures are approximate due to the randomisation algorithm, which produced variable quantities with each iteration.

4.5.2 Study 2: Small-Scale Sample Data Collection

This small-sample case study was conducted with “live” data to assess the efficacy of the protocol developed using the simulated data. It involved fieldwork with a sample of students in Years 5 and 7 – the NAPLAN cohorts – attempting Mathematics tests. This study again included only MC items, each with four options and only one correct answer. Mathematics was the subject used in the analyses of these field data and large-scale authentic data in subsequent studies.

The tests were developed using items that were pre-calibrated using items that had previously been used in a large-scale assessment, hence the relative difficulty of the items that comprised the test, and their psychometric properties, were already known. Only items that had functioned with acceptable item statistics and overall test characteristics in the large-scale administration were included in the sample tests. The Year 5 tests consisted of 40 multiple choice items ordered from easy to hard (as determined by the item location in the previous administration). The Year 7 tests consisted of 44 multiple choice items also ordered from easy to hard.

Consistent with the curriculum experiences of each academic year, the tests comprised a selection of items from the subdomains of Number, Geometry, Measurement, Chance, and Data.

For each Year level, two versions of the test were produced from the initial English version: one in English and the other in Arabic, each with the same items. The students participated in both versions of each test. This provided a baseline estimate of student ability and item difficulty with the English version of the test – effectively a “control” group. The Arabic version not only provided an “intervention” group in which the guessing was self-identified, it also generated statistics to compare with the baseline English version values. The tests were first administered to the students in the Arabic language, with the content and contexts of the items involving varying degrees of language dependence. In some cases, the items were language independent. In others, the context or processes needed to solve the items were familiar to the students. The remaining items were language dependent and, since none or very few of the students understood Arabic, the responses were generally either guessed or omitted. To gain further insight into whether students were guessing their responses, the students were asked to colour-code their responses as follows:

- *green* if the student believed they had answered the item correctly using their knowledge;
- *yellow* if the student had used partial knowledge or a process of elimination; and
- *pink* if the student had guessed their answer.

The same students were then administered the same items in English and were asked to use the same colour-coding system for each response. The English versions of the tests were designed to

- obtain a better estimate of student ability in a common language medium; and
- assess the veracity of the Arabic data by investigating the responses to non-language dependent items compared to the baseline statistics provided by the English versions of the items.

4.5.3 Study 3: Large-Scale Cohort Data Collection

The large-scale data study used data collected from a student achievement program grounded in the Abu Dhabi national curriculum. To ensure anonymity of student performance, students were de-identified by a unique system of IDs and, apart from gender, no other identifying information was provided. The test procedures and implementation were conducted under the authority of the Abu Dhabi Education Council and grounded in the curriculum expectations of Grades 4 and 8 at the time of the tests.

The tests were constructed by professional test developers and subjected to the rigours of a national assessment in regard to test framework, review protocols, final test assembly, and checks. The tests used a combination of MC items and constructed response (CR) items. The CR items were omitted from the analyses, and the data set re-calibrated using only the MC items. The tests comprised 22 or 23 MC items delivered in an on-line format, with 40 minutes allowed for completion. Due to the differences in the online resources available in the schools, some variation in administration procedure was expected, although all tests were expected to be conducted under the mandated administration protocols.

4.6 Procedures

4.6.1 Study 1: Simulated Data

The simulated data were structured in a Guttman pattern with a variety of correct responses, depending upon the nominated relative ability of the “student”, followed by the remainder of the responses being incorrect for each “student”. Each data set was then “corrupted” by introducing defined “guessed responses” at a random rate of approximately one in four ($\frac{1}{4}$), which is in the regions of the item/student response patterns that Guttman determines as incorrect for an individual. The random “correct guesses” were generated by a randomisation algorithm with four categories (7 through 10). One category of these random data was then defined as a “guess”. For convenience, the value “9” was selected to define random guessing within the regions of student responses beyond their ability. The remaining values in the “incorrect regions” were coded as incorrect responses.

4.6.2 Study 2: Small-Scale Field Study Data

Ethics approval for Study 2 was provided by the University of Wollongong’s Human Research Ethics Committee (HE15/450) and the NSW Department of Education’s SERAP process (2015296), which is required for research involving NSW schools. Working with Children clearances were also obtained for all invigilators, as required by the NSW government. To recruit schools, letters were sent explaining the intent of the study and the instruments to be used. Consent to participate was indicated by approval from the school principal and the teacher, after which the information was distributed to relevant families to seek parents’ or carers’ written consent and the student’s verbal assent.

The tests were administered in class groups in accord with the test’s administration guidelines: first with the Arabic administration and then the English version immediately after a short break. The tests were administered in pen-and-paper format, with students additionally responding to items using coloured markers as described above. Students were given up to 30 minutes to complete each version of the test, and typical examination conditions were followed.

4.6.3 Study 3: Large-Scale Study Data

As the assessments for Study 3 were implemented as a component of the Abu Dhabi Education Council’s assessment program, no ethics approval required for data collection. Permission was sought and received from the Director of Research and Assessment of the Abu Dhabi Education Council to conduct and report investigations specific to the identification of guessing in student response patterns, and for using item response time as a potential indicator of guessing. The permission allowed use of both the analyses and the de-identified data.

4.7 Plan for Analysis

Each of these five simulation data sets were analysed using the Rasch analysis technique according to the processes and sequences outlined in Table 4.1.

4.7.1 Study 1: Simulated Data

In the simulated data, all sets had defined “guessed” responses. The analysis plan applied to simulated data were those outlined in Table 4.1, with the exception that the plan for Study 1, the Simulated Data, used defined guesses formulated with the intent of determining GIP. The plan and sequence for analysis had three phases.

Phase 1 was the “Initial” (INIT) analysis. In the INIT analysis all the defined guesses were treated as correct responses, ignoring the fact that they were defined a priori as guesses. The item, student, and item-student interaction parameters from the INIT analysis of each simulated dataset were then extracted to determine an appropriate set of parameters that would indicate the defined, embedded guesses in these data.

As the INIT estimates were produced from data sets that were analysed without accounting for guessing, “defined correct guesses” were included both in the calibration of the item difficulty and the student raw score, and hence the student ability estimate. The measures extracted from the analyses included:

- INIT person ability (logits);
- INIT item difficulty (location in logits); and
- reliability statistics (Cronbach’s alpha).

The Rasch model was used to calculate:

- individual item/student probability of success; and
- individual item/student residuals based on the success of the student on the item.

Phase 2, the GS phase, involved a process in which the defined guessed items were re-coded as missing data to estimate the impact guessing had on the calibrations of item locations and student ability estimates. GS estimates were produced from the data sets in which the defined guesses were re-coded as missing data in the calibration stage, and the student raw score was reduced by the number of “correct guesses”, with consequent impact on the calculated student ability estimate. GS-conditioned data were analysed to determine revised item difficulties and student parameters, together with the item-student interaction parameters. These data represented “accounting for guessing”, as the guessing response had been defined a priori in each data set.

Differences in outcomes of the INIT and GS phases were determined with the parameters relating to items that had been generated in the INIT analysis compared to those that were generated from the GS analysis for responses a priori defined as “guessed”. These parameters were interrogated to develop a beta model for the identification of guessed items in the response patterns of students of differing ability ranges. The three parameters found relevant in differentiating items that were defined guesses were the location of the item; the calculated probability of the interaction; and the calculated misfit of a correct guess compared to

the expected outcome for the interaction of the student ability with the item of that difficulty. Following a series of iterations and interrogation of the simulated data, the GIP parameters were determined. This process is elaborated on in Chapter 5.

Phase 3 had two parts. Part 3A involved first using the parameters from the Phase 1 data to instigate a re-code of the specific items that had failed GIP indicators, determined from the simulated data, and indicated as probable guesses. The revised GIP was implemented to suppress indicated guessing in the calibration of item difficulties. This phase of the analysis (GIP3A in Table 4.1) was used to re-calibrate the item locations to be used in the GIP3B process, which was the re-calibration of student ability estimates.

In the next step of Phase 3A, after the items identified as “highly likely guesses” were re-coded to “missing data”, the conditioned GIP item sets were analysed to extract revised test performance, item statistics, and individual student ability estimates. These results were extracted for comparisons with the INIT outcomes and the GS outcomes of Phases 1 and 2. Comparisons were then made between “actual” guessed responses, as defined in the Phase 1 data set, and the guesses identified by the GIP process to determine the efficacy of the process in respect of “recovery rates” (where recovery rate reflects a comparison between the percentage of defined guesses in the simulated data and the percentage indicated by the GIP procedure).

Finally, Part 3B involved using the raw scores from the Phase 1 data with the re-calibrated item locations derived in Phase 3A. The conditioned GIP item sets were re-analysed to extract revised test performance, item statistics, and individual student ability estimates to reflect the impact of the GIP. In this final phase of the GIP evaluation, the guessed items were scored as correct (score for a guess = 1). This process provided an estimate of the degree to which ability was misreported in analyses that did not account for guessing, while maintaining the face validity of scoring probable guesses as correct. This version of the GIP was considered by the researcher to represent the reported ability of the student on the trait if guessing was taken into account.

4.7.2 Study 2: Field study

The data collected in the field study were analysed using the same processes outlined in Study 1. This involved an INIT phase, a SIG (self-identified guess) phase, and the GIP3A and GIP3B processes. The second phase of the analysis of the English version of the test took account of the colour-coded responses. This was the Self-Identified Guess (SIG) analysis. The efficacy of the proposed GIP model was assessed by comparing the success of the GIP model in identifying the item/student interactions that were indicated as guesses in the student data.

The data from this study were analysed initially with the English form of the test, taking into account the student answers without reference to the colour coding to determine a set of baseline data. The parameters obtained from these analyses were compared to the item parameters of the original Arabic tests from which the items were extracted in relation to the order of difficulty and the fit statistics (INIT). The items had performed uniformly with their original calibrations that had been used to determine the test content and item order.

The responses to the Arabic version of the tests were subjected to the same analysis phases (INIT, SIG, and GIP) for the purpose of comparison with the baseline data extracted from the English versions. These analyses were designed to provide further insight into the relationship between student ability and their propensity to guess. The analyses of the results from the Arabic version of the tests were predicated on three “known” parameters:

1. the relative mathematical ability of each of the students was “known” from the English version of the test.
2. the relative difficulty of each item was “known” from both the initial large-scale calibration statistic and the analysis of the English version of the instrument; and
3. there was knowledge of the degree to which each item was language dependent (i.e., there were no clues, and the item was totally encoded in Arabic).

The intent of this component of the study was to provide further insight into the response patterns of groups of students of differing ability when guessing was an expected and self-indicated outcome.

4.7.3 Study 3: Investigations of a Large-Scale Data Set

In the analysis of the large-scale data, an INIT, GIP3A, and GIP3B analyses were implemented using the final threshold parameters developed for the GIP process. Ability estimates from the initial analysis were generated and, following the implementation of the GIP procedure, new estimates were generated to provide an indication of the impact of guessing.

The MC items of the data sets were analysed in the same manner as the simulated data and the field trial data, with an INIT analysis ignoring response patterns. No GS analysis was possible with the large-scale data, as there was no process to identify or define guesses. In the case of the large-scale data, the second iteration of the analysis involved implementation of the parameters of the GIP model developed in phases GIP3A and GIP3B. Tests of fit were completed, and the individual and group results before and after data conditioning were evaluated to assess the efficacy of the process and the final outcomes of accounting for guessing using the GIP process. These are elaborated on in Chapter 7.

4.7.4 Investigation of Student Response Time as a Parameter in the Identification of Guessing

The technology associated with online assessments allows additional (unplanned) response time data at the individual item/student level to be collected alongside the item response data. Consequently, for this research it was possible to investigate average item response time across a cohort by item (using aggregation of individual response times) and by interrogation of the individual responses.

The researcher considered it plausible that the interaction between individual and cohort response times might give some indication that there was a factor (e.g., guessing) impacting a student’s response to a particular item. These instances were considered in association with the cognitive load required to engage with the stimulus, review the options in MC items, and perform calculations or operations demanded to answer the item. It was expected that there would be some variation in student responses to individual items depending upon item complexity, word count, item location within the test, and the ability of individual

students. However, it was reasonable to assume that a response pattern, in relation to item answer and item response time, might be observed for a student, and that when taken in tandem with their ability, it would give some insight into the level of engagement and commitment of a student at the individual item level. Response time data thus potentially provided an additional factor that could be incorporated into the model to identify guessing.

Chapter 8 details these investigations in which the response time for a student was well below the average response time of the cohort or ability group of the student, yet a correct response was recorded. These parameters were considered in conjunction with GIP parameters to determine if the timing parameter could be used to better identify a “guessed” item.

4.8 Summary

This chapter has outlined the overall methodology, the data sources, and the characteristics of the data used to develop and evaluate the GIP procedure. The methodology includes the logic used to address the problem of accounting for guessed items in a Rasch analysis. The major deviation of this methodology from those of previous researchers is that it addresses the identification of guessing at the individual item/student interaction level and maintains the complete set of response data. Previous methods either generalised a guessing parameter over the population or reduced the data set by removing responses that were deemed to be beyond the ability of students.

Chapter 5

Study 1: Analysis of the Simulated Data

5.1 Introduction to the Chapter

In Study 1, the simulated data were controlled with a specific definition of which student/item interactions were guesses. This allowed observation of results before and after controlling for guessing as a baseline for determining and refining a Guessing Indication Protocol (GIP) that could subsequently be evaluated in “live” data (see Study 2 in Chapter 6, and Study 3 in Chapter 7). The purpose of this chapter is to describe the process by which this GIP procedure was developed and initially evaluated. Specifically, this chapter provides a rationale for using the Guttman structure to develop Study 1 data and describes how the simulated data were generated. The conduct of Study 1 followed the methodology outlined in Chapter 4, and the plan for these analyses is elaborated in this chapter. The results of the analyses are then presented, summarising the several iterations of analyses that informed the development of the GIP, which is the subject of further evaluation in Study 2 and Study 3. The chapter concludes with an interim discussion of the outcomes of Study 1 and a definition of the GIP developed and used throughout the remainder of the thesis.

5.2 The Guttman Structure as a Starting Point for Investigations

5.2.1 Introduction to the Section

Until the late 1900s the dominant theory that underpinned test construction, multicomponent assessment instruments, and student achievement testing in the educational, social, and behavioural science domains was Classical Test Theory (CTT). CTT measures the reliability of the test, as determined by the difficulty of the test items and the score achieved by individual test-takers. The development of CTT dates to the work of Spearman (1904b), Fisher (1921), Thurstone (1927), and Cronbach (1951), and is typically represented by the equation:

$$\text{Observed Score } (X_j) = \text{True Score } (T_j) + \text{Error } (\epsilon_j) \quad \text{Eqn 5.1}$$

The development of CTT is grounded in three conceptualisations:

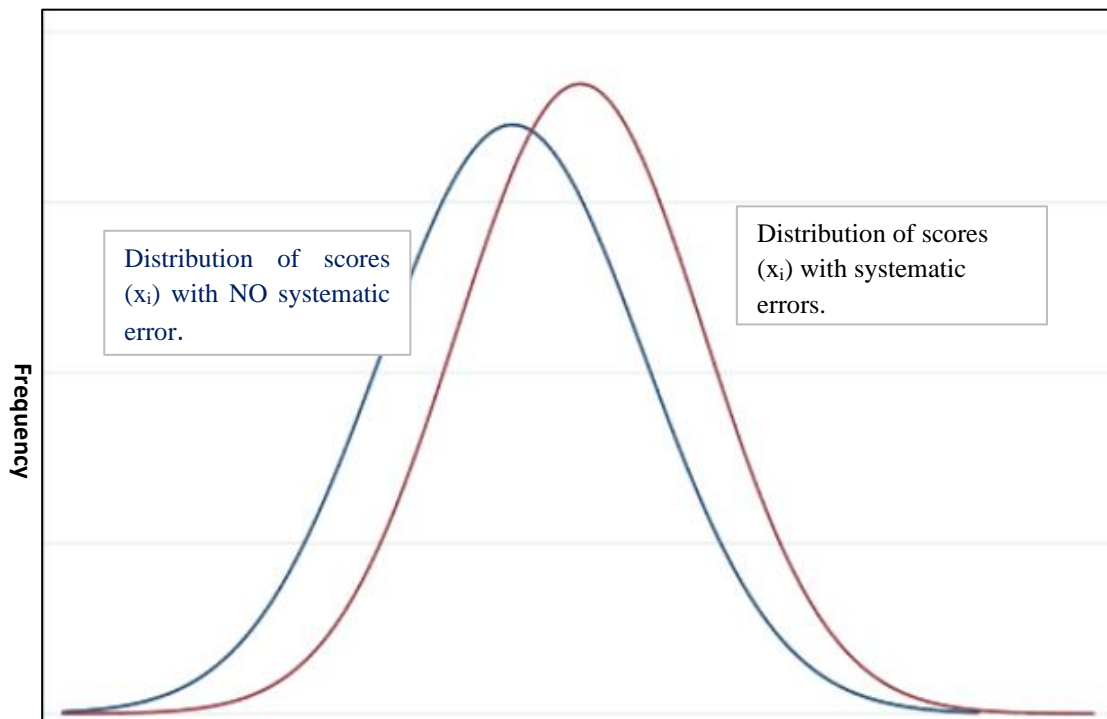
1. the acceptance that there are errors in measurement;
2. these errors are random variables; and
3. there are correlations between items that contribute to a test, and these correlations can be indexed.

The general initial aim of CTT is to minimise the impact of the errors by understanding and improving the reliability of the items used to measure the ability of students in the trait of interest. The source of the error in trait measurement may include systematic errors that reflect misclassification in the measures and random errors that may accrue due to arbitrary noise, careless mistakes, and/or misconceptions. Random errors that impact on individual test scores are independent of the individual test-taker and of the trait of interest, and hence will sum to zero over the entire sample of test-takers. Systematic errors will engender a positive or negative misclassification across the individual test-takers and the overall performance of the entire sample.

Although systematic errors will not affect the distribution of the scores (X_j) of test-takers, they will affect the average of the aggregated scores as represented in Figure 5.1. Given the assumptions regarding the relationship between item difficulty and student ability in the source of guessing, an error that accumulates due to guessing is considered systematic. The existence of systematic errors in the calibration of the parameters that contribute to the estimation of item difficulty, and consequently the quantification of student ability, contaminates the quality of the variable of interest and, by extension, imputes inaccuracy into the reported results of the students.

Figure 5.1

Impact of NOT Accounting for Errors on the Distribution of a Set of Raw Scores



Note. Adapted from Marias (2015, p. 125).

5.2.2 *The Guttman Scale*

Guttman's (1944, 1950) significant contribution to the development of measurement theory and the evolution of Modern Test Theory (MTT) was the conceptualisation and subsequent coding of the interrelationships between items and their difficulty, and between test-takers and their ability, along with the concept and construct of unidimensionality in the measurement of the trait of interest. The Guttman scale arose from an appreciation that although the reliability of an item in a test is a significant consideration in determining its contribution to the measure, it is not sufficient in determining if the total set of items are unidimensional. That is, item reliability provides a necessary, but not sufficient, indication of internal scale consistency and unidimensionality of the scale. The relationship between internal consistency and unidimensionality was investigated by Cronbach (1951) and Lumsden (1957), each of whom discussed the concept of a "pure factor" or "same thing" as a concept of unidimensionality, which is a feature central to the concept of measurement. Guttman's (1944, 1950) contribution to MTT stem from his work with sets of dichotomous test items that were initially designed to order a sample of students and position their ability

on a derived scale. His work advanced the concept of a unidimensional scale on which items and students could be placed in relative, comparable positions that are invariant for the set of conditions and variables. He developed a set of requirements that are implicit in measurement theory, the use of which generates data complying with conditions of the Guttman scale. This led to the concept of “items operating in the same direction” being superseded by “items belonging to a scale that measure the same dimension” (Guttman, 1950, p. 185). This concept is fundamental to the principles of “measurement”.

For a given data set, there are significant similarities between the conditions that satisfy the requirements of the Guttman scale and the requirements of the Rasch model (RM), which underpins this research. Andrich (1985) notes: “The Rasch models and the Guttman scale are both constructed **a priori** to the data whereas their respective rivals are governed more by the data” (p. 36, emphasis added). That is, a principal aim of CTT and some Item Response Theory models is to fit data to a bespoke statistical outcome to explain the data, rather than to conceptualise a measurement model, collect data, interrogate the results, and from that ascertain if the data fit the conceptualised model; “The Rasch models are consistent with the Guttman scale because they are generated from the same conditions and requirements for unidimensional scaling of measurement” (p. 36). Andrich also notes that the conditions in common with the RM that satisfy a Guttman scale include dichotomous responses, reproducibility, unidimensionality, data reduction, and invariance. The requirements of the Guttman scale (1950) will now be introduced.

5.2.2.1 Dichotomous responses

The use of dichotomous response formats provides a simple vehicle for the definition and implementation of a scale, and for providing a data structure on which unique items and students can be located on the conceptualised variable. The data structure allows for the ordering of items from easiest to hardest in an unequivocal structure based on percent correct as the measure of relative difficulty. The ordering of item difficulty, and the implicit assumption regarding the ability of the student who answers the item correctly, define that the knowledge required to answer all previous items has been mastered.

5.2.2.2 Reproducibility

Guttman (1950) makes reproducibility of scale development explicit in the perfect correlation between individual item responses and the total score. The item pattern is defined by the total score for each test, which provides the deterministic characteristic of a Guttman scale. This condition is highly problematic in the real-world application of student achievement tests and is a point of divergence with IRT models that calculate the probability of a positive response and a “likely” response pattern rather than a “certain” response pattern. However, in the simulated data of Study 1 the Guttman scale provides certainty in the relationship between the response pattern and the achieved score.

5.2.2.3 Unidimensionality

Unidimensionality is a requisite element of data for the valid use of the RM. Guttman (1950) defined the condition of unidimensionality as “a single continuum as a series of items each of which is a simple function of the scale scores permits a clear-cut statement of what is meant by rank-order based on a single variable” (p. 154). Given the relationship between the items, and the defined dependency of the correctness of item x being contingent upon a correct response to item $x-1$, the pure Guttman response pattern typically displays a high reliability index (KR-20 and/or Cronbach α), which indicates that all items are contributing uniformly to the trait and hence a high degree of unidimensionality.

5.2.2.4 Data reduction

The Guttman scale is cumulative in that if a student answers a given item correctly, the student must have correctly responded to all preceding items. For dichotomous items, the number of acceptable responses (scale items) is the number of possible item responses (item \times categories), less the number of items (incorrect responses) + 1 (all incorrect). The item combinations satisfying the Guttman scale are defined in Eqn 5.2:

$$\text{Number of scale items} = (n \times 2) - n + 1 \quad \text{Eqn 5.2}$$

The total of all possible item response combinations for a set of dichotomous items is 2^n , where n is the number of items in the set. For example, for a three-item test, the scale item combinations (assuming the items are ordered by difficulty) are: Scale items $(3 \times 2) - 3 + 1 = 4$ (see Table 5.1); and the possible total number of response combinations is 2^3 , being the scale items plus the following non-scale items = 8 (see Table 5.2).

Table 5.1

Guttman Scale Responses for a Three-Item Test

Item 1	Item 2	Item 3	Score
0	0	0	0
1	0	0	1
1	1	0	2
1	1	1	3

Table 5.2

Non-Guttman-Like Responses for Three-Item Test

Item 1	Item 2	Item 3	Score
0	1	0	1
0	0	1	1
0	1	1	2
1	0	1	2
1	1	0	2

For a 40-item test, the number of scale items is $(40 \times 2) - 40 + 1 = 41$ scale item combinations. The total number of possible response combinations is 2^{40} . Andrich (1985) notes:

The reduction of the data to scale types according to the total score is a major feature of the Rasch models (the raw score is a sufficient statistic) and provides an important link between the Guttman and Rasch scaling principles. (p. 41)

5.2.2.5 Invariance

The concept of invariance relates closely to the fundamental principles of the RM. This concept concerns the relative ordering of students on a defined dimension, irrespective of the items that have been used to define the test-taker's position on the scale of interest. Hence, irrespective of different sets and subsets of items that make up a unidimensional scale, the order of students on each of these tests will be invariant.

The major deviation between the conditions that define a Guttman scale and the requirements of a Rasch scale is the deterministic nature of the Guttman scale. Although both models contend that the raw score is a sufficient statistic to determine student ability estimates, the Guttman scale defines a specific item order and student response pattern required to achieve a particular score. This defines the deterministic nature of the Guttman scale. Rasch, on the other hand, defines an achievement scale as a probabilistic matrix in which the raw score is a random variable that can be achieved by several different item/success combinations.

Although the concept of a Guttman scale has declined in popularity due to its deterministic properties, the theoretical underpinnings specifically link item difficulty and student ability in determining student response patterns. According to Andrich (1982),

The ideal of a Guttman scale is difficult to achieve in real testing, the main obstacle being the requirements that the responses of a student to an item is governed in a determinable way. ... The realizations of the Guttman scale is enhanced if the items have a large spread in difficulty and no two items are close together on the scale. ... The probabilistic counterpart of the ideal Guttman response pattern is the simple logistic model (SLM) of Rasch. (p. 96)

In developing the SLM, Rasch ... in fact presents a pattern of ideal results for ordering students and items which take the Guttman form. The responses of students to items which conform to the SLM conform to the Guttman scale in terms of probabilities. ... Consistent with the SLM's being a probabilistic counterpart of the Guttman scale, if item difficulties are spread greatly, then the responses generated according to the SLM will reveal a Guttman pattern. (p. 96)

Guttman's scale was utilised in Study 1 to derive a practical, theoretically based initial data set that allowed guessing to be defined and identified in a Guttman-like data set. This provided certainty regarding the impact upon item parameters and consequent student ability estimates from applying different approaches to account for guessing. Specifically, the use of a Guttman-like data set was appropriate as a starting point as it represents "a limiting case of the probabilistic SLM pattern" (Andrich, 1982, p. 95). Hence, given the similar requirements of the Guttman scale and the Rasch measurement scale, as outlined in Chapter 4, it

was deemed appropriate to derive, evaluate, and refine an initial GIP using simulated data that essentially follows a Guttman scale, with defined relationships between item difficulty and person ability.

In the simulated data sets the Guttman-like data structures were corrupted by introducing randomly generated guessed items to develop the potential indicators of systematic guessing in a student response pattern. The following sections elaborate on these methods.

5.3 Elaboration of the Methods

5.3.1 Development of Simulated Data Algorithm

The intention of this research was to develop a protocol that is efficient and effective in identifying highly probable guesses in student responses. The plan outlined in Chapter 4 proposed that this could be achieved by evaluating the misfit in the student/item responses and suppressing those interactions that were considered indicative of a probable guess. Having suppressed the likely guesses, a new set of item parameters could be generated and used to develop a revised set of student ability estimates. Comparisons could then be made between the initial item locations and student ability estimates and the revised values to determine the impact of accounting for probable guessing in the estimates of student ability.

Aligned with this intention, this chapter describes the development of sets of simulated data in which the guessed responses are defined such that comparisons can be made with certainty over which items in the student response pattern were guesses to provide confidence in the developed protocol.

5.3.1.1 Overarching Method for the Development of the Simulated Data

The simulated data were created to serve two main purposes:

to produce a data structure that was effectively Guttman-like in design and provided the benefit of a defined relationship between item difficulty, student ability, and response regions in which items would be incorrect; and

to introduce defined “guessed” responses into those regions in which student responses would be incorrect in a Guttman structure.

The data structure thus had the following four features:

a number of students ranked from more able to less able, in a Guttman-like scale, based on a raw score;

a decreasing number of correct responses per student to reflect a population ordered by decreasing student ability;

a corresponding increase in incorrect responses to reflect a population ordered decreasing ability; and

a corruption of the zones of incorrect responses by randomly including correct guesses, which inflate the raw score of the lesser ability students.

The fundamental principles underpinning the Guttman scale are students will have varying degrees of ability in the trait of interest, that it is possible to construct a unidimensional scale of items that contribute to the measurement of the trait of interest, and that the items comprising the testing instrument can be ordered by difficulty. The dichotomous result (1,0) of the interaction between the test-taker and individual items is a function of the ability of the test-taker and the difficulty of the item. Hence, if the relative difficulty of item(i) is lower than the ability of the student(j), the response to the item(i) will be correct for student(j). Conversely, when the student's ability is exceeded by the difficulty of the item(i) the test-taker(j) will respond incorrectly to the item. The matrix in Table 5.3 is an example the Guttman scale generated by these basic principles.

Table 5.3

An Example of the Guttman Scale: Six items, Seven Test-Takers

	Easier			Harder			
Person	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	X_j
Case 1	0	0	0	0	0	0	0
Case 2	1	0	0	0	0	0	1
Case 3	1	1	0	0	0	0	2
Case 4	1	1	1	0	0	0	3
Case 5	1	1	1	1	0	0	4
Case 6	1	1	1	1	1	0	5
Case 7	1	1	1	1	1	1	6
Facility	85.7%	71.4%	57.1%	42.9%	28.6%	14.3%	
Increasing Difficulty \longrightarrow							

According to Guttman (1950), the student score (X_j) not only provides the estimate of the relative ability of the test-taker, assuming the validity of the items selected to measure the trait, but also defines the response pattern of the dichotomous answers to each item. For example, within this pattern, a score of 4 can only be achieved by a response pattern of 1,1,1,1,0,0 in relation to this test with the items ordered by increasing difficulty. The model is deterministic, with all the information about the student being defined by the score.

Correct guesses were introduced in the regions indicated in Table 5.3 that would be defined as incorrect responses (scores of zero) in a pure Guttman pattern. These are systematic in that they are in the region beyond the ability of the student. The introduction of these correct guesses was by the implementation of a randomisation algorithm in the "incorrect" region as shown by the values of '9' in Table 5.4. These values were defined as correct guesses.

To accommodate the paradox between purely Guttman-like data and the RM being a stochastic data response pattern, the following two minor variations were introduced to the simulated data patterns (see Table 5.4 for a demonstration of this structure):

1. Some random incorrect responses were included within the determined "correct" zone of the Guttman response pattern to reflect careless mistakes and/or misunderstandings of some able students with respect to individual items or concepts.

2. A “zone of uncertainty” was also introduced to the data. This was a region at about the intersection of the correct/incorrect boundary at which it is reasonable to expect that an individual item/student response has about a 50% probability of being correct (incorrect). This zone of uncertainty translated into random correct/incorrect responses (randomly coded 5, correct, and 6, incorrect) for each student at the boundary between the determined correct and incorrect response pattern.

Table 5.4

An Abridged Example of the Structure of the Simulated Data

Person	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	X _j
Case 1	0	0	0	0	0	9	1
Case 2	5	0	9	0	0	0	2
Case 3	1	5	0	0	9	0	3
Case 4	1	1	6	0	0	9	3
Case 5	1	1	1	9	9	0	5
Case 6	1	1	1	0	0	9	4
Case 7	1	1	1	1	0	0	4
Case 8	1	1	1	1	5	0	5
Case 9	1	1	1	1	5	9	6
Case 10	1	1	1	1	1	1	6
Obs Facility	90.0%	80.0%	80.0%	50.0%	50.0%	50.0%	
True Facility	90.0%	80.0%	60.0%	40.0%	30.0%	10.0%	

Note. Table 5.4 shows correct responses that are a function of ability coded as 1; responses in the zone of uncertainty coded as 5 and scored as correct (1); mistakes in the zone of uncertainty coded as 6 and scored as incorrect (0); and correct guesses in the zone of uncertainty coded as 9 and scored as correct (1).

The Observed (Obs) Facility is the sum of the items scored correct divided by the total number of responses, expressed as a percentage. The True Facility is the sum of the items scored correct, less those correct due to guessing (9), divided by the total number of responses. X_j is the raw score that would be reported, after inflation by the correct randomly generated guesses.

Table 5.5

Summary of Intended Targeting and Observed Traditional Statistics for Each Data Set

Simulation	Target	Cases	Max Score	Raw Score M (SD)	True Score M (SD)	Introduced Guess M (Total)
SIM 1	Normal	400	40	24 (8.7)	19 (11.1)	4 (1771)
SIM 2	Marginal easy	250	20	14 (3.8)	13 (4.8)	1 (360)
SIM 3	Normal	1000	40	24 (8.1)	19 (10.5)	4 (4304)
SIM 4	Easy	400	40	27 (8.5)	22 (11.1)	3 (1311)
SIM 5	Hard	400	40	20 (7.6)	14 (9.7)	5 (2164)

Note. Raw Score is the mean of the sample. The standard deviation about the mean for the sample is represented in the brackets. True Score is the mean of correct answers when all correct defined guesses have been removed. Introduced Guess is the mean number (per student) of defined random guesses (9) within each simulated data set. The number in brackets is the total number of randomly generated guesses introduced to each simulated data set.

For the reported outcomes of the simulated data to be generalisable across a wide range of distributions, five variants of the analysis of the data are provided. Since the emphasis in Study 1 was the student/item response interactions, the simulated data reflect a wide range of response patterns (see Table 5.5).

- The rationales that underpinned these five reported simulated data sets were, respectively:
- SIM1 was a normally distributed data set with a linear distribution of item difficulties, including a sufficient number of items (40) and a sufficient number of students (400) to generate stable item and student parameters.
- SIM2 was a data set with a variation in student numbers and test width, resulting in an insufficient number of items (20) and possibly too few students (250) to generate stable item and student parameters.
- SIM3 was a normally distributed data set with a wider separation between item difficulties, with a sufficient number of items and a larger sample of students to generate stable item and student parameters.
- SIM4 was developed as a negatively skewed distribution, with sufficient items that tended to be too easy for the sample students and hence there would be an expectation of a lower proportion of guesses in these data.
- SIM5 was developed as a positively skewed distribution, with sufficient items that tended towards too difficult for the sample students and hence there would be an expectation of a large proportion of guesses in these data among the more difficult items.

5.3.2 Plan for Analysis (PFA) of the Simulated Data

Using Rasch analysis, the following three primary factors were considered relevant to student guessing:

- the score of the student on the item (1,0);
- the ability of the student in the trait of interest; and
- the difficulty of the item in the scale of the trait of interest.

The result of the student/item interaction was determined by referencing the student response to the correct key defined in the test specification. All multiple-choice (MC) items had a dichotomous value: incorrect (0) or correct (1).

The analysis of the simulated data was a 3-step process:

- Each of the simulated data sets was first analysed using traditional techniques to review the facility performance of the items in situations in which defined guesses are ignored (current convention). These analyses sought to demonstrate the extent to which guessing could influence reported scores.
- Each simulated data set was then analysed using the RM, as described in Phases 1 and 2 of the Plan for Analysis (Table 4.1), which initially took no account of defined guesses (INIT). The data were then conditioned with the defined guesses recoded to produce revised parameters that accounted for the guesses, the Guessing Suppressed analysis (GS), to inform the development of the GIP. The purpose of this analysis was to determine the extent to which accounting for the defined guesses in each data set impacted on the item difficulty locations and the relative ability estimates of the students represented in the simulated data.
- The third step followed on from the overarching presumption guiding this proposed approach to account for guessing: the parameters of misfit and probability of a correct guess might provide threshold values

that accurately and consistently indicate high probability of a guess. Observations of the differences in the item locations, the probability of a successful outcome for each student/item interaction, and the value of the misfit residual were collated to determine reliable and defensible thresholds that could be assessed for their efficacy in identifying the defined guesses.

5.4 Results for the Simulated Data

5.4.1 PFA Step 1: Preliminary Results of Observed Item Performance in the Simulations

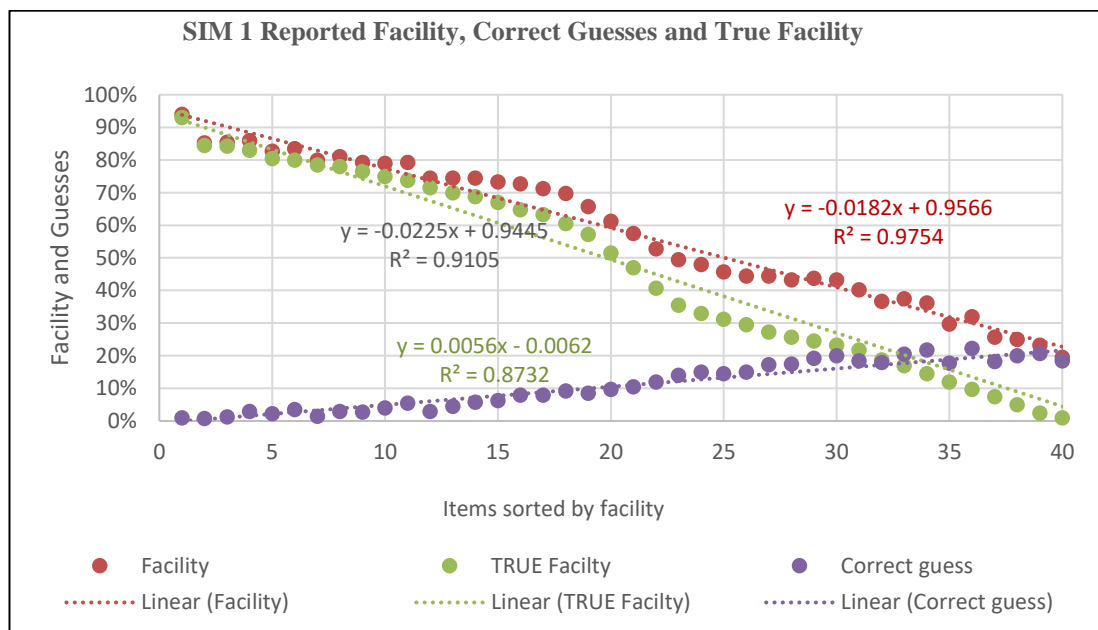
Results of the first step in the analysis confirmed that there was a strong relationship between item difficulty and the proportion of guesses as difficulty increased. This demonstrated the extent to which guessing could influence reported scores. These results are reported separately for each data set in the sub-sections that follow.

5.4.1.1 SIM1

Figure 5.2 shows the relationship between the proportion of correct random guesses and the true facility (correct responses minus guessed responses) for each of the 40 items. Specifically, the difference between the observed facility (red dots) and percentage of correct guesses (purple dots) is the true facility (green dots). For example, for Item 10 approximately 3% of the responses were correct guesses, the observed facility was 80%, and the true facility was approximately 77% (80%-3%). The inconsistency in these proportions, as indicated by the deviation from the trend line, reflects the random nature of the generated successful guesses in the simulation.

Figure 5.2

SIM1 Comparison of Facility With Correct Guesses



Note. The pattern of item facilities shown in Figure 5.2 follows the expected pattern as hypothesised in Chapter 1. The value of R^2 – the co-efficient of determination of the random guessing variable – indicates that 87% of the variation about the mean could be explained by the facility in the regression analysis.

In examining the full SIM1 dataset (detailed in Appendix A), the analysis of the mean true facility of the full test (47.2%) indicated that 52.8% of the responses were incorrect. When compared to the observed mean facility of the full test (58.3%), the impact of the correct random guesses was an overestimation of the mean facility of the test by 11.1%. The scale of the misclassification of the correct random guesses was of the order of 25% of the number of true incorrect responses, which was the expected rate of random guessing in a 4-distractor multiple choice item pattern.

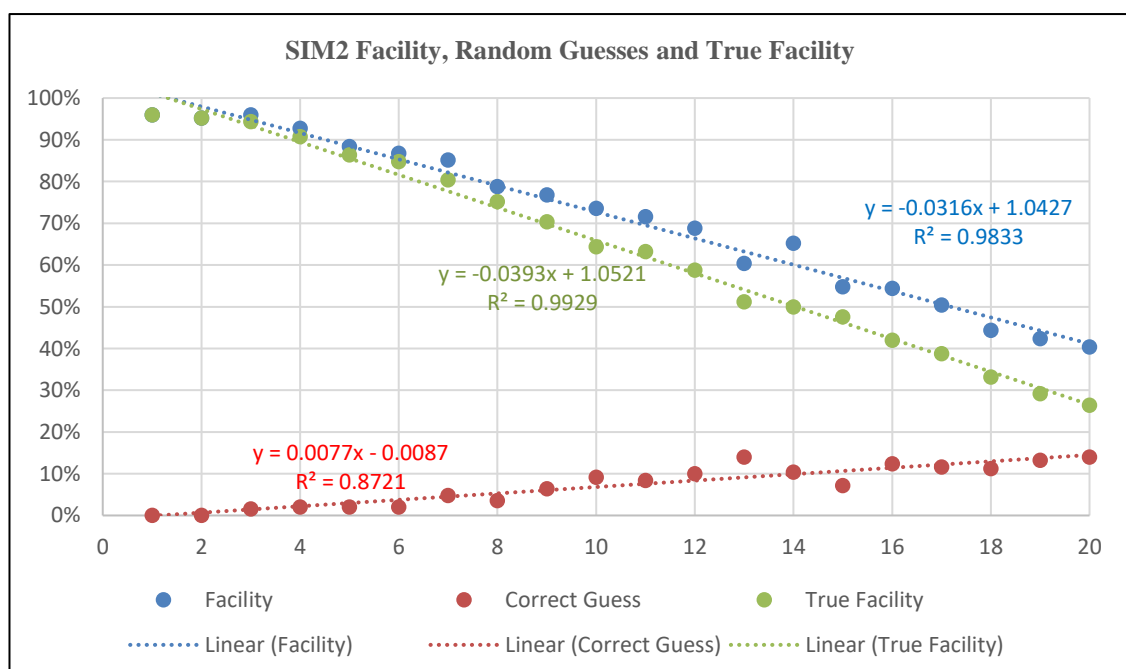
The dataset created for Simulation 1 confirmed the expectations in terms of rate of guessing and demonstrated the potential impact of reporting student results when guessing is not accounted for, given that scores were inflated by an average of 11.1% over the population.

5.4.1.2 SIM2

Simulation 2 was derived to provide a different sample size and true score distribution than that evaluated in SIM1. The mean of the true item facilities was 64%, which was the normal target for a test appropriate to a lower age group assessment. The mean of the observed item facilities of the overall test was 71%, indicating a difference from true facility of 7%. This represents an inflation due to guessing of approximately $\frac{1}{4}$ of the proportion of true incorrect responses. As observed in Figure 5.3, high correlations exist between the coefficients of determination of the item order (ordered from easy to hard) and the facility rate of the item as shown in Figure 5.4. These high rates are also shown in the rate of guessing and the facility of the item.

Figure 5.3

SIM2 Comparison of Facility With Correct Random Guesses



Note. The dataset created for Simulation 2 confirm the expectations in terms of rate of guessing and also demonstrate the potential impact of reporting student results when guessing is not accounted for. For this smaller, more homogenous set, the scale of the inflation of the scores was marginally reduced.

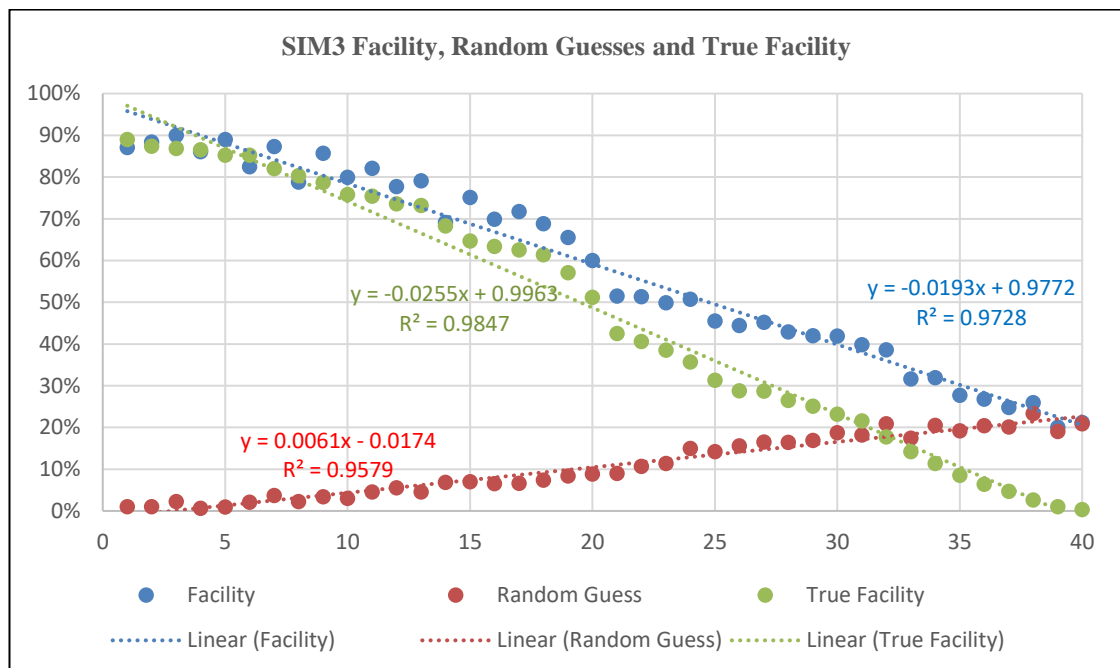
5.4.1.3 SIM3

Figure 5.4 reflects a data set with a significant increase in the sample size. The averages are 58.2% and 47.4% for observed and true facility rates, respectively, which indicates the raw scores were inflated by the random guessing misclassification by 10.8%. This value is very close to the expected value of 10.5%, ($\frac{1}{4}$ of 42%), indicating that the data are consistent with the design construct. The values of the R^2 (co-efficient of determination) are also closer to a value of 1. This result suggests that the relationship between item difficulty and the instance of guessing was more determined in this larger sample.

The results of SIM3 are consistent with those of the earlier simulations, with an indication that larger samples tend to conform more closely to the expected outcomes of the hypothesis.

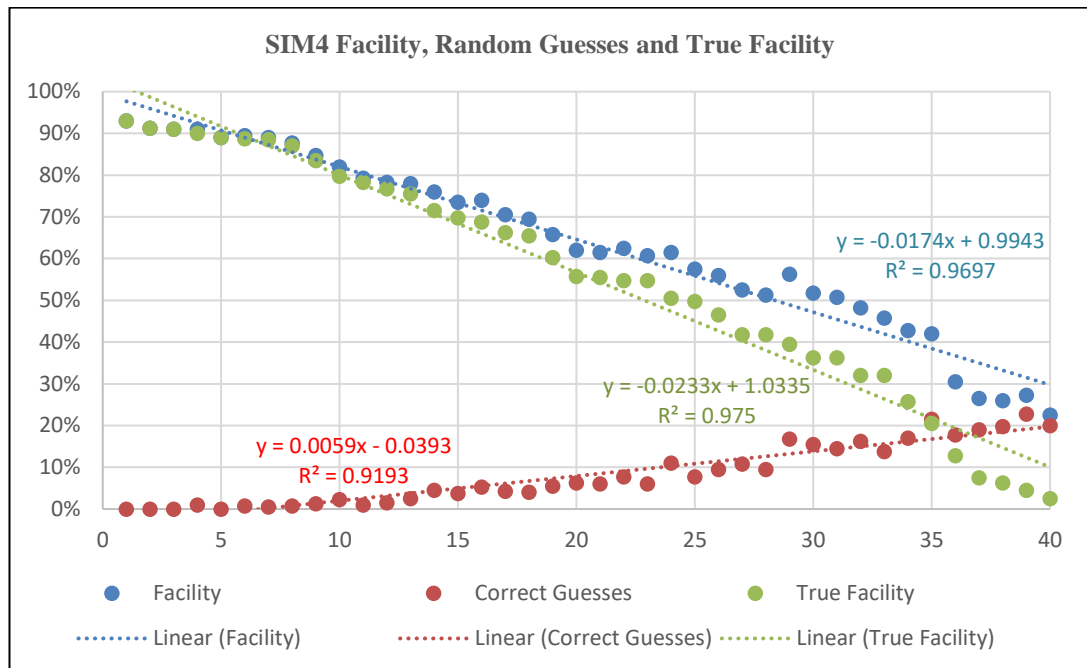
Figure 5.4

SIM3 Comparison of Facility With Correct Random Guesses



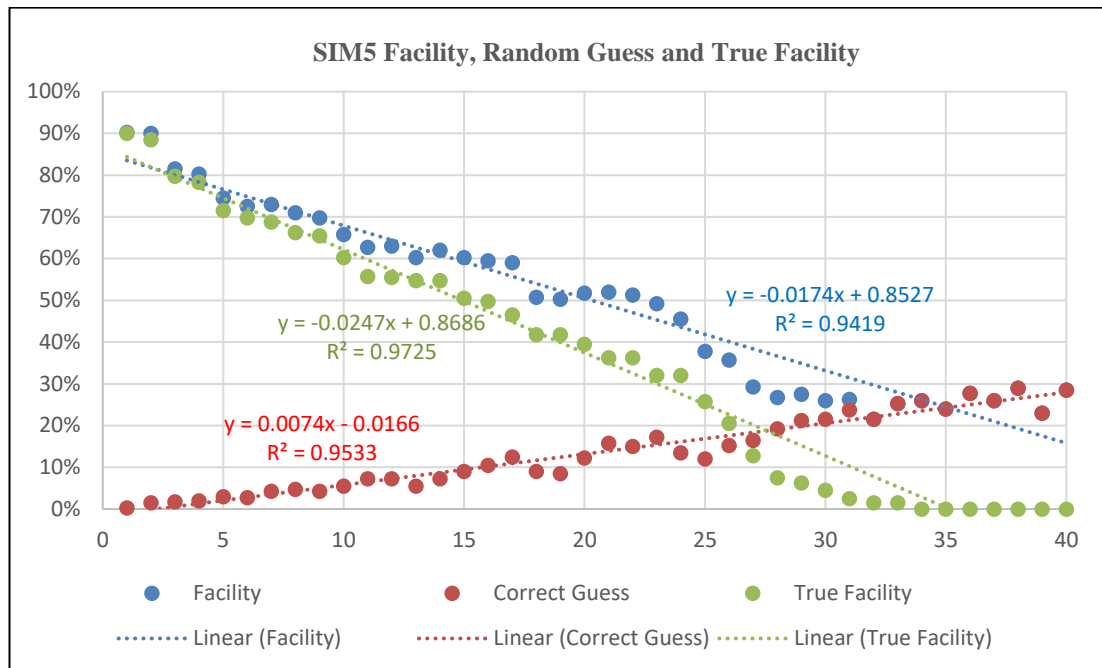
5.4.1.4 SIM4

Figure 5.5 displays the relationships for a data set that was designed to be relatively easy for the target group. It shows the equivalence of the facility rates for the true and observed values for the easy items, in which random guessing is negligible. As such, this sample design (an easy test) has relatively low impact on the R^2 coefficients compared to the previous sample simulation data designs. This statistic resulted from the higher number of easy items in the data, in which there were fewer instances of guessing to contribute to the co-efficient of determination. These data result in an observed average facility of 63.7%, with a corrected true facility of 55.5%. The misclassification due to the correct random guesses is 8.2%, which is in accord with the expected value of 9.1%. These results confirm that when a test is too easy for the target group there are fewer instances of guessing.

Figure 5.5*SIM4 Comparison of Facility With Correct Random Guesses*

5.4.1.5 SIM5

Simulation 5 was designed as a dataset that was too difficult for the target cohort. The average observed facility rate was 49.7%. Relatively little guessing was observed in the first 10 items, each of which had high facility rates (>70%). However, the SIM5 dataset was constructed with relatively few correct responses in the final six items, which was designed to simulate the discrimination in ability and knowledge in the higher region. Figure 5.6 displays the effect of a test which is too difficult for the target cohort. The facility of the final six items represents correct guesses, with the rate of success approximately 25% in each case. The randomness of the generated data delivered a range of up to 28% in observed randomly generated guessed items when both the random code (9) and the 'zone of uncertainty' code (5) were combined.

Figure 5.6*SIM5 Comparison of Facility With Correct Random Guesses*

The previous commentary described the methodology implemented to produce the simulated data and report the traditional statistical outcomes derived from the planned initial analyses those data. Those analyses sought to evaluate the impact of “known” guessing and the potential to extract parameters from “defined guessing” within the simulated data. The results provided confidence in these data as a baseline for the investigation of indicators of guessing in data in a Rasch analysis environment

In each of Figures 5.2 through 5.6 it was noted that the R^2 is higher for the true facility rate than for the observed facility rate. This confirms that the removal of random guessing improves the reliability of the data set. The analysis of distributions shown in these figures supports the overall contentions regarding the relationship between item difficulty and the proportions of random guessing in the simulated data.

The simulated data were “perfect” data sets in which the random guessing pattern was controlled. Hence the resulting outcomes represent an extreme example of “perfect” knowledge of student responses and its relationship to the trait of interest. Clearly this is unobtainable using real world data. However, within Study 1 the aim of the simulations was intended as the catalyst for generating a solution to the problem of guessing in large scale assessments.

5.5 Rasch Analyses to Inform the Guessing Indication Protocol (GIP)

5.5.1 Pfa Step 2: Introduction

The second step in the planned analyses involved an initial Rasch analysis (INIT) of each of the simulated data sets, followed by an analysis in which the defined guessing was conditioned in the data by recoding these instances as missing data (termed here as ‘Guessing Suppressed’ (GS) analysis). The purpose of these analyses was to provide evidence of the changes in parameters when guessing is accounted for, as a basis for development of the GIP.

5.5.2 Rasch Analyses Conducted on Simulated Data

5.5.2.1 Expected Outcomes from Rasch Analyses of the Simulated Data Sets

In line with the hypotheses outlined in Chapter 2, in relation to the comparisons of the parameters generated by the INIT and GS analyses, the suppression of the defined guesses embedded into the distribution of the simulated data were expected to have the following six results:

1. The item difficulties would vary, particularly for the more difficult items where the guessing had the greatest effect. The harder items would become even more difficult as a consequence of removing the guessed correct responses.
2. Because item locations are centred on zero (0), there would be a shift in the relative locations of the easier items, causing them to be relatively easier.
3. Consequently, the range of the distribution of item difficulties in the GS analysis would increase relative to the INIT analysis.
4. The interpretation of the changes in the distributions of student abilities would be more problematic because the defined guesses had been suppressed and treated as missing. Consequently, the student results would be impacted because they would have a reduced score based on the number of recoded items that had been indicated as defined guesses.
5. In the interaction between item difficulty and student ability, lower-ability students would have reduced ability estimates in the GS analysis than in the INIT analysis, and higher-ability students (who are less impacted by guessing in their responses) would have increased ability estimates.
6. The mean ability estimates, impacted by both the re-calibration of the item locations and the reduction caused by recoding correct responses to missing, would be lower for the GS analysis than for the INIT analysis.

The interrogation of these simulated data parameters aimed to provide the characteristics of student response patterns that were strong indicators in identifying guessing, and to subsequently develop a critical set of parameters that could be applied generally to other datasets to account for the misclassification caused by guessing.

5.6.2.2 Data Preparation for the INITIAL (INIT)

The INIT analysis was conducted on raw data with no accounting for guessing (although items identified by the code 9, correct guesses, were treated as correct responses = 1). There were no missing data. This analysis was the type of typical Rasch analysis carried out on any large-scale program of standardised educational assessment. From these analyses, using RUMM 2030 (Andrich et al., 2010), Rasch estimates were produced for the following parameters:

- item locations also known as “deltas” – Rasch difficulties (δ) expressed in logits;
- item category frequencies – facility expressed as a percentage;
- student ability estimates (β) – expressed in logits;
- the item difficulty (δ), which is the value at which the student cohort is demonstrated to have probability of a correct response of 0.5 ($p = 0.5$);
- item-student correlations (there must be less than 500 students for auto generation in RUMM 2030);
- the extraction of a raw score to ability conversion table; and
- the values of the item delta and student ability enable external calculation of the probability of a correct response to any item by any individual, using the Rasch model involving the item difficulty/student ability interaction.

5.6.2.3 Data Preparation for Guessing Suppressed (GS) Analysis

The second analysis took the original complete data sets and scored correct responses due to sufficient knowledge (1) and those at the threshold of knowledge (5) also scored correct (1). The randomly generated guess responses of “9” were recoded as “missing data”. All other responses were scored as incorrect, as in the INIT analyses. The incidence of missing data in these analyses resulted in the inability to calculate a Cronbach’s Alpha (α) for these data. In RUMM 2030 the Separation Index is a surrogate for the Cronbach’s Alpha.

5.5.3 Simulation Item/Student Maps

Figures 5.7 and 5.8 show comparative item/student maps of the simulation data sets generated in the Rasch analysis program Conquest (Wu et al., 1997). This program was used to highlight the extent of the changes in outcomes by providing a visual representation of the results. The figures have been arranged to provide comparison between the outcomes of the INIT analysis (which has not taken account of the defined guessing) and the GS analysis (for which the guessed responses have been recoded as “missing” data).

5.5.3.1 SIM 1 Analysis Results INIT and GS Analyses

Figure 5.7 compares outcomes of these analyses of the SIM1 data. The INIT analysis is shown in the left-hand column, with the effective scale ranging from -3 to +4 for the student ability estimates and the items, expressed in the same scale, ranging from -3 to +2. The outcomes of the GS analysis are shown in the right-hand column. Overall, the relative order of the locations of the items are maintained. Minor variation in the location order of the items in the GS analysis was a function of the randomisation process for the guesses in the “incorrect” zones of the student response patterns. This indicates that the randomisation algorithm was consistent in the allocation of uniformly increasing guesses as the difficulty of items increased over the range.

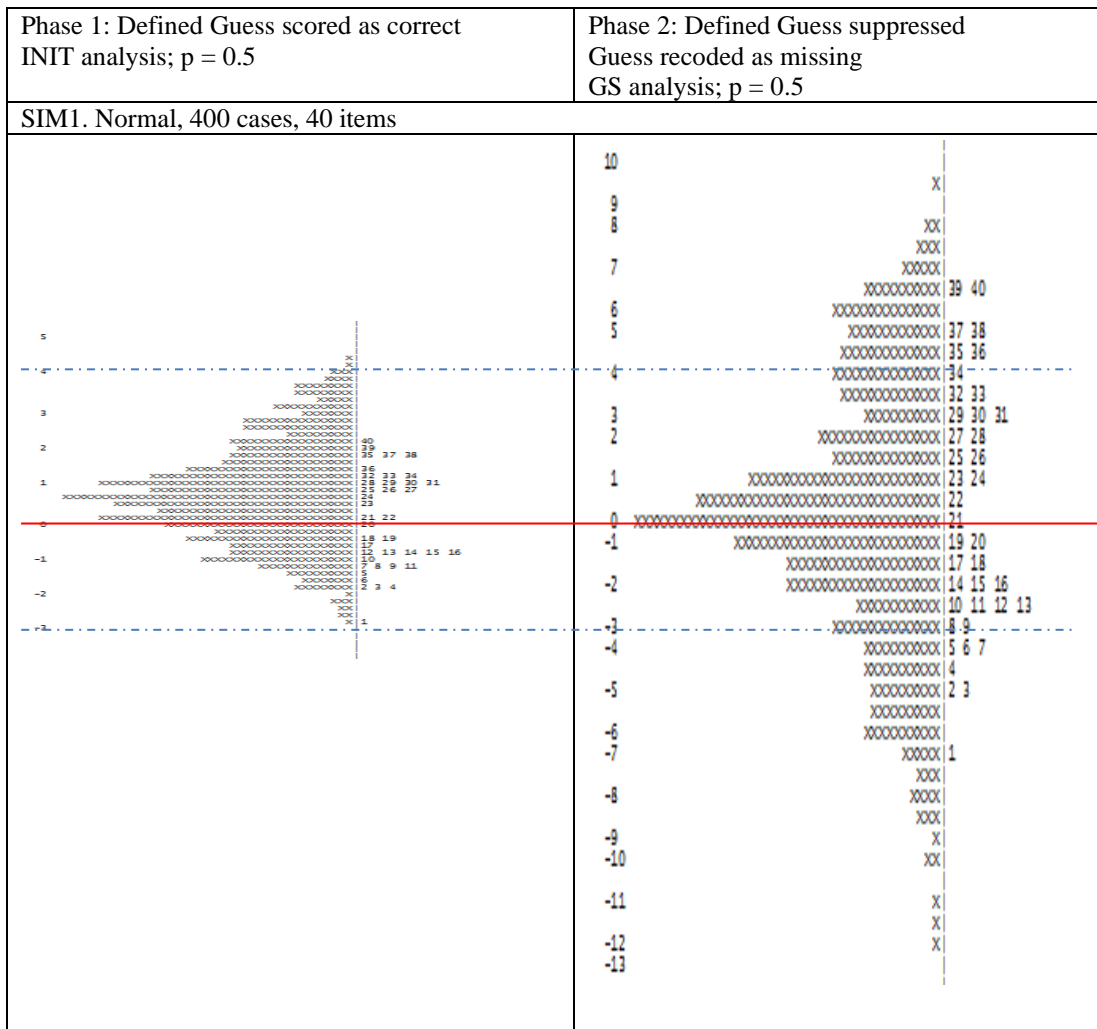
A major impact of the suppression of the defined guesses, as shown in the GS analysis, was the effect on the student ability estimates. The analysis discriminates between students more effectively in the GS analysis and reflects the “normal distribution” centred about zero of the student ability estimates. By comparison, the mean ability of the INIT analysis students is located at about 0.8 logits – an indication of the impact of not accounting for guessing for these data and the consequent inflation of the raw scores of all students, particularly the lower-ability students.

The other significant impact when guessing was suppressed was that the range of ability estimates of the GS analysis showed the the quartile of the higher-ability students having ability estimates above +4 logits, which was the maximum observed value in the INIT analysis. This observation highlights the detrimental impact of ignoring the presence of guessing on the estimates of higher-ability students. In the lower range of the ability estimates, the lower-ability students had estimates below -4 logits when guessing had been suppressed. By comparison, in the INIT analysis these students reported ability estimates in the range 0 to -3 logits, which suppressed the “true” estimates of the lower-ability students.

As expected, there was a significant increase in the distribution of the item difficulties centred about zero in each analysis. Whereas the locations in the INIT analysis are effectively within the range $-2 < \delta < 2$ (one outlier), in the case of the GS analysis the effective range of item locations is about $-5 < \delta < 6.5$, which provided the basis for the model to better estimate student ability (as inferred by Table 3.2). It is noted that the student ability estimates represented in the GS outcomes resulted in a reduction in the raw score for each defined guess. The results displayed in Figure 5.7 reflect the expectations of the hypothesis in that accounting for guessing will increase the distribution of item locations and consequently the increase the distribution of student ability estimates.

Figure 5.7

SIM1 Item/Student Map for INIT Analysis and GS Analysis



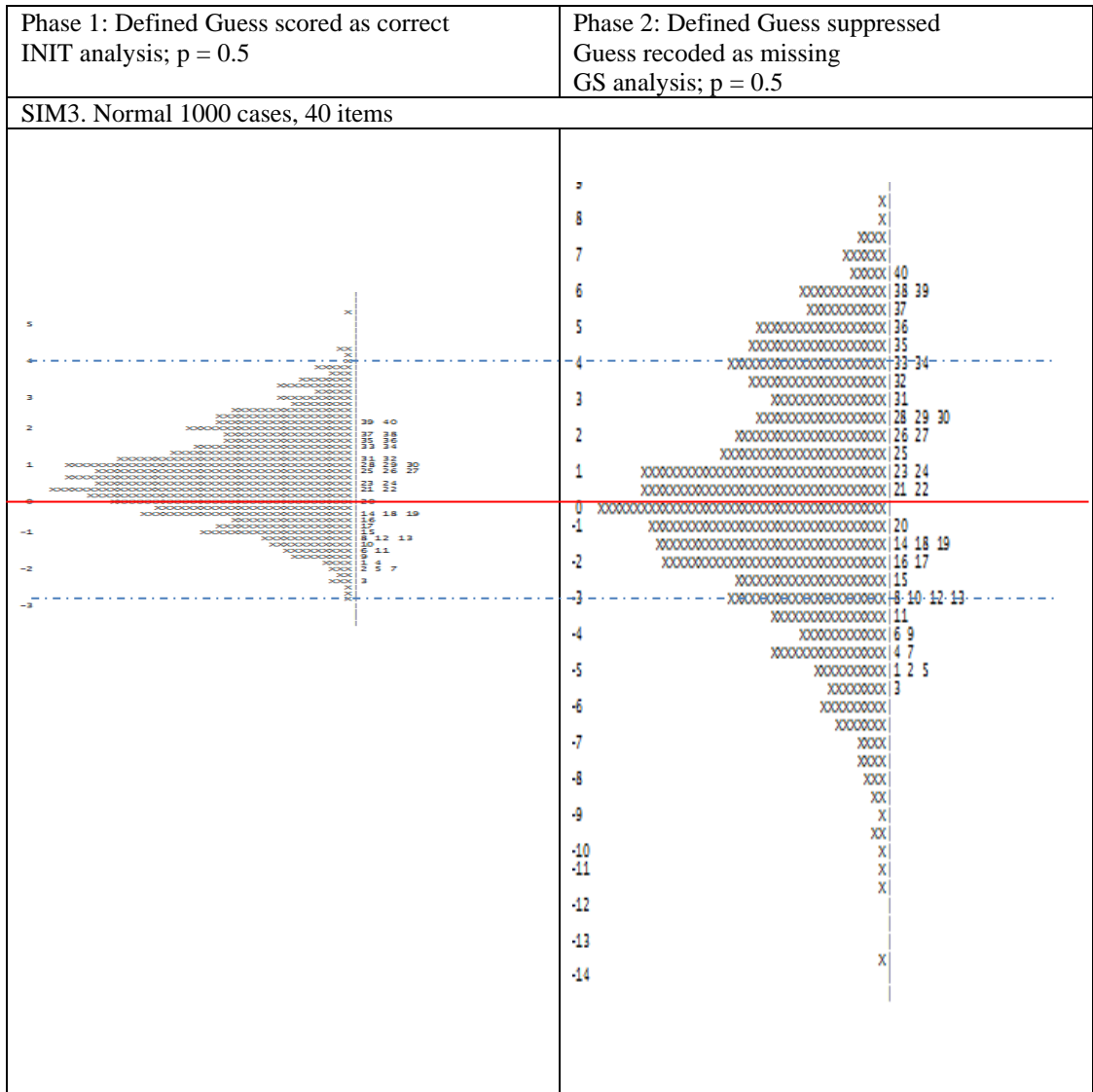
Note. The outcomes displayed in Figures 5.7 and 5.8 of the SIM1, SIM3 INIT, and GS analyses indicate the degree to which the reported performance of students differs when either ignoring or accounting for guessing. The figures have been aligned to be centred on a common value of zero (the defined mean of the item locations for each analysis) and the scales aligned (dotted horizontal lines). The purpose of this alignment is to highlight the impact on the range and distribution of item locations and student ability estimates from each analysis.

5.5.3.2 SIM 3 Analysis Results INIT and GS Analyses

Simulation 3 was designed as a sample of 1000 cases, with a target of a normal distribution about a mean “true” raw score of 20. The item/student maps shown in Figure 5.8 show a similar distribution to Figure 5.7, with the INIT distribution of the ability estimates of students considerably compressed compared to the range observed in the GS scales. This was a consequence of the compressed item difficulties in the INIT analysis that were impacted by guessing, with facilities inflated, and, as a consequence, item difficulty locations reduced. The impact was the contraction of the item locations, with the consequent contraction of the range of ability estimates of the students.

Figure 5.8

SIM3 Item/Student Map for INIT Analysis and GS Analysis



Note: The comparable figures for SIM2, SIM4, and SIM5 are provided in Appendix A. They demonstrate similar outcomes and have been added as appendices to reduce redundancy in this chapter.

Tables 5.6 and 5.7 provide a summary of the results from the five INIT and GS analyses for each of the simulated distributions. The results in each of these tables were consistent in highlighting the GS analysis, showing a greater distribution of item locations, and reflecting the “perfect” identification of the guessed items.

Table 5.6*Summary Results of Simulated Data Sets – INIT ANALYSIS – Defined Guesses (1)*

SET	Distribution	Sample N	Items (logits)			Students (logits)			
	Target		St. Dev	Range	Skew	Mean β	St. Dev	Range	α^*
SIM1	Normal	400	1.19	-2.6/1.9	0.65	0.72	1.39	-2.6/4.6	0.90
SIM2	Normal	250	1.37	-2.4/1.7	0.89	1.40	1.27	-1.4/4.0	0.79
SIM3	Normal	1000	1.26	-2.2/1.8	0.50	0.72	1.30	-2.2/3.9	0.89
SIM4	Too easy**	400	1.26	-2.3/2.2	0.34	1.27	1.48	-3.0/4.7	0.89
SIM5	Too hard**	400	1.19	-2.7/1.9	0.65	0.72	1.39	-2.6/4.6	0.89

Note 1. α^* = Cronbach computed for complete data sets.

Note 2. **Too easy/hard relates to the targeting of the items with respect to the ability of the cohort of students.

None of the INIT analyses summarised here takes account of guessing. The skew of the distribution reported in this study was an indicator of “shift” of the distribution of results as a result of the implementation of the data conditioning to account for guessing. Skewness relates to the asymmetry of the distribution relative to a normal bell curve. By definition, a normally distributed bell curve has a skew of zero. The mean in a positively skewed distribution has the mean to the right of the mode and median and have a longer “tail” to the right of the distribution; the change in the skew is an indicator of the impact on the distribution of the scores and the movement in the mean of the distribution.

Table 5.7*Summary Results of Simulated Data Sets – GS ANALYSIS – Guesses Scored Missing (9)*

SET	Distribution	Sample N	Items (logits)			Students (logits)			
	Target		St. Dev	Range	Skew	Mean β	St. Dev	Range	α^*
SIM1	Normal	400	2.88	-4.6/5.9	-0.18	0.03	3.20	-6.3/7.4	0.96*
SIM2	Normal	250	1.79	-3.1/2.4	0.64	1.27	1.69	-3.1/4.4	0.81*
SIM3	Normal	1000	3.08	-4.2/6.3	-0.82	0.02	3.12	-6.4/6.8	0.96*
SIM4	Too Easy	400	2.52	-3.5/5.8	-0.60	1.15	2.83	-5.7/7.2	0.95*
SIM5	Too Hard	400	3.27	-4.3/8.4	-0.21	-0.16	3.29	-6.1/10.2	0.95*

Note. α^* = RUMM Separation Index indicated for incomplete data sets (missing data).

The following six features of results were common across the analyses comparing the INIT and GS analyses of the simulation data:

1. The range of the item difficulties was increased when guessing was suppressed (GS).
2. The distribution of student ability estimates increased to reflect the wider distribution of item difficulties from which ability estimates were constructed, i.e., the more difficult items led to higher ability estimates and lower item difficulties reflected lesser ability in the trait (GS).
3. The item order with respect to difficulty was maintained.
4. The mean ability of the GS analysis was less than in the INIT analysis, due to the uniform reduction in the raw score for removed defined guesses.
5. The higher-ability students had increased ability estimates in the GS analysis.
6. The lower-ability students had reduced ability estimates in the GS analysis outcomes.

These outcomes are in accord with the hypothesis; that is, to ignore guessing is to deflate the range of item difficulties and student ability, creating a non-uniform increase in raw scores across student ability levels. The following important points were also noted:

7. The degree of skew in the data moved in a negative direction to account for the over-estimation of student ability in the data in which there were no adjustments for defined guessing.
8. The reliability indicator – Cronbach’s α in the case of the complete data sets and Separation Index in the case of the incomplete data – improved when guessing was suppressed.

These eight points, when taken in combination, indicate that accounting for guessing improved the reliability of the data and hence the quality of the variable and the subsequent reporting of item locations and student abilities. The process of recoding the defined guesses uncovered a degree of misclassification in the INIT data, which was to the order of 0.5 of a logit when estimating the mean ability of the students. This difference in ability relates to approximately one year of learning in a large-scale assessment such as NAPLAN (Choppin, 1983). For the domain of Reading, the difference between the “minimum standard” between Year 5 and Year 7 was 52 scaled score points. A value of 70 scaled score points represented one standard deviation in the initial calibration of the NAPLAN scale – or approximately 0.75 logits over two years (BEMU, 2010).

5.6 Initial Observations and Iteration of a Guessing Indication Protocol (GIP)

5.6.1 Initial Determination of the GIP Parameters

Critical to the analyses in Study 1 was that the guessed responses of the students were known. In each of the simulations the “guessed” items were defined and could be suppressed in the analyses. In cases where guessed items are not known, Andrich et al. (2012, 2015) proposed that the misfit of the item-student interaction of each student with each item can be used as an indicator of guessing. Andrich et al.’s (2012, 2015) research proposed that item/student residuals of 2.56 or greater suggested a less than 5% probability of a student of given ability being able to correctly respond to an item with difficulty location significantly higher than the student’s ability estimate. They reported similar directional outcomes to the current study’s findings reported earlier in this chapter, with improved results using the item-student correlation parameter as a criterion to determine which item/student interactions should be suppressed.

The intention of the current research is to improve upon the thresholds proposed by Andrich et al. (2012, 2015), as it is proposed that these values may be too parsimonious. To achieve this, a Rasch analysis of the responses of each of the simulations interactively produced two major statistics regarding each test:

1. an ability estimate for each student based on the total score achieved for the set of items in the test; and
2. a location of the difficulty of each item in the test, each expressed in the unit of logits.

The relationship between these two major statistics in relation to determining a student ability from the score achieved on the items of calculated difficulty was articulated in Table 4.2. The determination of a set of parameters that indicated likely guessing at the individual student/item level was grounded in calculating and observing two additional statistics:

3. the probability that a student of a given ability could achieve a correct response to an item of given difficulty; and
4. a measure of the degree to which the probability of a correct response by a student to an item of a determined difficulty, that exceeds the student ability estimate, differs from the expected result for that interaction. This measure was termed the student-item correlation index.

In interrogating the five simulated data sets, these two statistics were viewed as potentially complimentary in identifying correct responses due to guessing at the individual item-student interaction level.

The item/student interaction probability was derived from the item and student parameters using the RM. The matrix of the item-student probability was calculated from the initial estimates of difficulty and student ability, and it was derived in the INIT analysis by applying the RM to the relative difficulty of the individual item (δ_i) and the observed ability of the student on the overall test (β_j). The derivation of these values is defined in Eqns 5.3 and 5.4.

$$\Pr(1)_{ij} = \exp(\beta_j - \delta_i) / (1 + \exp(\beta_j - \delta_i)) \quad \text{Eqn 5.3}$$

where $\Pr(1)_{ij}$ is the probability of student (j) responding correctly to item (i) given
 β_j is the determined ability of the student, and
 δ_i is the determined difficulty of the item expressed in logits.

The item-student residual index as derived by Andrich et al. (1988) is calculated as:

$$\text{IF}(\text{Score}_{ij}=1) \text{ then residual} = (\text{SQRT}((1 - \Pr(1)_{ji}) / \Pr(1)_{ji})) \text{ else } (-\text{SQRT}(\Pr(1)_{ji} / (1 - \Pr(1)_{ji}))) \quad \text{Eqn 5.4}$$

where Score_{ij} is the observed scored response of student(j) on item(i); and
 $\Pr(1)_{ij}$ is the probability of student(j) responding correctly to item(i) (see Eqn 5.3).

The value of the residuals is the measure of the degree of the difference between the expected result (0,1) for an individual item/student interaction, given the item difficulty, student ability, and the observed result. Greater values of this index reflect greater misfit between the individual result and the expected result predicted by the model, and they were used as an indication of a guessed response. The idea proposed and advanced in this Study 1 is that these statistics can be used to generate data conditioning to account for indicated guesses, rather than simply reporting the misfit as demonstrated by Andrich et al. (2012, 2015).

5.6.1 Pfa Step 3: Initial Development of the Guessing Indication Protocol (GIP)

In attempting to determine a set of parameters that would be defensible in identifying an individual guess in a student response pattern, several combinations of values of the two indicative indices (using Eqns 5.3 and 5.4) were trialed to determine which values were accurate and consistent in “identifying” defined guessed responses in the simulated data. The implicit relationship between the calculation of the residual and the probability of a correct response, given the ability of the student (β_j) and the difficulty of the item, makes these calculations complementary.

5.6.2 Identification of Guesses Using Critical Indices

5.6.2.1 Interrogation of SIM1

By observing the values of the probabilities and residuals for item/student interactions that were both consistent with the defined guessed response and defensible as a logical result, the GIP was determined. The results of this process and series of analyses are the focus of this sub-section. Table 5.7 is an extract from Appendix A, which shows the observed response pattern (1. Scores), the derived item-student probability of a correct response (2. $Pr(1)$), and the item/student correlation statistic (3. Residual) for SIM1 for each student for SIM1. Table 5.7 is provided as an example of the multiple composite parameters generated to reflect the outcomes of the five final data sets used to determine a defensible and appropriate set of parameters that are consistent in identifying the defined guesses embedded in the data.

This abridged extract shows a range of the lower scores with the defined random guesses (coded as 9) highlighted red for convenience. Specifically, it presents a selection of 16 of the 40 items from SIM1, for students ranging from lower ability through higher ability (separated by the red lines). The table is sorted horizontally, left to right, from the easiest item to hardest the item. These item difficulties were derived from the INIT analysis. The table is sorted vertically by student ability estimate to display the response pattern of extracted parameters for each item/student interaction. The calculated ability estimates and the raw score (RawSc) from the INIT analysis are also included in the table for each student. The blue highlighted cells display the calculated values of the probability of a correct response ($Pr(1)_{ij}$) and the item-student residual calculated using Eqn 5.4. The table is partitioned (the red line) to show the interaction of the GIP procedure with different ability groups of students of increasing ability as indicated by their relative raw score. Identification of the most efficient GIP parameters was an iterative process involving reconciling both statistical values and face validity considerations in resolving the parameters that indicated a defined guess.

In Table 5.7 the lowest ability student in the simulated data set is student S390. The full table provided in Appendix A shows that this student answered no items “correctly”, yet successfully guessed four items to generate a raw score of 4. The fact that S390 could answer no items correctly does not necessarily indicate that the student has no ability in the trait. Rather, it is a reflection that the scope of items in this test did not include items of an appropriate level of difficulty that could be successfully accessed by this student. The following three general observations can also be made for S390 from Table 5.7:

1. The four correct defined guesses were items 19, 29, 31, and 34.
2. For each of the correct defined guesses, the probability of a correct response was less than 0.25.
3. The item-student residual was positive and had a value greater than 3.0 [Andrich (2012, 2019) proposed greater than 2.56 as the critical value].

Consider another student, S371, who had a raw score of 15, which derived an ability estimate of -0.647. This student had 14 defined correct guesses (items 4, 10, 17, 20, 22, 23, 25, 28, 31, 33, 34, 35, 36, and 40). Five items were considered very unlikely outcomes (calculated residuals greater than 2.56). Items 4 and 20 show the probability of a correct response to exceed 35%, which results in residual statistics less than 1.4. The impact of these correct guesses is to contaminate the statistics of other items in the student's response pattern by influencing the calibration of the item locations. These results show that student S371 only correctly answered item mkQ01, and the remainder of the observed scores were a result of 14 of 39 correct random guesses. This inflated assignment of ability for S371 (relative to S390, despite little difference in their true ability) directly relates to both the calculation of the individual probabilities of correct responses on all other items and the calculation of the item-student correlation statistic.

These observations highlight three challenges in identifying guesses in response patterns:

1. As student raw scores increase there is a decrease in the number of items that the student is likely to guess, or to be indicated as a probable guess.
2. Lower ability students have a higher probability of guessing the correct answer, with an observed randomising range of a correct guess about +/-6% in the SIM1 data. This estimate was derived by the variation observed in the number of guesses generated by the randomising algorithm. This can be seen by reference to the extract of students with a raw score of 16, ability of -0.511, in which S314 has had six successful guesses, whilst student S334 has had 11 successful guesses.
3. The greater the "capacity" of the student to guess correctly, the lower is the capacity to identify guessing in the response pattern using probability and residual indices.

Determining a set of parameters to apply to the analysed data to account for guessing was thus an iterative exercise that involved observing the minimum value for the item-student residual in conjunction with a probability of a correct response. It was initially considered that the lowest defensible criterion at face value, in relation to the probability of a correct response, was a probability less than 0.25. This value of 25% is consistent with the 1-in-4 construction typical of a multiple-choice structure. This constraint resulted in a reduced identified guess recovery rate compared to what would be identified if the 0.30 (which implies a ratio of approximately 1-in-3) criteria used in Andrich et al., (2015).

Having observed the interaction between the defined guesses of the simulation data (see Table 5.8 for SIM1) and the item-person residuals, a value of 1.75 in combination, with a probability of correct response of less than 0.25, was determined to be consistent in identifying the defined guesses. The residual value of 1.75 is less than Andrich et al.'s (2015) value of 2.56, but it represents a statistical confidence level of approximately 90%. In the analyses of each of the simulations the application of these thresholds did not produce any instances of Type 1 errors (i.e. indicating an interaction to be a guess that was not pre-defined in the data).

As a result of these observations and the iterations comparing the critical parameters and the defined guesses in the simulations, the parameters from the INIT analysis assigned for initial investigation to indicate a guess were defined as:

IF observed_response(ji)=1, AND
 Pr(1)_response(ji)<0.25, AND
 item-student correlation(ji)>1.75),
THEN 'classify observed_response(ji) as a guess (7), recoded as missing data
ELSE maintain observed_response(ji) as scored (1).

Syntax 5.1

5.7 Preliminary Analyses and Observations of the Initial Defined Protocol

5.7.1 Initial Analysis Results

As indicated in the Plan for Analysis (Section 5.3.2), an INIT analysis of each dataset was conducted to extract the primary statistics. From these INIT analyses the item/score interactions, item/student ability probabilities, and student/item residuals were calculated. Syntax 5.1 was then applied to indicate any individual student/item interaction that failed the GIP parameters: in the item/score interaction, the item had been scored as correct; in the item/student ability calculation, the probability of a correct response was less than 0.25; and the student/item residual index had a value greater than 1.75 (see Eqns 5.3 and 5.4). If an individual student/item interaction failed these GIP parameters, the individual item/student response was recoded from a correct response (1) to “missing data” (7) in the re-analysis of the data to “account for guessing” using the proposed GIP procedure.

As shown in Table 5.9, a comparison of the actual defined random guesses with the GIP-recovered random guesses indicates a relatively low percentage of defined items identified by the GIP process.

Table 5.9

Summary of the GIP Recovery Rates by Simulation $p = 0.5$

Simulation	Count of Actual Defined Random Guesses	Count of Guesses recovered by GIP strategy	Recovery rate Identified/actual
SIM1, Normal, 400,40	1771	572	32.3%
SIM2, Normal, 250,20	360	55	15.3%
SIM3, Normal, 1000,40	4304	1397	32.5%
SIM4, Easy, 400,40	1311	337	25.7%
SIM5, Hard, 400,40	2164	754	34.8%

Note. “Count of actual random guesses” is the number of defined guesses from the simulation data sets. These interactions were initially coded as correct in the INIT analysis and recoded as missing in the GS analysis. The actual defined random guesses that were identified using the GIP procedure are recorded in the “count of guesses recovered by the GIP strategy”.

Although this result was initially considered a low rate, further disaggregation of the cohorts revealed that this “recovery rate” was not consistent across all student ability ranges. Follow-up analyses sought to evaluate these recovery rates by student ability, for each data set. Table 5.10 shows the variation in the capacity of the GIP procedure to identify the defined guessed responses across various ability groups within the cohort for each simulation. The proposed GIP parameters with a default p value of 0.5 produced relatively low rates in the capacity of the GIP to recover instances of defined guessing. Across the simulations an overall recovery rate of about 30% was achieved when comparing the GIP identified “guesses” with the actual defined guesses (Table 5.10). It was noted that the procedure was more efficient in indicating the defined guesses in the lower-ability groups, which is the student group in which guessing is more likely to occur.

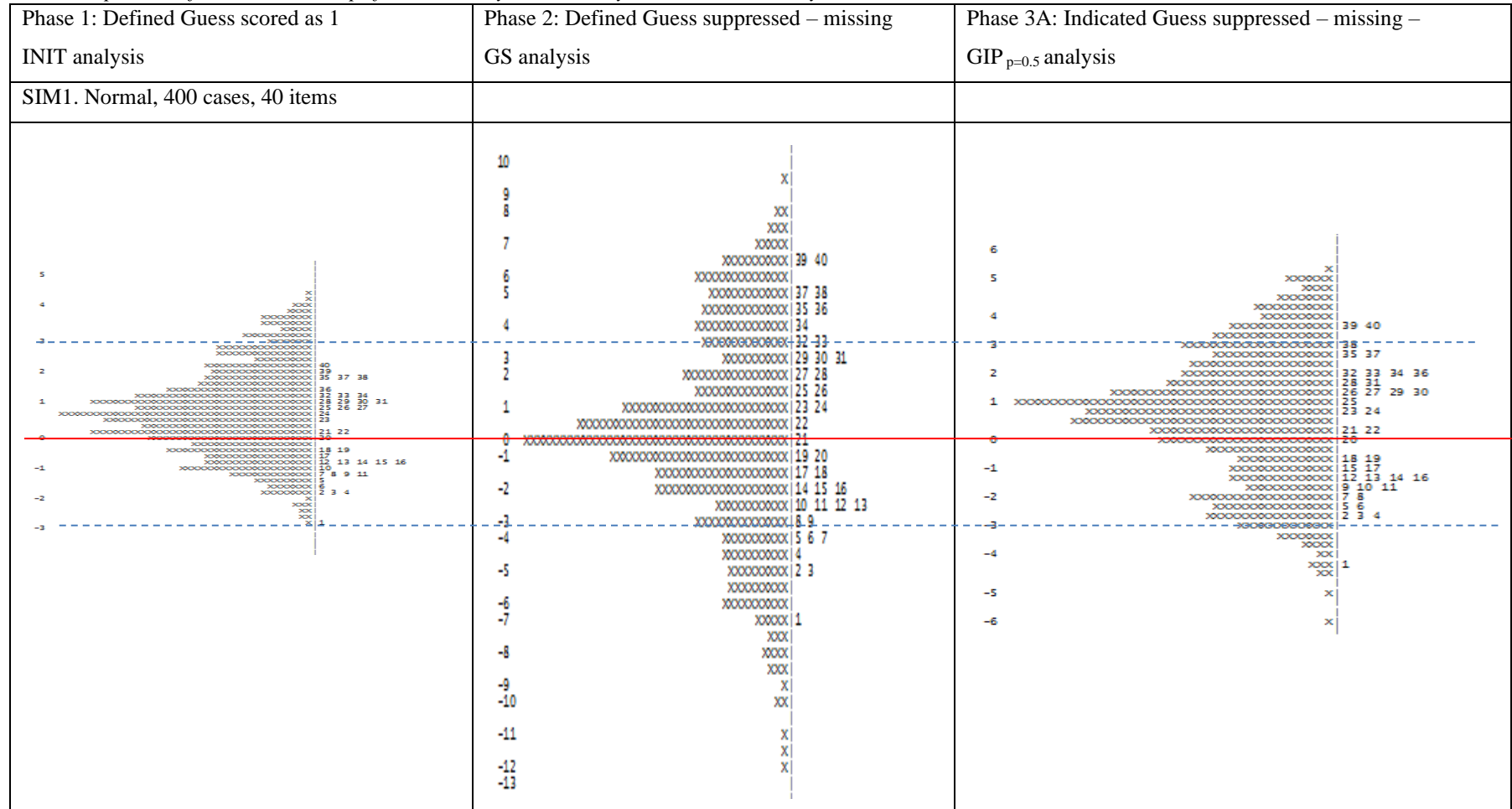
The item/student maps of the analyses for the data sets of SIM1 and SIM3, respectively, which are representative of the outcomes of all simulations for the INIT, GS, and GIP3A phases of analysis are provided in Figures 5.9 and 5.12. The probability of a correct response was set at the default level of $p = 0.5$. For the GS analysis each of the defined guesses in the dataset were recoded as “missing” data. For the GIP3A analysis, the GIP identified guesses were recoded as “missing” data. Hence the student ability estimates in Figures 5.9 and 5.12 reflect a raw score for each student that has been reduced by the number of guesses defined (GS) or identified by the GIP process. Figure 5.9 appends the GIP3A analysis to the outcomes shown in Figure 5.7 and thereby shows the degree to which the results are misclassified by ignoring guessing in the GIP3A analysis compared to the values of the defined guesses reported in Figure 5.7.

Table 5.10*Elaboration of the GIP Recovery Rates by Simulation $p=0.5$*

Simulation	Group (Quartile)	Proportion of items guessed (%)	Count of Actual Random Guesses	Count of Guesses recovered by GIP strategy	Recovery rate Identified/ actual
SIM1, Normal, 400,40	Q4, most able	1.8	72	0	0.0%
	Q3, able	11.2	446	9	2.0%
	Q2, less able	12.0	478	116	24.3%
	Q1, least able	19.4	775	447	57.7%
	Overall			1771	572
SIM2, Normal, 250,20	Q4, most able	1.1	14	0	0.0%
	Q3, able	6.0	74	0	0.0%
	Q2, less able	9.2	114	1	0.9%
	Q1, least able	12.7	158	54	34.2%
	Overall			360	55
SIM3, Normal, 1000,40	Top decile	0.4	14	0	0.0%
	Decile 9	2.5	98	0	0.0%
	Decile 8	6.5	259	0	0.0%
	Decile 7	10.0	400	0	0.0%
	Decile 6	12.0	481	12	2.5%
	Decile 5	11.0	438	60	13.7%
	Decile 4	11.8	473	137	29.0%
	Decile 3	14.8	591	266	45.0%
	Decile 2	19.8	791	452	57.1%
	Decile 1	19.0	759	470	61.9%
Overall		10.8	4304	1397	32.5%
SIM4, Easy, 400,40	Q4, most able	1.8	72	0	0.0%
	Q3, able	5.4	217	0	0.0%
	Q2, less able	9.7	389	62	15.9%
	Q1, least able	15.8	633	275	43.4%
	Overall			1311	337
SIM5, Hard, 400,40	Q4, most able	7.6	303	0	0.0%
	Q3, able	10.4	417	10	2.4%
	Q2, less able	18.1	722	293	40.6%
	Q1, least able	18.1	722	451	62.5%
	Overall			2164	754

Figure 5.9

SIM1 Comparison of Item Student Maps for INIT Analysis, GS Analysis, and GIP3A Analysis



Note. The left-most plot shows the distribution of items and students on a common scale derived from the INIT analysis. This analysis took no account of the defined guesses in the data set of SIM1. The middle plot shows the distribution of items and students derived from the GS analysis (guesses suppressed). The right-most plot shows the item/student map of the analysis when the GIP parameters had been applied to the data. All items were centred on a location of zero (by definition).

5.7.2 Observations Regarding the Initial Analysis Results

In the INIT analyses, the range of item difficulties was between -2.66 and + 1.87, a range of 4.53 logits. The student ability estimates varied between -2.62 and + 4.64, a range of 7.26 logits. By comparison, when all the defined guess items had been suppressed and recoded as missing data (GS analysis), item locations varied between -4.64 and + 5.89, a range of 10.5 logits. Yet the distribution of student estimates is more “revealing” in relation to the impact on the distribution when guessing is ignored. The GS results show that the mean of the ability estimates was close to zero (0.03). This was in accord with the intended distribution of the data set in its simulated construction. The student estimates varied from -6.31 to +7.45, a range of 13.76 logits. It is noticeable that the distribution of the lower-ability students in this figure reveals a greater discrimination among the students of SIM 1, compared to the INIT analysis figure. The GS analysis reveals a negative skew not reflected in the INIT analysis.

In the GIP3A analysis, correct responses were suppressed as “missing data” by recoding the item/student interactions whenever the item failed to achieve a location/ability interaction defined by the GIP. As noted in Table 5.9, the application of the GIP protocol was sub-optimal in identifying all defined guesses in student response patterns. Overall, the GIP3A phase of the protocol resulted in a rate of approximately 30% in recovering the actual defined guesses. However, in relation to the lower-ability students, the recovery rate of the defined guesses was above 50%. This is reflected in the wider distribution of ability estimates for the lower-ability students. The GIP3A analysis increased spread of item locations and student ability estimates compared to the INIT analysis (see Figure 5.9).

Overall, the application of the GIP procedure to the SIM1 data resulted in a minor reduction in the mean of the ability estimates of the cohort compared to the INIT analysis (0.61), but a greater distribution of ability estimates, with the most able students recording an ability estimate of +4.66 and the least able students recording an ability estimate of -5.38. These analyses show that the impact of applying these GIP parameters to the INIT data was to discriminate more effectively among the lower-ability students. It can also be seen that the impact on the results of, and discrimination between, the average ability estimates of higher-ability students was not as significant as in the lower-ability students.

Figure 5.10 shows the comparison of the item locations for the three datasets. The variations in the item locations for each Phase shown in Figure 5.11 are reflected in Figure 5.10. Specifically, the GS analysis provides the greatest range of item locations with both significantly lower locations for the easier items and higher locations for the more difficult items than the INIT estimates. The INIT analysis has the least discrimination in the item locations and the GIP3A analysis tends to be mid-range, but still noticeably different from the INIT locations at the extremes.

Figure 5.10

SIM1 Comparison of Item Locations INIT Analysis, GS Analysis and GIP3A $p=0.5$ Analysis

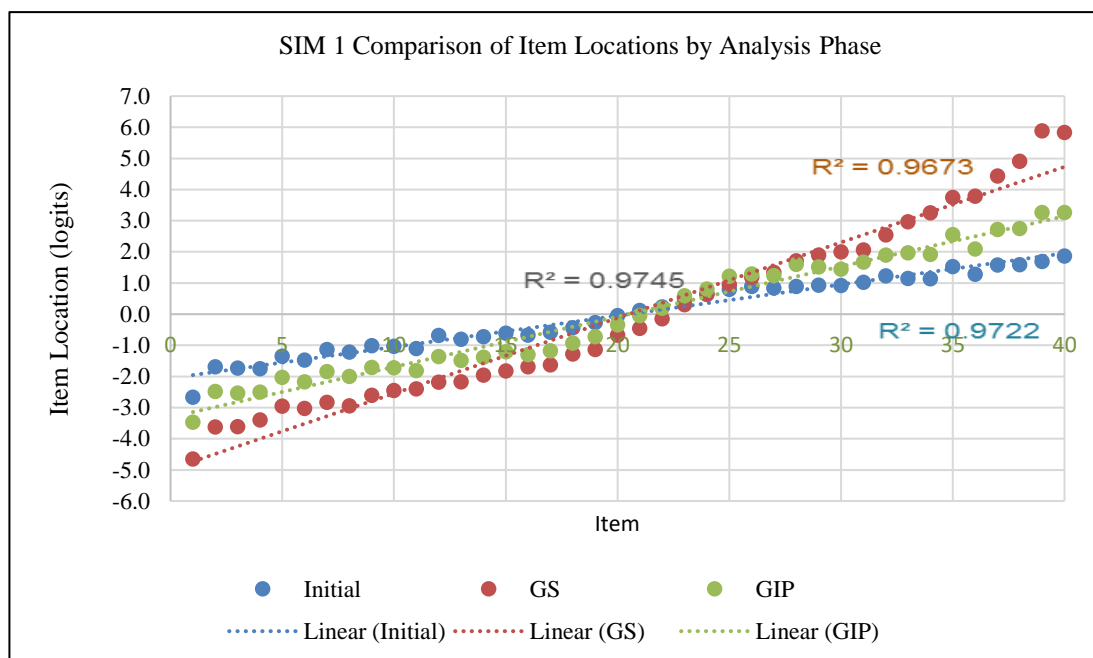
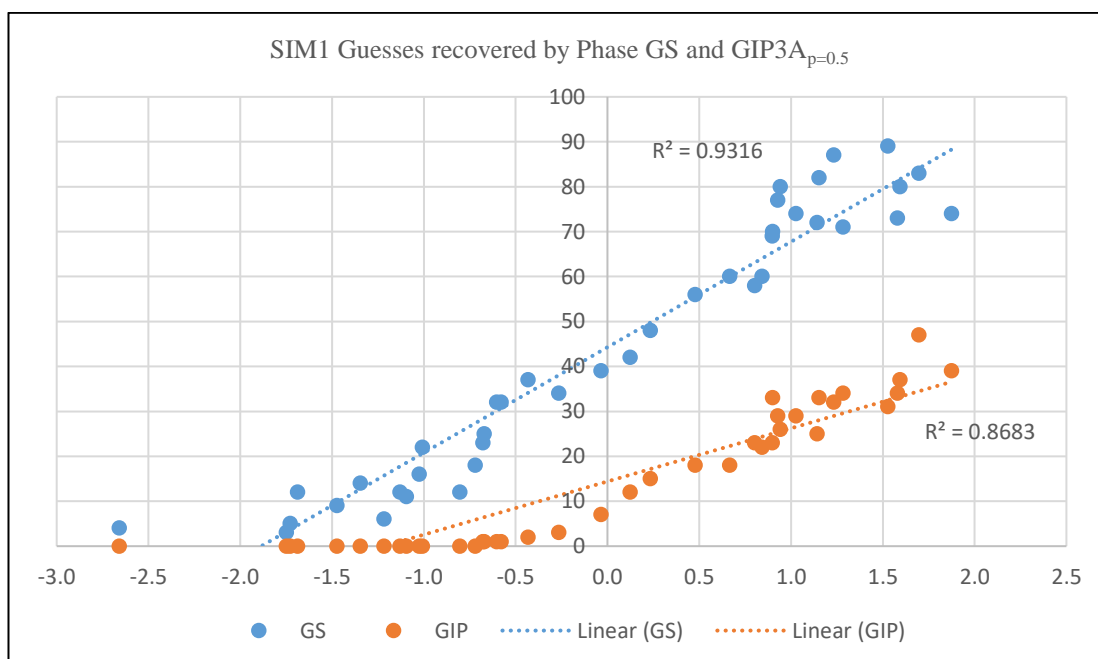


Figure 5.10 shows the comparison of the recovery rate of the implementation of the GIP procedure compared to the actual defined guesses in the 40-item data set of SIM1. Results of the SIM1 Phase 2 and Phase 3A analyses show that for the easiest items the protocol had a relatively low efficiency in identifying the guessed items (see Figure 5.10), which impacted the R^2 statistic. This is because the ability required to correctly answer easy items was relatively low. As items became more difficult, the protocol was more effective.

Figure 5.11

SIM1 Comparison of Count of Defined Guesses Recovered From GS and GIP3A $p=0.5$ Analysis



The SIM3 data were designed to be a normal distribution, which is shown in the graphic in the middle of Figure 5.13 for which the defined guesses have been suppressed. These results mirror the outcomes of SIM1, with the outcomes of the GIP3A procedure having a minor impact on the overall mean compared to the INIT results. The figure likewise shows marginal impact in the upper ability ranges but a significant impact on the distribution of abilities of the lower-ability groups. The range of the item distribution of the GIP3A analysis is also noted as greater than that observed in the INIT analysis.

Figure 5.12

SIM3 Comparison of Item Student Maps for INIT Analysis, GS Analysis, and GIP $p=0.5$ Analysis

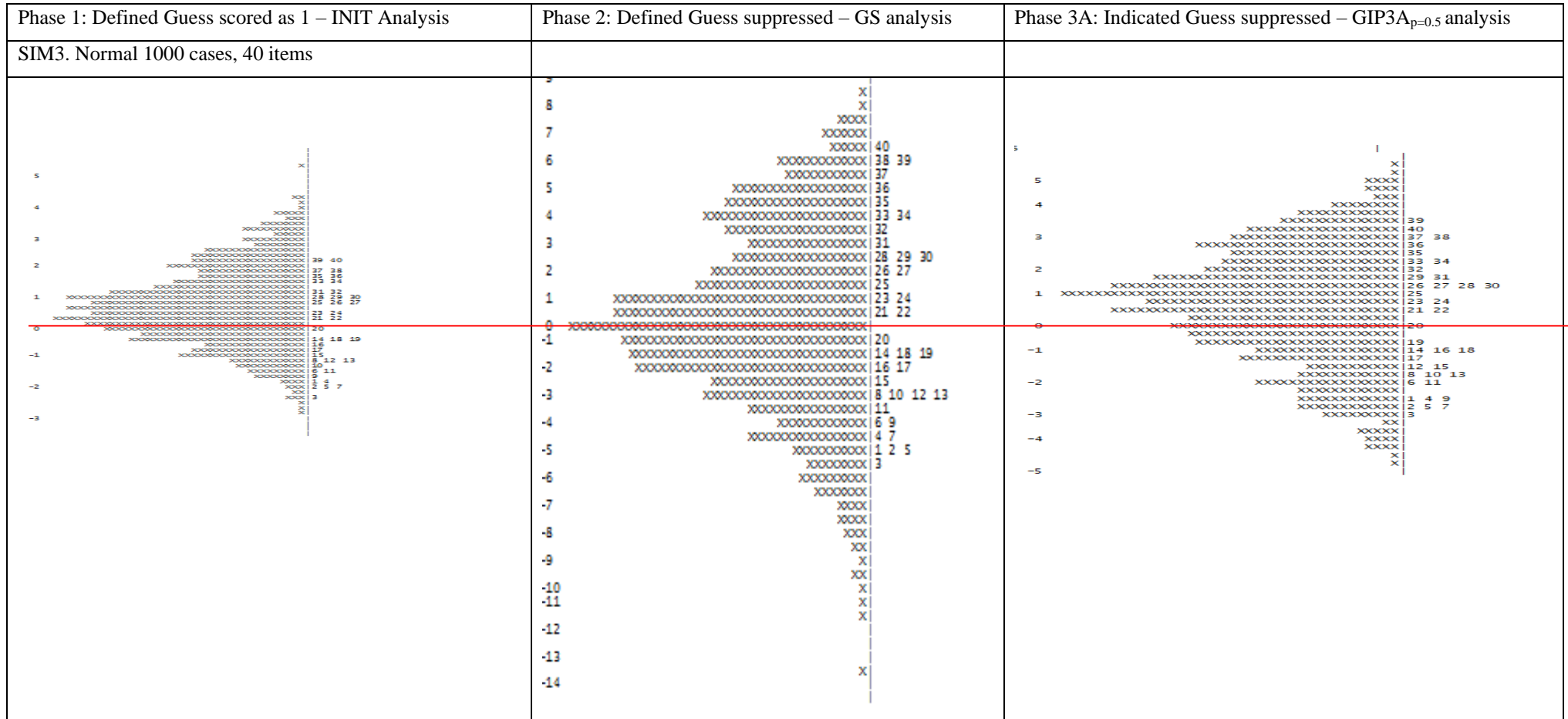


Figure 5.13 follows a similar pattern to that shown for SIM1 data in Figure 5.10. The R^2 value of 0.814 indicates that about 66% of the variation in the mean of the GIP analysis can be explained by the frequency of defined guesses. The GIP implementation was relatively inefficient for items with locations less than zero (the easy items), but it improved for more difficult items because of the suppression of the guessed items, with greater differentiation between the item locations in the more difficult items.

Figure 5.13

SIM3 Comparison of Item Locations – INIT Analysis, GS Analysis, and GIP3A_{p=0.5} Analysis

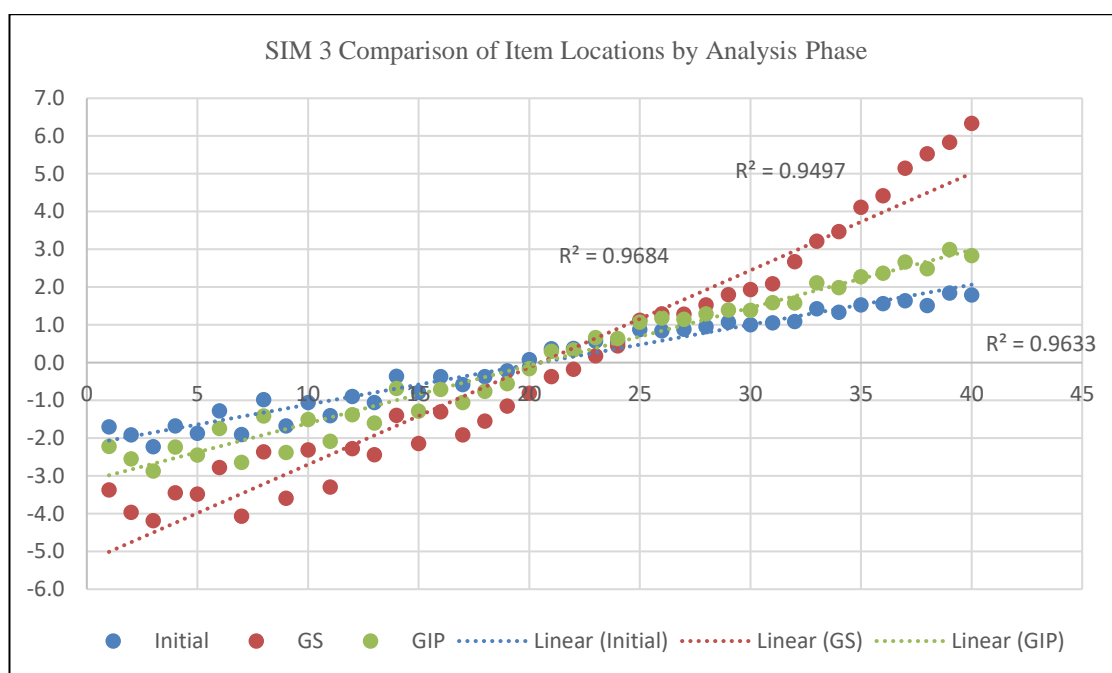
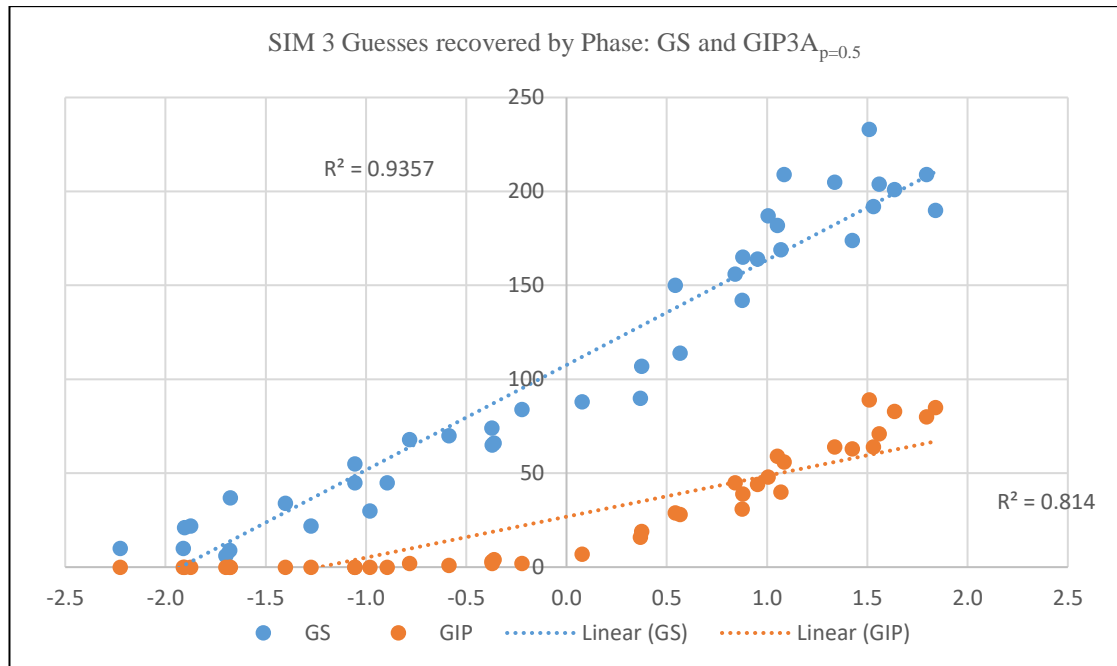


Figure 5.14 shows a similar pattern to that observed in Figure 5.11, with relatively few instances of guessing identified by the process for items with locations less than zero. This observation precipitated a review of the components of the protocol as described in the following section.

Figure 5.14

SIM3 Comparison of Count of Defined Guesses Recovered From the GS and GIP3A_{p=0.5} Analyses



5.8 A Second Iteration of the Protocol: Adjusting the p Value

5.8.1 Modification of the Protocol Parameters

In the initial iteration of the GIP, in which the default value of response probability (RP or p value) of $p = 0.5$ was applied, an insufficient capture of guesses was observed and so the default value of $p = 0.5$ was reconsidered. In evaluating the possible (theoretically and practically) sound revisions that could be made, it seemed reasonable that to effectively identify a correct, ability-based response to an item a better than 50% probability of responding with a correct answer would be more appropriate. Since all parameters are stochastic and, to an extent, an application of a subjective threshold, it was deemed reasonable and defensible to increase this threshold. The theme of this research is conceptual, and consequently this iteration of the protocol was investigated to observe the impact of the modification. Consideration of an “optimal” p value is a possible topic of future research.

To improve upon the identification rates achieved in the preliminary results for the GIP displayed above, an investigation of the effectiveness of the GIP procedure when a RP value of 0.6 was introduced and the efficacy in the analysis of data for each simulation was evaluated. It is noted that both RUMM 2030 (Andrich et al., 2013) and Conquest (Wu et al., 1998) have default values that compute the probability of a correct response as 0.5; that is, the dichotomous state of a response is right or wrong (0/1). This constraint defines that at the intersection between the ability of a student being equal to the difficulty of the item in the RM, the probability of a successful response is 0.5.

This can be resolved mathematically that by simplifying the Equation 5.5

$$0.6 = \exp(\beta - \delta) / (1 + \exp(\beta - \delta)) \quad \text{which derives}$$

$$-0.405 = \beta - \delta \quad \text{Eqn 5.5}$$

Hence by adjusting the p value to 0.6, the student ability (β) required to have at least a 0.6 probability of success on an item of difficulty δ is ($\delta + 0.405$). Effectively, this adjustment requires the ability of the student to be 0.405 logits greater than the difficulty of the item to indicate a correct response with at least a 60% probability. In relation to the calculation of the probability of success in the item/student interactions, the “adjustment” of the ability of the INIT student estimates by the constant -0.405 allows the recalculation of the $\text{Pr}(1)$ and the item/student residual by applying the revised ability estimate to Equations 5.3 and 5.4 respectively.

5.8.2 Result of the Modification of the p Value With the Protocol Parameters

5.8.2.1 The Response Probability $p=0.6$ Recovery Rates

The process reported in Section 5.6 was repeated after adjusting the values of the residuals and the Response Probability (RP) of a correct response ($\text{Pr}(1)$) with the revised p value. These revised analyses with $p = 0.6$ gave rise to a new set of GIP data, with the number of items failing the defined GIP parameters increased. Table 5.11 shows the impact of the implementation of a p value of 0.6 to the results of this iteration of the GIP implementation. It also shows a significant improvement in the recovery rates of the GIP procedure compared to Table 5.9.

Using a p value of 0.6 for the probability of a correct response for the student produces a greater efficiency in the recovery of the defined guessing in the simulated data (Table 5.11). Further analysis revealed there were no Type 1 errors generated by increasing the p value to 0.6; that is, there was no incidence of a defined correct response being recoded as missing under this protocol. Only defined guesses, although not all, were identified when this modification to the analysis and calculation of the GIP parameters was introduced.

As shown in Table 5.11, the effect of changing the probability of a correct response from 0.5 to 0.6 was to increase the capacity of the GIP procedure to identify the defined guesses without compromising the accuracy by introducing any Type 1 errors on average by approximately 15%. The comparison of the analyses shows that the use of the $p = 0.6$ constraint tended to maintain the discrimination among the lower-ability students and increase the range of both item difficulties and student abilities in the simulated responses.

Table 5.11*Comparison of Guess Recovery Rates $p = 0.5$ and $p = 0.6$*

Simulation	Count of Actual Defined Random Guesses	Count of Guesses recovered by $p = 0.5$ GIP strategy	Count of Guesses recovered by $p = 0.6$ GIP strategy	Increased number recovered	Recovery rate Identified/ actual $p = 0.5$	Recovery rate Identified/ actual $p = 0.6$	Rate of increase
SIM1	1771	572	791	219	32.3%	44.7%	12.4%
SIM2	360	55	119	64	15.3%	28.9%	13.6%
SIM3	4304	1397	2049	652	32.5%	47.6%	15.1%
SIM4	1311	337	589	252	25.7%	44.9%	19.2%
SIM5	2164	754	1031	277	34.8%	47.6%	12.8%

A feature of the disaggregation of the recovery rates by group is the region in which the increased recovery is most apparent (see Table 5.12). In each simulation the increase in the p value resulted in an increased proportion recovered in the lower-ability groups than in the higher-ability groups. In all cases the increased p value did not uncover any guesses in the highest ability group, although in all cases there were some defined guesses in this group.

The evidence presented earlier in this chapter, and the underlying hypothesis of the research, is that guessing is more likely to occur among the lower-ability students. In this respect, the increase in the p value increased the capacity of the GIP to identify probable guesses in this group, without error. This was a more efficient and, consequently, a more valid representation of the ability of this group in the relative scales.

Table 5.12*Elaboration of the Comparison Between Defined Guess and GIP3A_{p = 0.6} by Quartile/Decile*

Simulation p = 0.6	Group	Proportion of items guessed (%)	Count of Actual Random Guesses	Count of Guesses recovered by GIP3A strategy	Recovery rate Identified/ actual
SIM1, Normal, 400,40	Q4, most able	1.8	72	0	0.0%
	Q3, able	11.2	446	59	13.2%
	Q2, less able	12.0	478	195	40.8%
	Q1, least able	19.4	775	537	69.3%
	Overall		1771	791	44.7%
SIM2, Normal, 250,20	Q4, most able	1.1	14	0	0.0%
	Q3, able	6.0	74	0	0.0%
	Q2, less able	9.2	114	20	17.5%
	Q1, least able	12.7	158	84	53.2%
	Overall		360	119	28.9%
SIM3, Normal, 1000,40	Top decile	0.0	14	0	0.0%
	Decile 9	0.7	98	0	0.0%
	Decile 8	2.0	259	0	0.0%
	Decile 7	3.3	400	28	7.0%
	Decile 6	4.0	481	128	26.6%
	Decile 5	4.6	438	177	40.4%
	Decile 4	5.0	473	237	50.1%
	Decile 3	6.6	591	407	68.9%
	Decile 2	7.5	791	502	63.5%
	Decile 1	9.3	759	572	75.4%
Overall		4304	2051	47.6%	
SIM4, Easy, 400,40	Q4, most able	1.8	72	0	0.0%
	Q3, able	5.4	217	7	3.2%
	Q2, less able	9.7	389	150	38.6%
	Q1, least able	15.8	633	432	68.2%
	Overall		1311	589	44.9%
SIM5, Hard, 400,40	Q4, most able	7.6	303	0	0.0%
	Q3, able	10.4	417	90	21.6%
	Q2, less able	18.1	722	409	56.6%
	Q1, least able	18.1	722	532	73.9%
	Overall		2164	1031	47.6%

5.8.2.2 Analysis of the Pattern of the GIP Indicated Guess

Although in the simulation design the definition of guessed responses was generated by a random algorithm, the pattern of GIP-identified probable guesses was very Guttman-like (see Tables 5.13 and 5.14); in other words, as item difficulty location increased so did the identified number of guesses as ability decreased (indicated by the raw score). This supports the contentions that there are relationships between ability, item difficulty, and guessing, and that the GIP procedure generated the expected outcomes regarding the interactions between the ability, item location, and guesses identified.

Given the guiding premise of this research – that guessing will increase as item difficulty increases – it was expected that the GIP procedure would identify an increasing number of item/student responses as item difficulty increased. Tables 5.13 and 5.14 confirm this expectation. Specifically, they highlight the relationships between the item locations, the student abilities, and the rate of probable guessing identified by the GIP process, showing that as the item difficulty location increased the GIP procedure identified an increased number of probable guesses.

Further there tended to be an increased number of guesses identified as student ability decreased (i.e., the protocol has a greater capacity to identify guesses among the lower-ability students compared to those of higher ability). This is consistent with the hypothesis that there will be an increasing proportion of guessed responses in the responses of lower-ability students.

As a confirmatory strategy, a supplementary analysis was undertaken to investigate the interaction between the GIP procedure with the student item responses, to assess the relationships between student ability, item difficulty, and guesses identified by the protocol. These analyses were interpreted with respect to the quality in the fit of the revised data to the RM (i.e., did the protocol provide a better fit of the data to the model than an INIT analysis). The results of these analyses of the SIM1 and SIM3 data are shown in Tables 5.13 and 5.14.

5.8.2.3 Alternate Response Probability (p Value) Investigations

The selection of $p = 0.6$ as the response probability (RP) was arbitrary. Other values may have been selected however the purpose of defining an alternative response probability was to assess the efficiency of the action. This component of the research was conceptual rather than a definitive resolution.

It was also appreciated that, given the probability of success for a correct response was defined as 0.25, in adjusting the RP there was a fundamental change in the proportion of the student success that was related to the changed RP and a lessening of the proportion of the success that may be attributed to student knowledge or learning or other factors. The impact of the increase in the RP is detailed in Appendix D and discussed in the Areas for Further Research section of Chapter 11.

Appendix D has been provided to show the outcomes of introducing RP values of 0.62, 0.65 and 0.70 respectively for the responses defined in SIM1 and SIM3. This allows comparison with the outcomes shown in Table 5.13 and Table 5.14 below.

Table 5.13

SIM1 Extract of GIP3A Indicated Guesses by Ability Quartile and Item Location $p = 0.6$

δ	-2.66	-1.69	-1.73	-1.75	-1.35	-1.47	-1.13	-1.22	-1.01	-1.03	-1.1	-0.68	-0.8	-0.72	-0.6	-0.67	-0.58	-0.43	-0.27	-0.04	0.123	0.233	0.477	0.665	0.801	0.899	0.842	0.898	0.941	0.928	1.026	1.233	1.153	1.142	1.527	1.283	1.58	1.594	1.697	1.874						
Quartile	Q01	Q04	Q03	Q02	Q06	Q05	Q08	Q07	Q11	Q10	Q09	Q13	Q14	Q12	Q16	Q15	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q27	Q28	Q26	Q30	Q29	Q31	Q34	Q33	Q32	Q36	Q35	Q37	Q38	Q39	Q40	Total					
Q4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Q3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	3	0	5	9	9	12	17	59					
Q2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	2	1	6	3	6	7	7	14	14	9	19	17	15	25	25	18	195						
Q1	0	0	0	0	0	0	0	0	1	1	1	3	2	0	4	1	1	7	9	16	20	23	23	17	24	26	33	23	29	29	28	29	27	23	26	23	26	20	24	18	537					
ALL $p=0.60$	0	0	0	0	0	0	0	0	1	1	1	3	2	0	4	1	1	7	9	16	20	23	25	22	26	27	39	28	35	36	35	44	42	35	45	45	50	54	61	53	791					

Note. Items are ordered by item location (δ) from easiest to hardest, and students are grouped in quartile groups ranked by raw score from highest to lowest.

Table 5.14

SIM3 Extract of GIP3A Indicated Guesses by Ability Decile and Item Location $p = 0.6$

δ	-2.23	-1.91	-1.9	-1.88	-1.7	-1.68	-1.68	-1.4	-1.27	-1.06	-1.06	-0.98	-0.89	-0.78	-0.59	-0.37	-0.37	-0.36	-0.22	0.078	0.369	0.375	0.542	0.567	0.841	0.876	0.88	0.953	1.005	1.051	1.07	1.085	1.337	1.426	1.51	1.531	1.56	1.636	1.795	1.84	Total					
Decile	Q03	Q02	Q07	Q05	Q01	Q04	Q09	Q11	Q06	Q10	Q13	Q08	Q12	Q15	Q17	Q18	Q16	Q14	Q19	Q20	Q21	Q22	Q24	Q23	Q26	Q25	Q27	Q28	Q30	Q31	Q29	Q32	Q34	Q33	Q38	Q35	Q36	Q37	Q40	Q39	Total					
Decile 10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Decile 9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Decile 8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Decile 7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	12	28		
Decile 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	19	17	19	24	25	19	128					
Decile 5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	25	17	26	23	24	19	18	177					
Decile 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	1	4	9	7	17	16	21	22	22	25	19	14	18	19	18	237						
Decile 3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	9	19	25	31	20	30	20	32	30	23	18	27	17	24	27	21	21	407						
Decile 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	8	28	25	31	18	23	19	23	28	23	31	24	28	22	20	34	26	26	21	24	18	502					
Decile 1	0	0	0	0	1	1	1	0	3	3	3	4	2	9	7	21	14	23	15	20	17	20	24	27	34	17	22	24	23	27	15	26	22	22	25	19	17	25	19	20	572					
ALL $p=0.60$	0	0	0	0	1	1	1	0	3	3	3	4	2	9	7	21	14	23	17	28	45	45	68	54	81	62	80	81	83	95	87	105	116	110	147	124	123	139	143	126	2051					

Note. Items are ordered by item location (δ) from easiest to hardest, and students are grouped in decile groups ranked by raw score from highest to lowest.

Note. This table provides an extract of the count of GIP-identified guesses in the same structure as Table 5.12, for the SIM3 data. The outcomes are very similar to the Guttman-like pattern and the increasing number of items identified as guessing as item difficulty increases and ability decreases.

5.9 Implementation of Phase 3 of the Plan for Analysis

5.9.1 *The Phase 3, GIP Process*

The results reported in Section 5.8 show the increase efficiency of the $p = 0.6$ constraint in association with the GIP thresholds. The following sections show the results of implementing the procedure with each of the data sets.

To reiterate, the developed GIP procedure that was implemented in this third phase of analysis had two components:

1. Phase 3A: the identification of probable guesses using the protocol, with a primary goal of re-calibrating the item locations. This process has the capacity to produce a student estimate in which the identified guesses are suppressed and “scored” as missing, resulting in reduced student raw scores compared to those in the INIT data.
2. Phase 3B (also termed GIPINIT3B) was the re-calibration of student ability estimates using GIP item locations produced in the GIP3A procedure in conjunction with INIT response data (raw score). This analysis was achieved by “anchoring” the item locations with the GIP3A-determined parameters and implementing the Rasch analysis with the student INIT data.

Given the outcomes of the GIP item identification process detailed in Section 5.8.2.1, it was anticipated that the GIP process would result in more difficult items being re-calibrated as even more difficult than in the initial calibration. This would be as a result of the identification of guessing in the lower-ability student responses, which are typically associated with the more difficult items.

In relation to the measurement scale generated, the re-calibration of the item locations was expected to increase the distribution of the item locations, with a consequent impact on student ability estimates. The increase in the distribution of items (centred on zero by definition) means the harder items would be re-calibrated with higher difficulty locations and the easy items display lower locations. It was also expected that the item location/student ability estimate of the RM would result in the ability of lower-ability students to be reported at a lower value than the INIT estimate, and the ability of the higher-ability students being reported at a higher value than the INIT estimate. The outcomes of these analyses were based on conditioned data with the $GIP_{p=0.6}$ application of the GIP procedure, thus drawing comparisons between all analysis phases (from INIT to GIPINIT3B).

5.9.2 *PfA Step 3: Comparison of Rasch Analysis Results: INIT vs GIP3A and GIP3B*

In this section, only the results of the INIT analysis and each of the GIP3A and GIPINIT3B analyses are discussed. The impact of the GS analysis has been previously demonstrated in Figures 5.10 and 5.13, and it is not repeated in the figures below. The item locations shown in the GIP3A and GIPINIT 3B graphs were identical because the GIP analysis has these locations anchored.

Given the outcomes reported, it was expected that the suppression of the indicated guessed item from the GIP process would have a consistent impact on the item calibration and relative difficulty locations and a varied impact with respect to the ability estimates of the GIP3A and GIPINIT3B outcomes. The expected impact on the student ability estimates was likely to be varied. It was expected that the GIP3A analyses would result in a lower mean ability estimate of the group accompanied by a significant increase in the distribution of estimates about that mean. Both the maximum and minimum observed ability estimates for the GIP3A analysis were anticipated to be greater than the INIT analysis. The GIPINIT3B results were likely to be varied because of the relative target of the test and the capacity of the GIP process to indicate guesses, particularly with more easy test items. Consequently, the impact of reintroducing the INIT raw score data was expected to have a varied impact on the final outcomes.

Figure 5.16 shows the comparison between the INIT analysis outcomes with the GIP3A outcome (identified guesses suppressed) and the GIPINIT3B outcome (anchored GIP locations with INIT data) for the SIM1 data. The GIP3A analysis suppressed the GIP identified guesses and consequently reduced the GIP3A raw score for each student by the total of the guesses identified. This resulted in the lower-ability students being reported with a considerably lower mean ability estimate – up to approximately three logits lower. By comparison, the higher-ability students benefitted from a more accurate estimate of item locations by up to approximately two logits.

The GIPINIT3B analysis enabled the higher-ability students to have their true ability recognised as a consequence of the re-calibrated item locations. However, re-introduction of the INIT data means that the lower-ability students were credited with the probable guesses as correct answers. Consequently, the degree to which the distribution of the lower ability estimates increased was reduced compared to the GIP3A analysis. Nevertheless, the lower-ability students were still reported with a lower ability estimate than the original INIT analysis.

Figures 5.16 to 5.20 confirm the expectations of item location re-calibrations and their impact on the distribution of student ability estimates for each of the simulated data sets.

5.10 Results of the Implementation of Phase 3 of the Plan for Analysis

Figures 5.15 to 5.19 and the subsequent tables (Tables 5.15 to 5.19) detail the results of the analyses of the simulated data with a p value of 0.6.

Figure 5.15

SIM1 Comparison of Item/Student Maps for INIT Analysis, GIP3A Analysis, and GIP3B Analysis

Phase 1 – Defined Guess scored as 1 INIT analysis	Phase 3A – indicated Guess suppressed - Item calibration analysis GIP3A _{p=0.6}	Phase 3B – indicated Guess scores correct – INIT data with GIP _{p=0.6} item locations (GIPINIT3B)
SIM1. Normal, 400 cases, 40 items		

Note. These item/student maps compare the INIT analysis outcomes with the outcomes of the GIP3A analyses with $p = 0.6$ and GIPINIT3B locations with the item locations of the GIP3A analysis anchored with the data from the INIT student responses. Hence, the item locations of the GIP3A analysis are the same as those in GIPINIT3B but the student ability distributions are different, as anticipated.

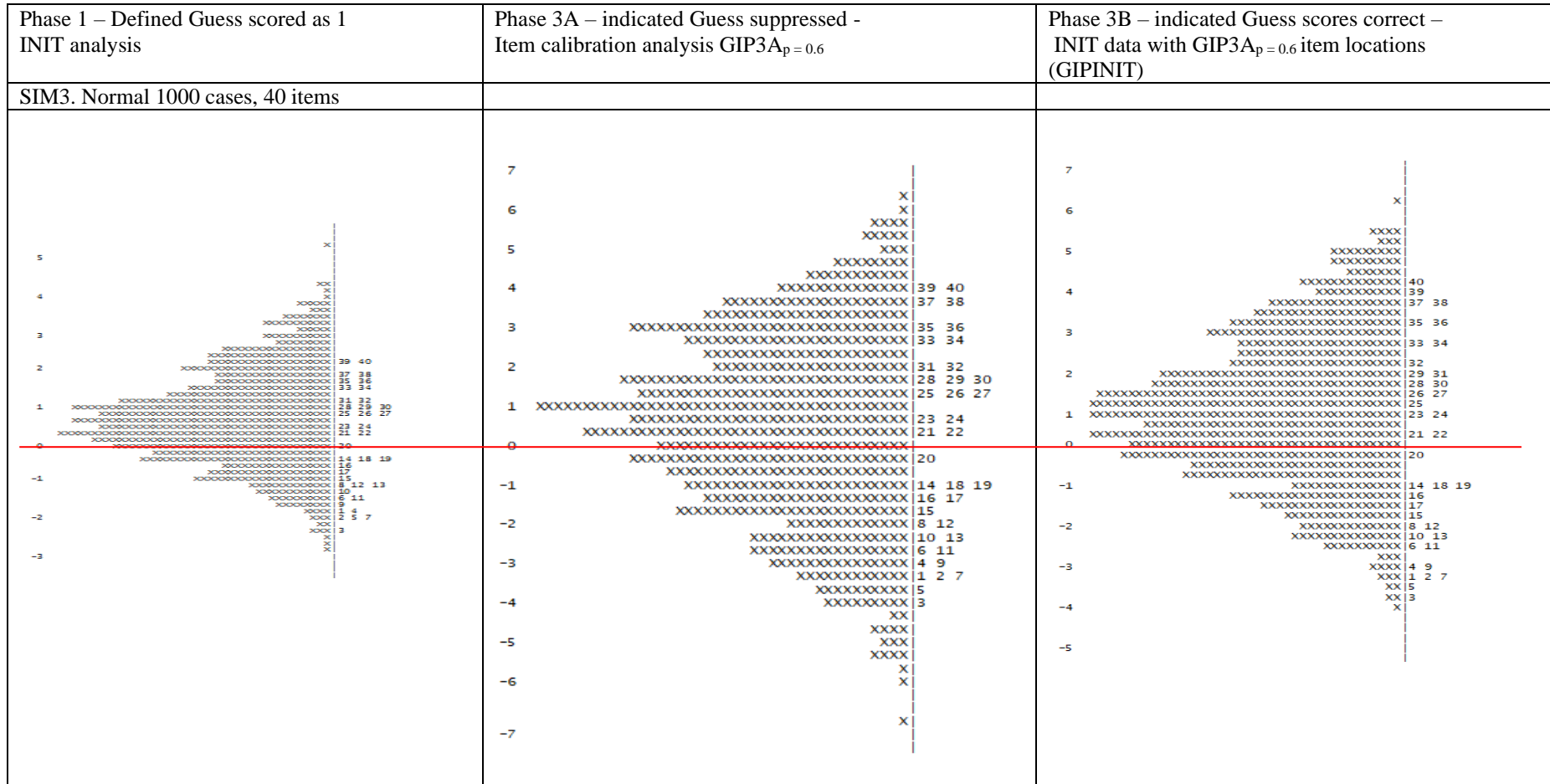
Figure 5.16

SIM2 Comparison of Item/Student Maps for INIT Analysis, GIP 3A Analysis, and GIP 3B Analysis

Phase 1 – Defined Guess scored as 1 INIT analysis	Phase 3A – indicated Guess suppressed - Item calibration analysis GIP3A _p =0.6	Phase 3B – indicated Guess scores correct – INIT data with GIP3A _p =0.6 item locations (GIPINIT)
SIM2. Normal 250 cases, 20 items		

Figure 5.17

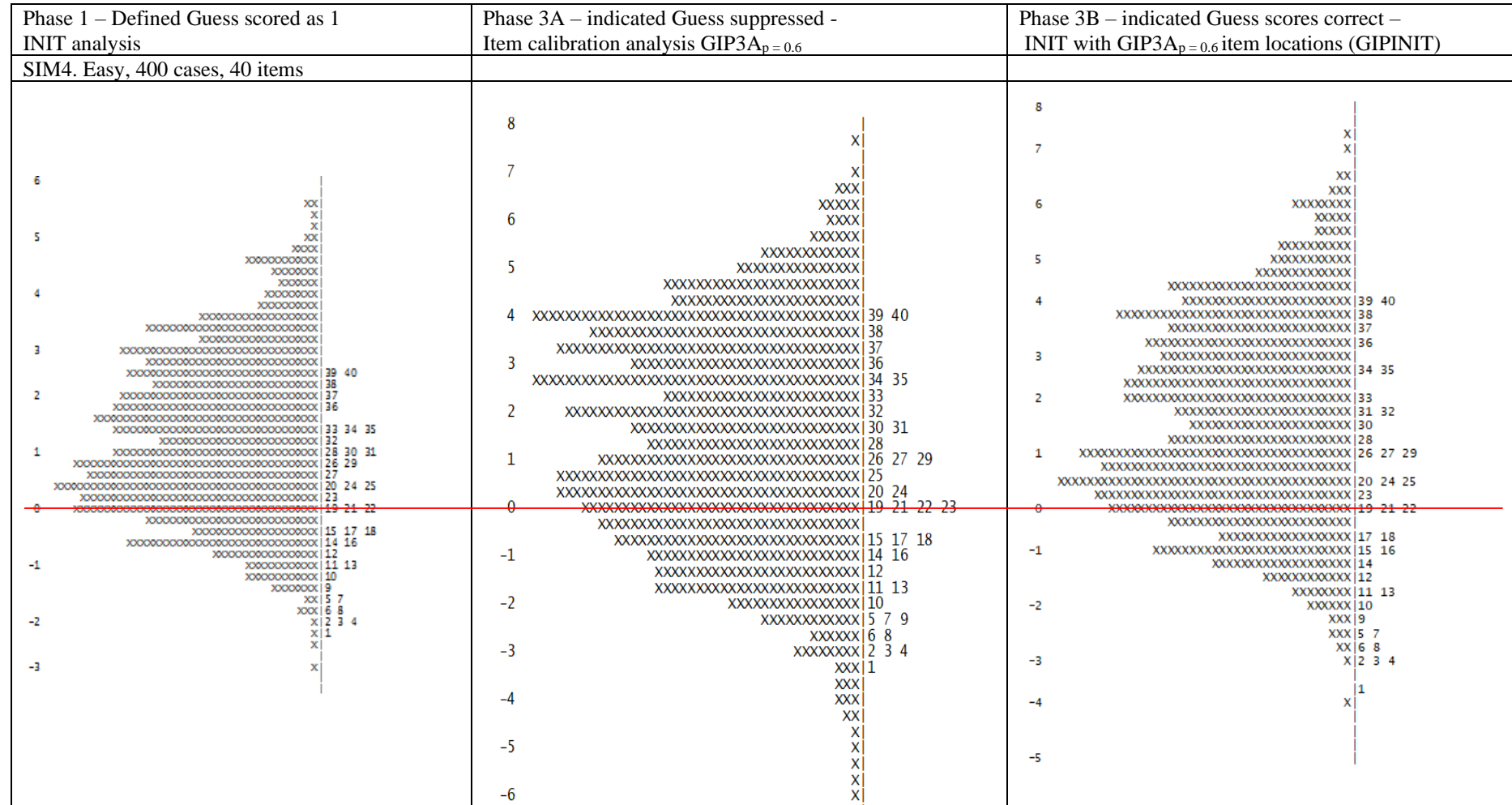
SIM3 Comparison of Item/Student Maps for INIT Analysis, GIP Analysis, and GIPINIT Analysis



Note. This figure shows that the $p = 0.6$ constraint had a similar impact in the SIM3 data as that observed in the SIM1 data. The range of the item locations of the GIP $p = 0.6$ increased compared to the $p = 0.5$ INIT analysis, and the range of the student ability estimates was also increased. There was a greater discrimination and a higher maximum achieved among the higher-ability students than was found in the results summarised in Figure 5.13.

Figure 5.18

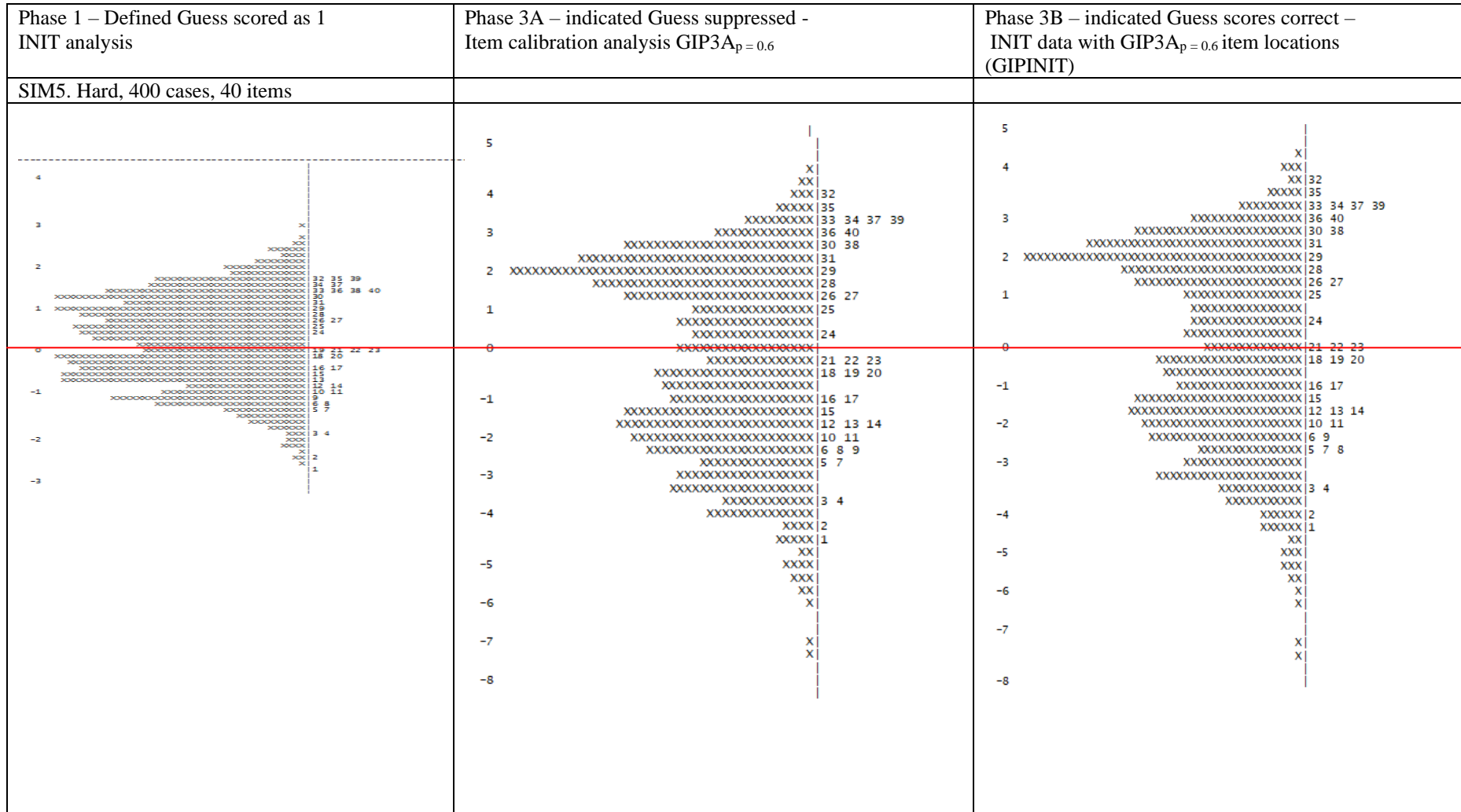
SIM4 Comparison of Item/Student Maps for INIT Analysis, GIP Analysis, and GIPINIT Analysis



Note. The SIM4 data reflect a test that was too easy for the cohort. The results shown for the GIP $p = 0.6$ analysis mirror the results of SIM2, which had a similar test/cohort interaction. The impact was not as marked in SIM4 as in SIM2. This relationship may indicate that the effect was limited as the sample size increased.

Figure 5.19

SIM5 Comparison of Item/Student Maps for INIT Analysis, GIP Analysis, and GIPINIT Analysis



Note. This figure shows the introduction of the $p = 0.6$ probability of a correct response constraint, when combined with the GIP parameters, for a hard test. It illustrates that this implementation had the effect of increasing the range of item difficulties and the range of student ability estimates in the GIP3A and GIPINIT3B outcomes compared to the INIT Rasch analysis.

5.10.2 Comparison of Rasch Results for the Four Analysis Phases

5.10.2.1 Item Statistics

Table 5.14 provides a summary of the differences in item statistics for the SIM1 data for each of the four analyses applied: the INIT analysis; the GS analysis; and GIP analyses for each of the $p = 0.6$ probabilities. It is notable that the mean location and standard deviation are identical for each of the GIP Phase 3A (GIP3A) and GIP Phase 3B (GIPINIT3B) analyses, because of the anchoring of the GIPINIT locations to the GIP values. However, the Fit Residuals of the GIPINIT3B analysis were consistently higher than the GIP comparable statistics as a result of the interaction with the INIT student response raw score data.

Comparisons between “perfect” identification of data (GS) and the initial summary statistics (INIT analysis) in Tables 5.15 to Table 5.19 show a significant improvement in the skewness of the data and an increase in the range of the item difficulties in the GS analysis, which accounts for the defined guessing. The comparison between accounting for guessing achieved by implementing the identification parameters (GIP3A $_{p=0.6}$ and GIPINIT3B $_{p=0.6}$) to the original data analysis (INIT) also show a reduction in the skew of the data as a result of the suppression of the indicated guesses and an increased range of difficulties, in the approximate range of $-3 < \delta < 3$ compared to the INIT values, which provides for location estimates up to three deviations from the mean.

The GIP process provides a measurable improvement in the scale, which is a function of item difficulty independent of the participants, (upon which student ability estimates are calibrated) over the INIT scale which takes no account of guessing in the student responses.

Table 5.15

Comparison of SIM1 Item Statistics for the Four Analyses

Item Statistics	ITEMS: INIT analysis – Guessing ignored		ITEMS: Guess Suppressed (GS) Defined guess missing		ITEMS: GIP3A analysis – Indicated Guess suppressed $p = 0.6$		ITEMS: GIPINIT3B analysis – INIT data scores $p = 0.6$	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	0	0.407	0	0.615	0	0.317	0	3.550
SD	1.188	2.169	2.883	1.688	2.169	1.856	2.169	2.181
Skewness		0.646		-0.175		0.023		0.465
Kurtosis		-0.927		-1.590		-1.468		-1.386
Correlation		0.703		-0.735		-2.673		0.801
Min δ	-2.660		-4.642		-3.749		-3.749	
Max δ	1.874		5.888		3.996		3.996	
range δ	4.534		10.53		7.745		7.745	

Note. The INIT column shows the results of the initial analysis that took no account of guessing. The Guess Suppressed (GS) column shows the results when all defined guessing is recoded as missing. The GIP3A $_{p=0.6}$ and GIP3B $_{p=0.6}$ columns display the results when the resolved identification parameters have been applied to the data.

Table 5.16*Comparison of SIM2 Item Statistics for the Four Analyses*

Item Statistics	ITEMS: INIT analysis – Guessing ignored		ITEMS: Guess Suppressed (GS) Defined guess missing		ITEMS: GIP3A analysis – Indicated Guess suppressed $p = 0.6$		ITEMS: GIPINIT3B analysis – INIT data scores $p = 0.6$	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	0	-0.682	0	-0.427	0	-0.674	0	1.580
SD	1.365	2.674	1.785	2.622	2.240	0.931	2.240	2.753
Skewness		0.894		0.636		0.954		0.908
Kurtosis		-0.507		-1.280		0.102		-0.924
Correlation		0.677		0.651		0.336		0.779
Min δ	-2.373		-3.113		-3.368		-3.368	
Max δ	1.664		2.356		3.199		3.199	
range δ	4.037		5.469		6.567		6.567	

Table 5.17*Comparison of SIM3 Item Statistics for the Four Analyses*

Item Statistics	ITEMS: INIT analysis – Guessing ignored		ITEMS: Guess Suppressed (GS) Defined guess missing		ITEMS: GIP3A analysis – Indicated Guess suppressed $p = 0.6$		ITEMS: GIPINIT3B analysis – INIT data scores $p = 0.6$	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	0	0.370	0	0.937	0	0.216	0	5.226
SD	1.260	4.029	3.084	2.318	2.254	2.818	2.254	3.671
Skewness		0.504		-0.082		0.472		0.361
Kurtosis		-0.648		-0.736		-0.963		-1.355
Correlation		0.601		-0.721		-0.734		0.775
Min δ	-2.226		-4.187		-4.832		-3.246	
Max δ	1.870		6.333		5.150		5.150	
range δ	4.096		10.520		9.982		8.396	

Table 5.18*Comparison of SIM4 Item Statistics for the Four Analyses*

Item Statistics	ITEMS: INIT analysis - Guessing ignored		ITEMS: Guess Suppressed (GS) Defined guess missing		ITEMS: GIP3A analysis – Indicated Guess suppressed $p = 0.6$		ITEMS: GIPINIT3B analysis – INIT data scores $p = 0.6$	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	0	0.206	0	0.600	0	0.267	0	2.697
SD	1.263	2.399	2.520	2.142	2.001	2.224	2.001	2.356
Skewness		0.343		-0.600		-0.376		0.695
Kurtosis		-0.470		-1.210		-1.291		-0.131
Correlation		0.141		-0.731		-0.762		0.568
Min δ	-2.310		-3.526		-3.049		-3.049	
Max δ	2.211		5.813		3.961		3.961	
range δ	4.521		9.339		7.110		7.110	

Table 5.19*Comparison of SIM5 Item Statistics for the Four Analyses*

Item Statistics	ITEMS: INIT analysis - Guessing ignored		ITEMS: Guess Suppressed (GS) Defined guess missing		ITEMS: GIP3A analysis - conditioned data $p = 0.6$		ITEMS: GIPINIT3B analysis - conditioned scores $p = 0.6$	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	0	-0.735	0	0.023	0	-0.329	0	3.210
SD	1.136	4.244	3.273	1.471	2.358	1.886	2.358	3.155
Skewness		0.240		-0.213		0.028		-0.035
Kurtosis		-1.457		-1.231		-0.783		-1.306
Correlation		0.579		-0.586		-0.532		0.694
Min δ	-2.590		-4.272		-3.813		-3.813	
Max δ	1.550		8.352		3.233		3.231	
range δ	4.140		12.624		7.046		7.044	

Results in relation to student ability estimates (summarised in Tables 5.21 to 5.25) are consistent across the four analyses in the following five ways:

1. The distribution of ability estimates was greatest for the GS phase, and in the GIP3A analysis the distribution of ability estimates was greater than the INIT analysis to a lesser degree than the GS result.
2. The distribution of ability estimates for the GIP3A analysis was greater than for the GIPINIT3B analysis, both of which were greater than for the INIT analysis.
3. The reliability statistic for the GIP3A and GIPINIT3B_{0.6} analyses were higher than for the INIT analysis.
4. The mean ability estimates for the GIP3A_{p=0.6} analysis tended to be lower than for the INIT analysis.

5. The mean of the GIPINIT3B analysis was greater than for the INIT analysis, reflecting introduction of the INIT data that credits all correct responses in a more refined GIP scale.

5.10.2.2 Ability Estimate Statistics

Tables 5.20 to 5.24 summarise the observed ability estimate statistics for each of the four analyses of the simulated data. They show a consistent pattern that makes explicit the outcomes observed in Figures 5.15 to 5.19. For each of the data sets, the GS analysis shows significant increases in the distribution of ability estimates around a lower mean value compared to the INIT analysis. This was a consequence of the reduction in the raw scores of the GS analysis compared to the INIT analysis.

The GIP analyses show two consistent patterns. The GIP3A result has a lower mean result for the group as a result of the suppression of the indicated guesses, and a higher distribution of ability estimates about that mean. Both the maximum and minimum estimates are greater than the INIT outcome. By comparison, the mean of the GIP3B analysis generally shows a mean ability estimate higher than the INIT outcome because of the recalibration of the item difficulties and the reintroduction of the initial response data. However, the both the maximum and minimum GIPINIT3B estimates are greater than the INIT outcome.

Table 5.20 displays the outcomes for SIM1. It shows that the GS analysis for the defined guessing resulted in a mean ability approaching zero, compared to 0.72 logits in the INIT analysis. The GS analysis also resulted in a wider distribution of abilities calculated – a range of 13.7 logits – and the extremities being almost 3 logits further distant from without accounting for guessing.

The two components of the GIP analyses (3A and 3B) show a deviation in their ability estimates to a lesser degree than does the GS analysis. Both GIP analyses show a greater range of abilities calculated, a higher maximum ability reported (recognising the ability of higher-ability students), and a lower minimum ability reported (which more accurately positions lower-ability students than parameters of the INIT analysis). Tables 5.21 to 5.24 provide a comparison of the item parameters for each of the analyses and the two versions of the GIP procedures ($p = 0.6$) in relation to SIM2 to SIM5 data sets.

Tables 5.20 to 5.24 provide the results for each of the simulations, with the general pattern of results similar to those observed for SIM1.

Table 5.20

Comparison of SIM1 Student Ability Estimate Statistics for the Four Analyses

Statistics	STUDENTS: INIT analysis - Guess is correct		STUDENTS: Guess Suppressed (GS) - Guess is missing		STUDENTS: GIP3A analysis – Indicated Guess suppressed $p = 0.6$		STUDENTS: GIPINIT3B analysis – INIT responses with GIP3A $p = 0.6$ item locations	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	0.724	-0.310	0.034	-0.265	0.543	-0.254	0.950	0.839
SD	1.390	1.304	3.196	0.647	2.388	0.780	1.963	1.095
Skewness		0.473		0.328		-0.022		0.203
Kurtosis		-0.298		0.196		0.356		-0.379
Correlation		-0.362		0.368		-0.410		-0.523
Min β	-2.615		-6.314		-5.392		-3.417	
Max β	4.646		7.448		5.495		5.004	
range β	7.261		13.762		10.887		8.421	
RELIABILITY INDICES								
Separation Index		0.899		0.958		0.950		0.932
Cronbach Alpha		0.919		N/A		N/A		N/A

Table 5.21

Comparison of SIM2 Student Statistics for the Four Analyses

Statistics	STUDENTS: INIT analysis - Guess is correct		STUDENTS: Guess Suppressed (GS) - Guess is missing		STUDENTS: GIP3A analysis – Indicated Guess suppressed $p = 0.6$		STUDENTS: GIPINIT3B analysis – INIT responses with GIP3A item locations $p = 0.6$	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	1.403	-0.329	1.271	-0.327	1.534	-0.331	1.862	0.136
SD	1.274	0.889	1.691	0.835	2.119	0.515	1.653	0.878
Skewness		1.609		1.028		0.574		1.258
Kurtosis		2.275		1.279		3.670		0.667
Correlation		-0.473		-0.355		0.027		-0.626
Min β	-1.420		-3.140		-3.364		-1.989	
Max β	4.029		4.401		5.004		5.004	
range β	5.449		7.541		8.368		6.993	
RELIABILITY INDICES								
Separation Index		0.725		0.815		0.799		0.797
Cronbach Alpha		0.800		N/A		N/A		N/A

Table 5.22*Comparison of SIM3 Student Statistics for the Four Analyses*

Statistics	STUDENTS: INIT analysis - Guess is correct		STUDENTS: Guess Suppressed (GS) - Guess is missing		STUDENTS: GIP3A analysis – Indicated Guess suppressed $p = 0.6$		STUDENTS: GIPINIT3B analysis – INIT responses with GIP3A item locations $p = 0.6$	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	0.717	-0.326	0.019	-0.223	0.517	-0.299	0.950	0.838
SD	1.302	1.286	3.118	0.621	2.302	0.837	1.859	1.105
Skewness		0.489		0.532		0.088		0.104
Kurtosis		-0.129		0.051		-0.179		-0.519
Correlation		-0.349		0.365		0.402		-0.522
Min β	-2.419		-6.408		-5.640		-3.246	
Max β	3.887		6.828		5.150		5.150	
range β	6.306		13.236		10.790		8.386	
RELIABILITY INDICES								
Separation Index		0.891		0.958		0.948		0.927
Cronbach Alpha		0.907		N/A		N/A		N/A

Table 5.23*Comparison of SIM4 Student Statistics for the Four Analyses*

Statistics	STUDENTS: INIT analysis - Guess is correct		STUDENTS: Guess Suppressed (GS) - Guess is missing		STUDENTS: GIP3A analysis – Indicated Guess suppressed $p = 0.6$		STUDENTS: GIPINIT3B analysis – INIT responses with GIP3A item locations $p = 0.6$	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	1.267	-0.166	1.144	-0.168	1.313	-0.185	1.577	0.786
SD	1.479	1.071	2.834	0.696	2.294	0.810	1.979	1.112
Skewness		0.411		0.311		0.238		0.292
Kurtosis		-0.027		-0.296		-0.867		-0.350
Correlation		-0.175		0.529		0.571		-0.452
Min β	-2.989		-7.471		-5.344		-3.530	
Max β	4.724		7.222		5.926		5.926	
range β	7.713		14.693		11.270		9.456	
RELIABILITY INDICES								
Separation Index		0.889		0.950		0.939		0.921
Cronbach Alpha		0.923		N/A		N/A		N/A

Table 5.24*Comparison of SIM5 Student Statistics for the Four Analyses*

Statistics	STUDENTS: INIT analysis - Guess is correct		STUDENTS: Guess Suppressed (GS) - Guess is missing		STUDENTS: GIP3A analysis – Indicated Guess suppressed $p = 0.6$		STUDENTS: GIPINIT3B analysis – INIT responses with GIP3A item locations $p = 0.6$	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	0.240	-0.217	-0.157	-0.063	-0.257	-0.224	0.305	1.166
SD	1.313	4.244	3.290	0.489	2.249	0.810	1.720	1.224
Skewness		0.240		0.806		0.649		-0.309
Kurtosis		-1.457		1.476		0.413		-0.822
Correlation		0.579		0.156		0.322		-0.678
Min β	-2.599		-6.146		-5.855		-3.642	
Max β	2.520		8.260		3.797		3.797	
range β	5.119		14.406		9.652		7.439	
RELIABILITY INDICES								
Separation Index		0.865		0.952		0.949		0.923
Cronbach Alpha		0.884		N/A		N/A		N/A

Taken together, the improved item performance and largely improved estimates of overall student ability provide evidence to support an evolving conclusion that the GIP procedure yields an improvement in estimating test performance and ability estimates more than a Rasch analysis that does not attempt to take account of guessing.

5.11 Analysis of Fit Statistics

In the analyses reported in Tables 5.20 to 5.24, the GIP_{p=0.6} analysis also had an improved Separation Index statistic compared to the comparable INIT analysis. This statistic is a pseudo Cronbach α that measures the internal reliability of the data in which there were missing values. As a first indicator, this feature suggests that the GIP procedure was an improvement in the measurement of the student estimates compared to the INIT unconditioned data.

In the following the methodology of Andrich et al. (2015), measures of improvement in the fit to the RM are demonstrated through comparisons of the mean square statistic.

$$\text{Mean Square} = \chi^2 / \text{d.f.} \quad \text{Equation 5.6}$$

Where χ^2 is the sum of the squares of the deviations of the observed student/item interactions from the model; and

d.f. is the number of degrees of freedom of the sample.

Applied to the current data, in each data set the mean square fit statistic of the GIP3A_{p=0.6} analysis was lower than the INIT analysis statistic (see Table 5.24 to compare the results of the analyses in terms of fit statistics and calculated mean square statistic for each simulated data set). Andrich et al. (1988) describe this outcome as an improvement in the fit of the data to the model. In cases where the GIP mean square approaches half the INIT statistic, this represents a “substantial” improvement (Andrich et al., 2012) in the fit of the data used to re-calibrate the item locations. These results provide evidence that the GIP3A process produced a more “pure” variable (i.e., a better measure of the construct under investigation).

Table 5.25

Test of Fit for the INIT, GS, and GIP Analyses

Analysis	INIT			GIP3A _{p = 0.6} Guess Suppressed			GIPINIT3B _{p = 0.6} with INIT data		
	Total Chi-Sq	d.f.	Mean Square	Total Chi-Sq	d.f.	Mean Square	Total Chi-Sq	d.f.	Mean Square
SIM1	1519	200	7.6	707	200	3.5	812	200	4.1
SIM2	312	60	5.2	116	60	1.9	793	60	13.2
SIM3	4540	360	12.6	3713	360	10.3	20474	360	56.9
SIM4	1305	200	6.5	1239	200	6.2	4369	200	21.8
SIM5	1859	200	9.3	807	200	4.0	13528	200	67.6

When the initial student responses (INIT data) were re-introduced with the GIP3A item locations (analysis GIPINIT3B), the mean square fit statistic was universally greater than both INIT and GIP3A outcomes.

To reiterate, the GIP3A analysis takes account of guessing and suppresses the probable guessed responses. This means they were scored as missing in the analysis. The impact of that process was to refine the variable by reducing the “noise” caused by the misfitting guessed items and create a better fit of the conditioned data to the model. By anchoring the GIP3A items locations (which redefines a more “pure” measurement scale) and introducing the INIT data, the degree of misfit of the guessed items in the GIPINIT3B analysis was increased and then exacerbated by the summing of the squares of the residuals. Yet, as Andrich (2012) noted:

Within the Rasch measurement paradigm, misfit between the data and the model is seen as an anomaly in the data. ... In the case of random guessing to multiple choice items, to remove the source of anomalies requires better alignment of the difficulty of the items to the proficiency of the students. (p. 420)

Tables 5.15 to Table 5.25 and Figures 5.14 to 5.19 presented above clearly accord with this statement. The INIT analyses contain defined randomly generated guessing and display misfit as an anomaly in the data. By comparison, the GIP3A_{p=0.6} results display better alignment of item difficulty to the achievement of the students, following the removal of the source of the anomalies (guessing as a source of systematic error). Consequently, the because the raw scores vary in the GIP3A procedure, students have revised, and more reliable, ability estimates in a scale that is a better representation the relative difficulties of the items in the trait. The reintroduction of the systematic error by using the INIT raw scores is an artifact of the face validity required when a student correctly responds to an item

irrespective of the ability related source of that response. This anomaly may be addressed in the way results are reported to various stakeholders.

However, it is important to note that the GIPINIT3B outcomes are grounded in an improved measurement scales produced in the GIP3A process. Hence, although the re-introduction of the systematic errors in the INIT data have a negative impact on the fit statistics, the student estimates are based on a better measurement scale and hence an improvement over a scale that does not attempt to account for guessing.

Tables 5.26 and 5.27 summarise the impact of implementing the GIP parameters with each simulation data set. They were designed to provide a summary of the INIT and final products of the application of the proposed GIPINIT3B_{p=0.6} procedure to the data sets. They provide an abridged summary of the relevant parameters from the two analyses.

Table 5.26

Summary of the Results of the INIT Simulated Data Sets

Distribution Target	Sample N	Items				Students			
		St. Dev	Range	Skew	Kurtosis	Mean β	St. Dev	Range	α
Normal	400	1.19	-2.7/1.9	0.65	-0.93	0.72	1.39	-2.6/4.6	0.90
Normal	250	1.37	-2.4/1.7	0.89	-0.51	1.40	1.27	-1.4/4.0	0.72
Normal	1000	1.26	-2.2/1.8	0.50	-0.65	0.72	1.30	-2.4/3.9	0.89
Too easy	400	1.26	-2.3/2.2	0.34	-0.47	1.27	1.48	-3.0/4.7	0.89
Too hard	400	1.14	-2.7/1.9	0.24	-1.46	0.24	1.05	-2.6/2.5	0.86

Table 5.27

Summary of the Results of the GIPINIT3B_{p=0.6} Simulated Data GIP3A Locations with INIT Raw Scores

Distribution Target	Sample N	Items				Students			
		St. Dev	Range	Skew	Kurtosis	Mean β	St. Dev	Range	α
Normal	400	1.90	-3.4/3.3	-0.09	-1.26	0.54	2.39	-3.4/5.9	0.95
Normal	250	1.88	-3.2/2.6	0.09	-1.45	1.86	1.65	-2.0/5.0	0.80
Normal	1000	1.81	-2.9/3.0	0.76	-0.41	0.95	1.86	-3.2/5.2	0.93
Too easy	400	1.81	-2.8/3.5	-0.19	-0.71	1.58	1.98	-3.5/5.9	0.92
Too hard	400	2.36	-3.4/2.8	-0.03	-1.31	0.31	1.72	-3.6/3.8	0.92

Note. α = RUMM Separation Index indicated for incomplete data sets (conditioned data) with missing values (guesses).

Eight conclusions can be drawn from the overall pattern of results summarised in Tables 5.25 and 5.26:

1. The range of item difficulty locations derived from the simulated data was greater when guessing was suppressed by applying the GIP, compared to the INIT values. In these simulated data, the defined guesses of each student in the original response pattern had been indicated and suppressed with an overall recovery rate to the order of 33%. However, in the region where there is likely to be most guessing (the lowest quartile, or lower deciles of the student above), the recovery rate of “indicated” defined guesses was of the order of 60%.

2. The wider distribution of GIP item difficulty locations generated both lower and higher item locations than in the INIT analysis. The consequences of the lower difficulty location for easier items and higher difficulty location for harder items is demonstrated by the method of computing student ability estimates. That is, the interaction between the raw score and the item difficulty, as determined by the RM, results in success on lower difficulty items and thus in a lower ability estimate than when the item difficulty is overestimated. Conversely the interaction between higher item locations and the raw score, in applying the RM, generates higher ability estimates for students who have been successful in a greater number of items.
3. The application of the GIP procedure tended to impact the skew in the data in a negative (beneficial) direction, which accounted for the overestimation of student ability in the unconditioned data.
4. The mean ability of the sample decreased in response to suppression of the guessing in the response patterns. This was not simply a function of the reduced scores produced by re-coding correct guessed responses, it was also a function of the interaction between the ability estimate and the item difficulty (Table 3.2). The GIP procedure resulted in an improvement in the accuracy of estimating student abilities, but to a lesser extent than was observed by suppressing all “defined” guesses.
5. The spread of ability estimates increased to reflect the wider distribution of item difficulties, from which the ability estimates were constructed. The more difficult items led to higher student ability estimates, and success in only the items of lower difficulty reflected a lower ability in the trait.
6. The reliability indicators generated by suppressing guessing – Cronbach’s α in the case of the complete data sets, and Separation Index in the case of the incomplete data – improved when guessing was accounted for using the GIP procedure, compared to the unconditioned initial data.
7. Although the fit of the ability estimates for the GIPINIT3B analysis showed greater misfit than for the INIT analyses, the ability estimate and item fit analysis of the GIP3A analysis, which defines the GIPINIT3B item scale, was an improvement over the INIT analysis. Hence the measurement scale upon which the student achievement was compared appears to be a better indicator of ability than a scale which makes no attempt to account for the guessing in the data.
8. The GIPINIT3B highlighted the misfit introduced by the presence of guessing, which is consistent with the expectations and theory that underpins this research. The misfit observed in the GIPINIT3B analyses does not negate the findings. Rather, it highlights the impact of guessing and supports the contention that this two-stage analysis of the data provides a better measure of the students in the trait, provided the items that are producing the misfit – the GIP indicated probable guessing – are reported to the stakeholders for further consideration.

The application of the GIP process to account for guessing did not represent a procedure to adjust a model to fit the data, as is the cases in the 3PL IRT model discussed in Chapter 4. Rather, it represents the application of a logic to refine the variable by omitting responses which are highly likely to be a corruption of the requirements of the model, and hence develop an improved measurement variable – a removal of the systematic error.

The impact of the application of the GIP to the simulated data provided a better alignment between the item difficulties and the abilities of the students, as summarised in Tables 5.25 and 5.26. However, there were limiting factors in the capacity of the GIP that restricted the practical identification of every case of defined guessing generated in the simulated data. These will be discussed in Chapter 11.

5.12 Summary

The aim of Study 1, the analysis of the simulated data, was to develop a set of parameters that consistently identified a defined guess in the simulated data. The developed GIP could then be used in investigations of live data. This chapter detailed the process by which the GIP process was determined. The outcome of the analyses identified a set of parameters that were efficient in identifying previously defined guesses without any Type 1 errors. These processes and parameters constitute the GIP that will be the focus of the next chapters, which will evaluate its ability to indicate and account for guessing in student data. The results from implementing the GIP procedures (3A and INIT3B) to the simulated data sets are consistent with the expected outcomes the hypothesis of the research study as a whole.

Given the pattern of results derived from the sequence of the planned analyses, the parameters tested and reported in this chapter were used for subsequent investigations and development of the GIP. In summary, likely guessing is inferred in the GIP process when:

- the $p = 0.6$ constraint is applied in estimating the probability of a correct response and in calculating parameters 2 and 3 below;
- $\Pr(1)_{p=0.6} < 0.25$; and
- item/student residual > 1.75 .

Chapter 6

Study 2: Investigating the Proposed GIP Process Using Small-Scale Field Data

6.1 Introduction

The purpose of Study 2 was to review the application and effectiveness of the GIP parameters – developed with simulated data – for indicating probable guessing when using a set of real-life educational assessment data. As outlined in the Methodology of Chapter 4, this Study involved the administration of an Arabic version of a mathematics achievement test to students of Year 5 and Year 7 designed to increase the rate of guessing while introducing a mechanism to identify the self-indicated guesses. Following the administration of the Arabic tests, a English versions of the same tests was administered to the same students (adopting a within-subjects design), to enable the collection of a set of baseline data for comparison with the Arabic version that was designed to engender guessing.

To briefly reiterate the methodology, two convenience samples (total $N = 492$ students) sat a Mathematics test developed from previously trialled items, which had functioned uniformly when administered in large-scale assessments. Consistent with recommendations of Wright and Stone (1988), prior administration made it possible to order the items in the tests by item facility, from easy to hard. This structure was intended to maximise the potential for Guttman-like response patterns if the assumptions regarding the interaction between students and items that underpin Rasch Theory were observed in these real-world data. The tests were constructed using only multiple-choice items with four distractors. In responding to the items, students were asked to answer each item by filling in a response “bubble”. The only deviation from “normal” practice was that the students were required to answer with different coloured pens to self-identify guessing in the response data (Green (G) = known answer, Yellow (Y) = unsure response, and Pink (P) = Guess). Reference to the likely guessed items nominated by the GIP process uses the verb “indicate” in the reporting of these results. A differentiation is made between students “identifying” their self-identified responses and the guesses “indicated” by the GIP process.

For each stage of Study 2, this chapter follows the structure of Chapter 5.

6.2 Plan for Analysis (PfA) for the English Versions of the Tests

1. The analysis of the English versions of the tests are presented before the Arabic analysis outcomes, as the English versions were assumed to provide a better estimate of baseline data against which the results of the Arabic version could be compared. Analyses of these field data followed a similar plan as for the simulated data; separately for each language version of the tests and for each of the Year 5 and Year 7 students who engaged with the items. These followed processes similar to those developed with the simulated data:

3. For each test and Year level, data were first analysed using traditional techniques to investigate the item performance and the relationships between the self-identified guesses and the item difficulty expressed as facility (percent correct).
4. Each data set was then analysed using the Rasch model (RM) (see Table 4.1), which initially took no account of the self-identified guesses (INIT), followed by an analysis that accounted for the self-identified guesses (SIG, analogous to the GS analyses of the simulated data). The purpose of this analysis was to determine the extent to which accounting for the self-identified guesses in each data set impacted the item difficulty locations and relative ability estimates of the students. It also enabled comparison to the results from the simulated data to observe if these live data generated similar outcomes, as a further evaluation of the GIP procedure.
5. The third step involved the application of the GIP procedures (Phases 3A and 3B) with the original data. In this step, a correlation between the items indicated by GIP were compared to the self-identified guessing to evaluate the efficacy of the protocol.
6. Supplementary analysis of the frequencies of the GIP-identified items and item fit were conducted to assess the degree to which the outcomes accorded with the expectations of the hypotheses and the outcomes found with the simulation data.
7. Concluding comments are provided regarding how Study 2 contributes to the overall program of research, in terms of the prior findings and plan for subsequent studies.
8. The remainder of this chapter reports on the background, specifications, and results for each of the analytical steps.

6.3 Study 2, Stage 1, English Versions of the Tests

6.3.1 Pfa Step 1 Analysis – Investigation of Self-Identified Guesses

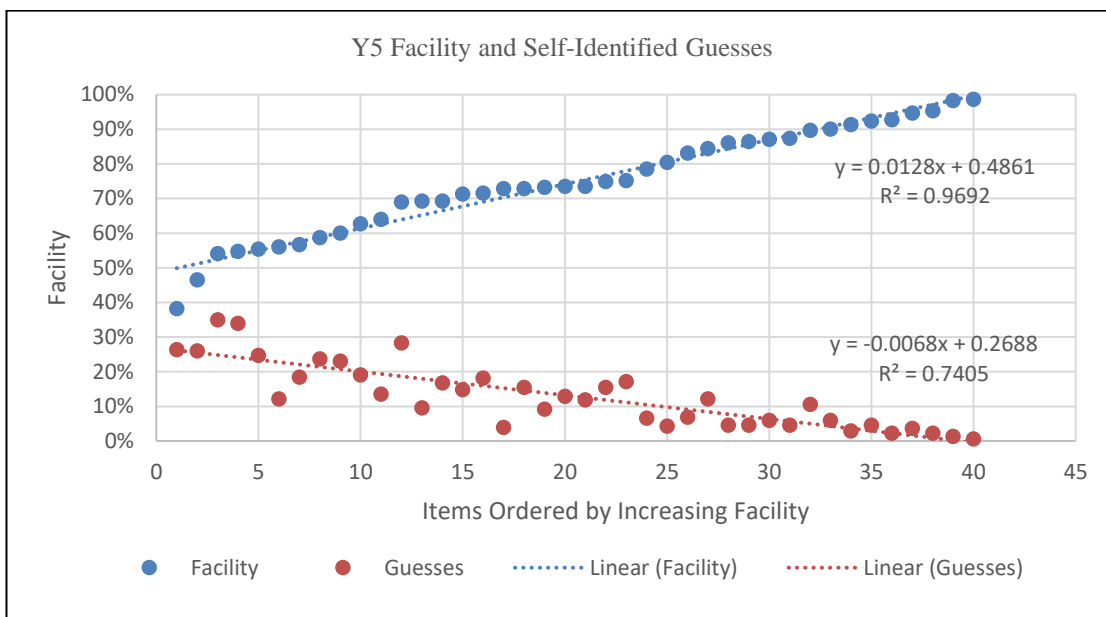
For both the Year 5 and Year 7 levels, the relationship between item difficulty and self-identified guesses was generally consistent with results from the simulated data analyses and the hypothesis. That is, as item difficulty increased so did the proportion of guesses in the student response patterns.

6.3.1.1 Year 5

Results of the relationship between the proportion of guesses and the facility of each item for the Year 5 sample conformed to the expected pattern (Figure 6.1). Although there was some variation in the proportion of guessing relative to the facility of each item (Figure 6.2), the R^2 of the guessing distribution indicated that more than 70% of the variation in the self-identified guessing was explained by the facility of the item. Figure 6.1 shows the variability in the proportion of guesses and the relative proportion of successful guesses for each item.

Figure 6.1

Year 5 Relationship Between Item Facility and the Percentage of Self-Identified Guesses



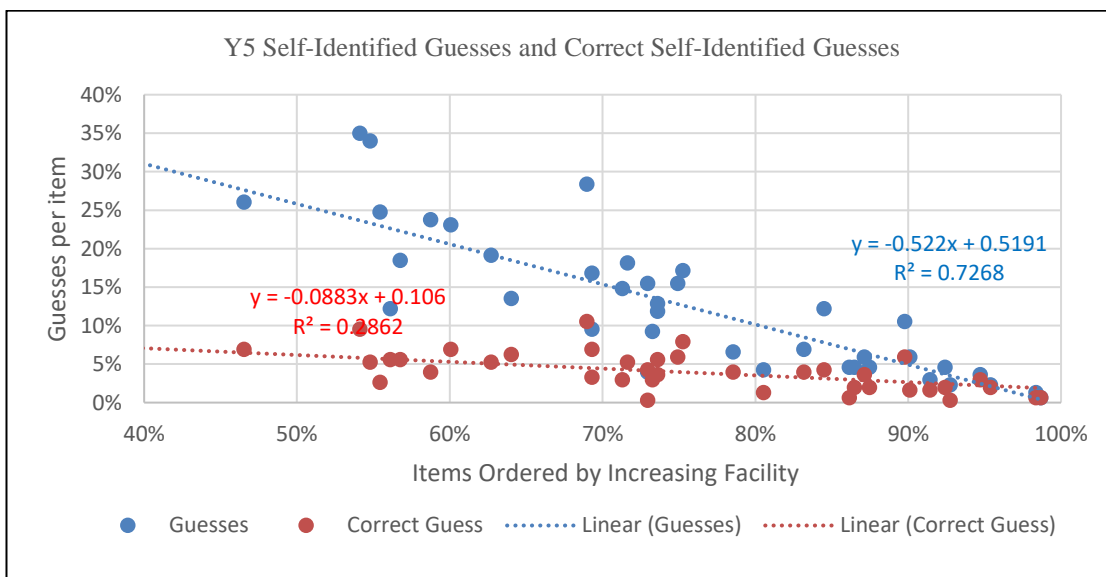
Note 1. The observed Self-Identified Guesses (SIG) represents the proportion of self-identified guesses.

Note 2. The items have been ordered from most difficult (lowest facility) to least difficult.

In reviewing the self-identified guesses of Year 5 students that were successfully answered (Figure 6.2), it is interesting to note that for the hardest items the success rate of self-identified guesses was approximately the expected chance rate (25%). As the items became easier, the success rate was considerably higher, which suggests that as item difficulty decreased the self-identified guesses were not random.

Figure 6.2

Year 5 Relationship Between the Percent of Self-Identified Guesses and the Percent of Correct Guesses



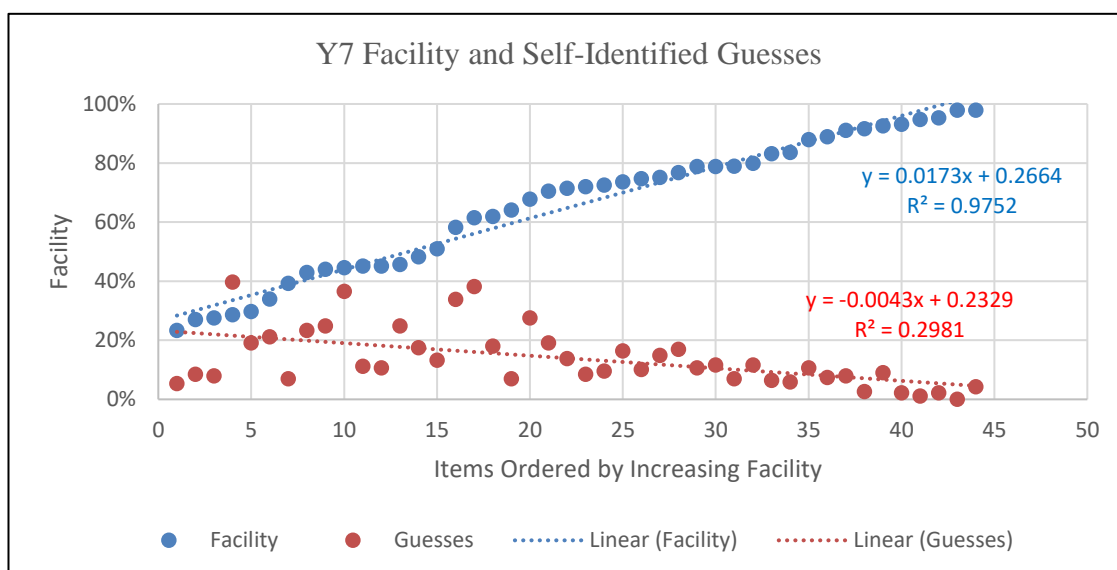
Note. Items ordered from hardest (lowest facility) to easiest

6.3.1.2 Year 7

Whereas the Year 5 patterns of the item/guess interactions were relatively uniform, there was greater variation within the Year 7 response data, especially in relation to the most difficult items (Figure 6.3). However, this may reflect the higher proportion of non-attempts (13%) associated with those items. It is notable that for the 20 most difficult items the proportion of self-indicated guesses tended to be at or above 25%. As items became easier, this proportion reduced considerably with less variability but was well below the expected rate of 25%. This supports the inference from the Year 5 results that as items became easier the guessing was less random and possibly reflected partial knowledge.

Figure 6.3

Year 7 Relationship between item Facility and the percentage of Self-Identified Guesses

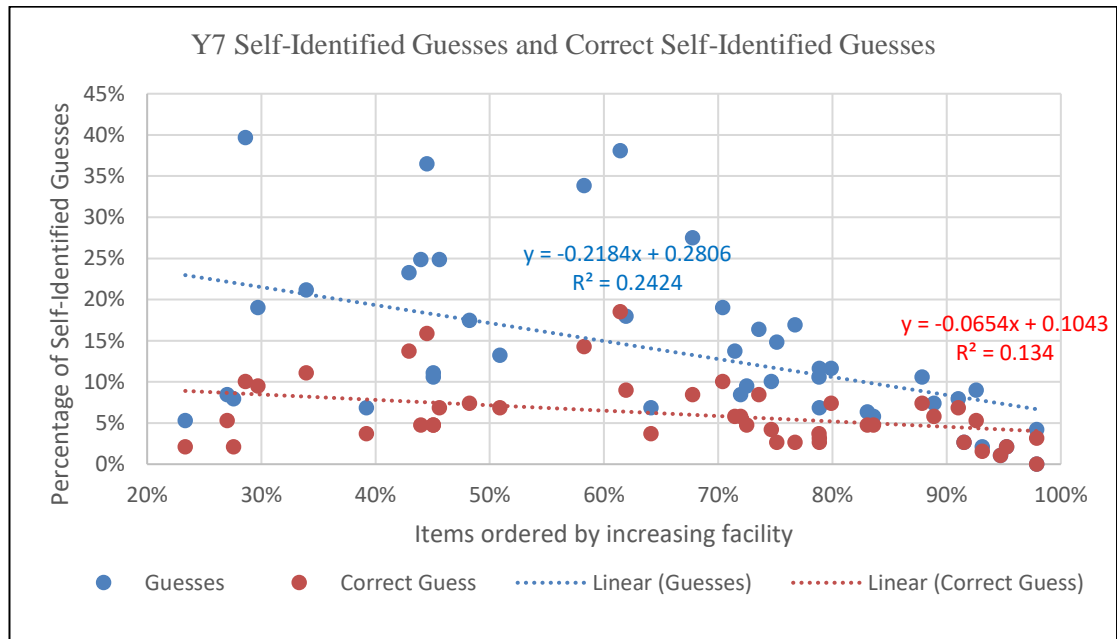


Note. This figure shows the relationship between the total number of guesses and the number of correct guesses for Year 7. It highlights the variability in response patterns for “authentic data”, in which the students’ pattern of guessing did not necessarily accord with theoretical expectations.

For the Year 5 sample, the value of the correlation between the observed raw score and the number of incorrect “known” answers was $r = -0.78$, indicating that lower-ability students suggested they knew the answer when it was in fact incorrect more frequently than the higher-ability students. For Year 7, the comparable statistic was $r = -0.18$; however, the correlation between raw score and non-responses (missing) was $r = -0.73$, indicating that higher-ability students tended to omit items more frequently rather than attempting a guess. This was not evident in the Year 5 responses.

Figure 6.4

Year 7 Relationship Between the Percentage of Self-Identified Guesses and the Percentage of Correct Guesses



These data reflect the real-life responses that do not accord exactly with theoretical patterns demonstrated in the simulated data, and they highlight the need to evaluate theoretically derived solutions with authentic data. In order to further understand the patterns of student responses observed in these data, the items were further disaggregated by quartile, based on the INIT values of item difficulty and student ability (Figures 6.5 and 6.6). These figures display responses irrespective of whether items were guessed correctly or incorrectly.

For the Year 5 data, the trend lines displayed in Figure 6.5 show that the higher-ability students (most-able quartile and Q3) generally had an increasing propensity to guess as items became more difficult, yet this propensity was less than that shown in the trend lines that represent the Q2 and lower-ability students. However, across the range of item difficulties, the proportion of guesses tended to be hierarchical, with the lower-ability students consistently guessing on more occasions than the higher-ability students. The results of these analyses show that there was an increasing propensity to guess as items became more difficult, and that increased guessing was observed as student ability declined.

Figure 6.6 shows a similar pattern in the Year 7 data to that displayed in Figure 6.5, although the variation in the patterns was more defined, with a clearer delineation between the trend lines of each of the quartiles. The frequency analysis conducted on the Year 7 responses clearly showed the relationship between propensity to guess and item difficulty for all ability groups. These data demonstrate the inverse relationship between the ability of the students and the number of guessed items, as item difficulty increased. These Year 7 data thus reflect the expected patterns and support the assumptions that underpinned the constructs of the simulated data.

Figure 6.5

Year 5 Relationship Between INIT Item Location and the Proportion of Self-Identified Guesses Irrespective of Result

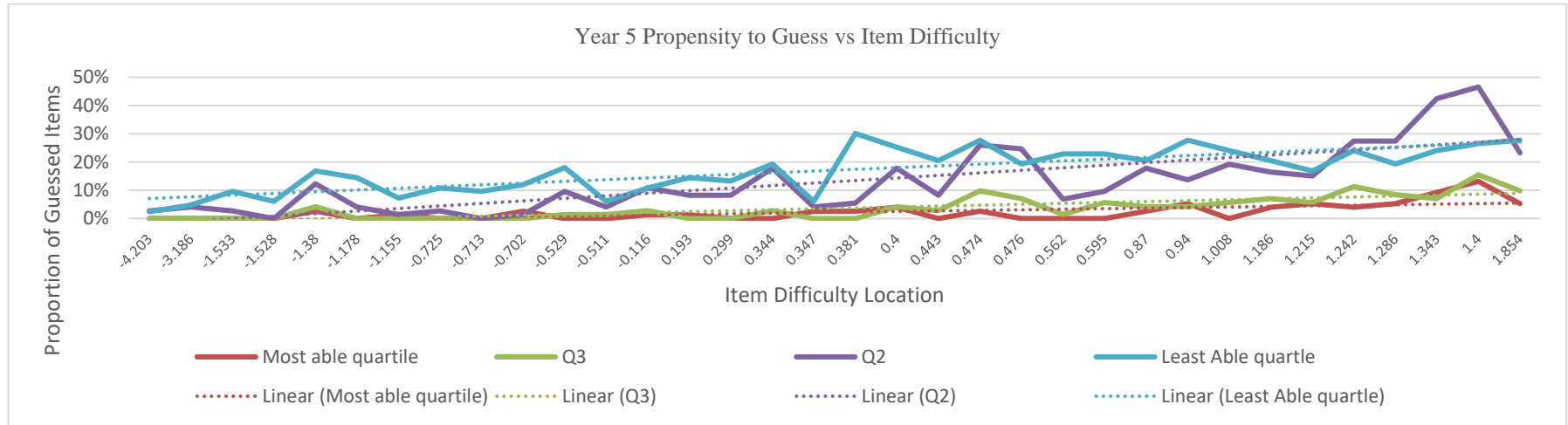
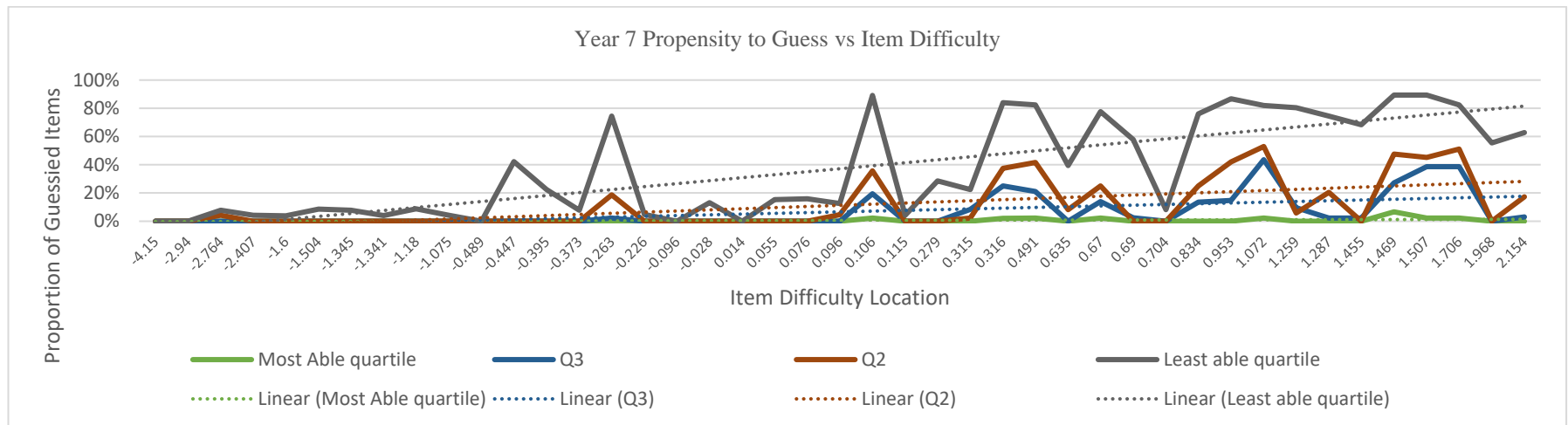


Figure 6.6

Year 7 Relationship Between INIT Item Location and the Proportion of Self-Identified Guesses Irrespective of Result



Note. The percentage of self-identified guessed responses is shown for each ability quartile as the item difficulty locations increase.

6.3.2 Pfa Step 2 and Step 3 Rasch Analysis – English Version Test Data

In a similar manner to the methodology applied to the simulated data in Chapter 5, the following three phases of data analysis using Rasch analyses techniques were conducted:

1. The INIT involved a simple analysis of the response data, taking no account of the self-reported guessing identified by the students. Item parameters and student ability estimates were obtained from these data.
2. The SIG analysis re-coded the student-identified, correctly guessed items as “missing data” in order to recalibrate the item parameters and student ability estimates (analogous to the GS analysis).
3. The GIP, developed from the results of the simulation data analyses (see Chapter 5), revised the student estimates using the GIP by re-coding items indicated as “probable guessing” and re-calibrating the item locations and the consequent raw score to ability estimates. The GIP3A procedure ignored the student-indicated self-identified guesses. The raw score to ability conversion parameters from the “GIPINIT3B conditioned” treatment were then applied to the original response raw score analysed in Phase 1 (the INIT analysis) to determine the ability estimate achieved by each student following the GIPINIT3B analysis.

6.3.2.1 Probable Guessing in the Year 5 Data – Comparison of the SIG Items and GIP Outcomes

The initial investigation focused on the relationship between the student-identified guesses and those indicated by the GIP process based on the INIT analysis. The first observation was the degree to which student success in items was associated with “known” responses (G) and those that were self-identified as a guess (Y and P).

The response pattern for a selection of lower ability Year 5 students is provided in Table 6.1 to illustrate the response pattern of this group of students. This group is displayed because it is the group for which guessing is most likely to occur (see Figure 6.5). For this group, there were many instances of incorrect responses that were indicated by the student as “known” answers (green colouring for Mode and a response score of zero). The frequency of incorrect answers in the responses of the lowest ability group that were self-identified as “a known answer” supported the earlier contention that students, especially the younger age group, may have difficulty in identifying a guess or a known answer (750 of a total of 1850 cases of “known” incorrect answers) or a false sense of their knowledge. Given that the GIP procedure is grounded in an algorithm that considers the difference between the ability of the student and the difficulty of the individual item, the frequency of the incorrect answers that were nominated as “known” responses suggests a high likelihood that these incorrect responses were guessed or a function of incomplete knowledge.

The number of cases in which the GIP parameters indicated a probable guess, for which the student considered he/she knew the answer, were initially considered a challenge to the efficacy of the GIP process. However, observation of the incorrect responses in these tables provides some insight into this apparent anomaly. On multiple occasions, across all ability levels, students answered items incorrectly for which they believed they “knew” the answer. Hence these responses may indicate incomplete knowledge, a misconception by the student, or perhaps a careless mistake. Hence, the coding of student-identified

“known answers” that the GIP procedure indicated as probable guessing may not be an indication of a miscoding, but rather could be construed as evidence of incomplete knowledge of the student for the concept/contexts being assessed. These inconsistencies challenged the intent of this component of the study, in which it was presumed that the self-identified guesses could provide a baseline for evaluating the GIP process; however, the process described in the PfA was completed to investigate the efficacy of the GIP process with live data.

An extract of the Year 5 results for mid-range ability students was reviewed to compare and assess the consistency of the inferences made in relation to the lower ability interactions (Table 6.1). This extract shows that the only items indicated by the GIP procedure to be likely guesses were the harder items (highlighted in pink). A feature of these response patterns was the presence of non-responses as student ability increased, coded in Table 6.2 with the value “9”. This characteristic of the data suggests that the lower-ability students were more likely to guess in a random manner when items were too difficult for their current understanding and/or ability, rather than omit the item, as evidenced in the responses of the higher-ability students. There were no instances of omission in the lower ability quartile of students (approximately 75 students). The increase in the proportion of missing items in the results of the higher-ability students contributed to the decrease in the proportion of student/item interactions indicated as probable guesses by the GIP process.

6.3.2.2 Probable Guessing in the Year 7 Data

Three items have been indicated as probable guesses for student G5E212 (ability estimate 0.086), namely, items Math5Q31, Math5Q35, and Math5Q33, which have difficulty locations that exceed the student ability by about 1.1 logits in this Year 5 scale. In each case, the student identified these responses as “a known answer”. These items, when the set are ordered by difficulty, are in the region that the Guttman scale would expect incorrect answers. By comparison, when student G5E048 (ability estimate 0.214) attempted item Math5Q29 (difficulty location of 1.418), a probability of a correct response was less than 0.25 and the residual greater than 1.75. The GIP indicated this response by as a probable guess. The Guttman-like pattern of responses, generated as a by-product of ordering the items by increasing difficulty, support the contention that item Math5Q29 was beyond the ability of student G5E048. These examples highlight the variability in live data that challenge the GIP process in indicating probable guesses, but in general confirm the consistency of the application of the process.

An extract of the higher-ability students is not presented because in the responses of the students in the top-two ability quartiles the GIP procedure did not indicate any probable guesses, even though there were more than 200 instances of SIG in the upper two quartiles. This highlights a possible challenge encountered by the GIP procedure, when items prove to be too easy for the participating students.

The previous analyses were also performed with the Year 7 data to evaluate consistency with Year 5 patterns. In the Year 7 responses there was an increased occurrence of missing responses (“9”) in these data (Table 6.3). It is also noted that in the majority of cases where a non-response was observed, the probability of a correct response was less than 0.25. The general Guttman-like pattern of correct responses is also notable. This suggests that the older age group (Year 7) of students seemed more likely to not respond to an item that was beyond their knowledge/ability compared to the younger age group (Year 5). There were also several items that were correctly answered and indicated by the GIP procedure as probably guessed, that the student indicated they “knew” the answer (9%). Yet many of their incorrect responses were also reported as “known” answers (15%). This tends to confirm the assertion that students were unable to distinguish between a response derived from a misconception and a guess. An extract of the lower ability Year 7 students (based on INIT analysis) is provided in Table 6.3. This group was selected as it was in this group that guessing was self-indicated as most prevalent and it also contributed the most number of GIP-indicated interactions.

This phase of the analysis highlights two issues that are relevant to the current study:

1. the higher-ability students had a greater propensity to omit harder items, whereas the lower-ability students were more prone to guessing; and
2. as indicated by the presence of the incorrect “known” responses, it is likely that in the case where students had partial skill/knowledge of the concept being assessed by an item they were more likely to respond than to omit the item.

6.3.3 Pfa Step 3 Investigations of the Field Data

The reporting that follows details the Rasch analyses and outcomes. The tables and figures report on the Rasch analyses of the student responses for the INIT (raw data), SIG (self-identified guesses suppressed), and GIP analyses. Given the similarities in the outcomes, the Year 5 and Year 7 data are reported together within each sub-section.

6.3.3.1 Student Ability Estimate Statistics

The increased proportions of GIP-indicated items as item difficulty increased (see Figure 6.4) affected the distribution of item locations and student ability estimates in the GIP analysis. However, the mean of the student ability estimates tended to be negated in the GIPINIT3B estimations (see Figure 6.7). These results were influenced by the relative ease of these test for the sample students, which highlighted the relationship between item and test targeting to the students when attempting to identify probable guessing in the student responses. Despite this impact on the mid-range student estimates, the GIP process did provide a better discrimination among the higher and lower-ability students, which had been anticipated from the simulation studies.

The results of the GIP3A analyses indicate that student-identified guesses were more accurately identified by the GIP for lower-ability students than for higher-ability students (see Figures 6.6. and 6.7) While this is not an ideal situation, it was an improvement over existing approaches (reflected in INIT analysis) that did not provide even this benefit. That guessing could be indicated among students with the highest proclivity to guess is thus encouraging.

As anticipated from the simulation analyses, the mean ability estimates were reduced in the GIP3A analysis, reflecting the recoding of probable guessing and its impact on the raw scores. However, as expected, the standard deviation (as an indicator of the spread of ability distributions) had increased (Tables 6.6 and 6.7). The differences between the student ability statistics of the INIT and GIPINIT3B analyses, as well as the improvement in the separation indices increased for all analyses compared to the INIT analysis, were small. However, the analyses displayed a similar pattern to those observed in SIM4 data analyses, which represented a test too easy for the sample and confirmed the manner in which the GIP process functions with data of these type. Given the premise of the current research is that the commonly applied Rasch analysis – which takes no account of guessing – does not accurately reflect the ability estimates of students in the higher and lower ability regions, the consistency of these outcomes is encouraging as confirmation of the manner in which the GIP process functioned for these distributions.

Despite the challenge imputed by the less than optimal targeting, the results of the planned analyses are revealing. The results show that the SIG conditioning yielded a higher degree of differentiation in student ability estimates, as indicated by the increase in the standard deviation of the mean statistic. Both the SIG and the GIP analyses provided a higher degree of reliability for the assessment, as indicated by the relative Person Separation statistics. Thus, even with these challenging data, these analyses demonstrate a superior performance of the GIP process relative to current (guessing agnostic) practice. The sections that follow describe the analyses conducted and their results.

The results show that for each test the mean ability estimate of the INIT analysis was well over one logit: 1.735 and 1.421, respectively (Tables 6.6 and 6.7). When tests are well targeted these values would trend towards zero, so that there is an alignment between the mean difficulty of the test and the mean ability of the students. These statistics, together with the graphical representations in Figures 6.7 and 6.8, clearly demonstrate that these tests proved too easy for the participating samples.

Table 6.4

Year 5 Summary Analysis – Student Ability Estimates by Analysis Phase

STUDENTS								
n = 303	Phase 1		Phase 2		Phase 3A (GIP)		Phase 3B (GIPINIT)	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
	INIT	INIT	SIG	SIG	GIP3A p=0.6	GIP3A p=0.6	GIPINIT p=0.6	GIPINIT p=0.6
Mean	1.735	-0.145	1.649	-0.186	1.752	-0.192	1.816	-0.019
SD	1.285	0.642	1.427	0.656	1.453	0.604	1.354	0.645
Skewness		1.405		1.040		1.108		1.432
Kurtosis		5.006		4.210		4.709		5.176
Correlation		-0.052		0.118		0.302		-0.128
Separation Index	0.829		0.855		0.858		0.839	
Max β	4.70		4.87		4.90		4.90	
Min β	-5.43		-5.50		-5.70		-5.70	

These data, when considered with the low range of item difficulties, which, if the outliers are ignored, were of the order of approximately 3.0 logits in each year level (Figures 6.7 and 6.8), highlight a challenge for the GIP to identify a guess when the assessments are relatively easy for the target group. This phenomenon was also observed, albeit to a lesser degree, in the Simulation 4 analysis.

Table 6.5

Year 7 Summary Analysis – Student Ability Estimates by Analysis Phase

STUDENTS								
n = 189	Phase 1		Phase 2		Phase 3.A (GIP)		Phase 3.B (GIPINIT)	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
	INIT	INIT	SIG	SIG	GIP3A p=0.6	GIP3A p=0.6	GIPINIT p=0.6	GIPINIT p=0.6
Mean	1.421	-0.176	1.212	-0.231	1.332	-0.174	1.543	0.197
SD	1.016	0.625	1.193	0.676	1.419	0.406	1.220	0.639
Skewness		1.217		0.883		0.394		1.541
Kurtosis		2.080		1.008		0.535		2.759
Correlation		-0.215		-0.001		0.344		-0.328
Separation Index	0.769		0.815		0.858		0.811	
Max β	4.86		5.03		5.81		5.81	
Min β	-5.48		-5.55		-6.85		-6.85	

6.3.3.2 Item Difficulty Statistics

Summary statistics for items across all three analysis phases are presented in Tables 6.4 and 6.5. The item statistics largely conformed to the same pattern found with SIM 4, a test of 40 items that was too easy for the cohort. In these analyses, the mean item locations were constant (default value of zero); however, the standard deviation increased when the guessed items had been accounted for in the SIG and GIP analyses. The range of item difficulties, including maximum and minimum item locations of the SIG and GIP analyses, tended to increase compared to the INIT locations, as anticipated by the simulation studies (Tables 6.4 and 6.5). This is consistent with the outcomes of the simulation studies and indicates that the procedures, when applied to the English-language field study data, functioned fairly uniformly with the outcomes of the simulation.

Table 6.6

Year 5 Item Summary Statistics by Analysis Phase

ITEMS								
n = 40	Phase 1 - INIT		Phase 2 - SIG		Phase 3A (GIP)		Phase 3B (GIPINIT)	
	Location INIT	Fit Residual INIT	Location SIG	Fit Residual SIG	Location GIP3A p=0.6	Fit Residual GIP3A p=0.6	Location GIPINIT p=0.6	Fit Residual GIPINIT p=0.6
Mean	0	-0.181	0	-0.228	0	-0.322	0	0.366
SD	1.338	1.480	1.417	1.624	1.488	1.614	1.488	1.432
Skewness		0.706		0.612		0.160		0.757
Kurtosis		-0.039		-0.086		-0.170		0.138
Maximum	2.37		2.43		2.89		2.89	
Minimum	-4.23		-4.35		-4.54		-4.54	

Table 6.7

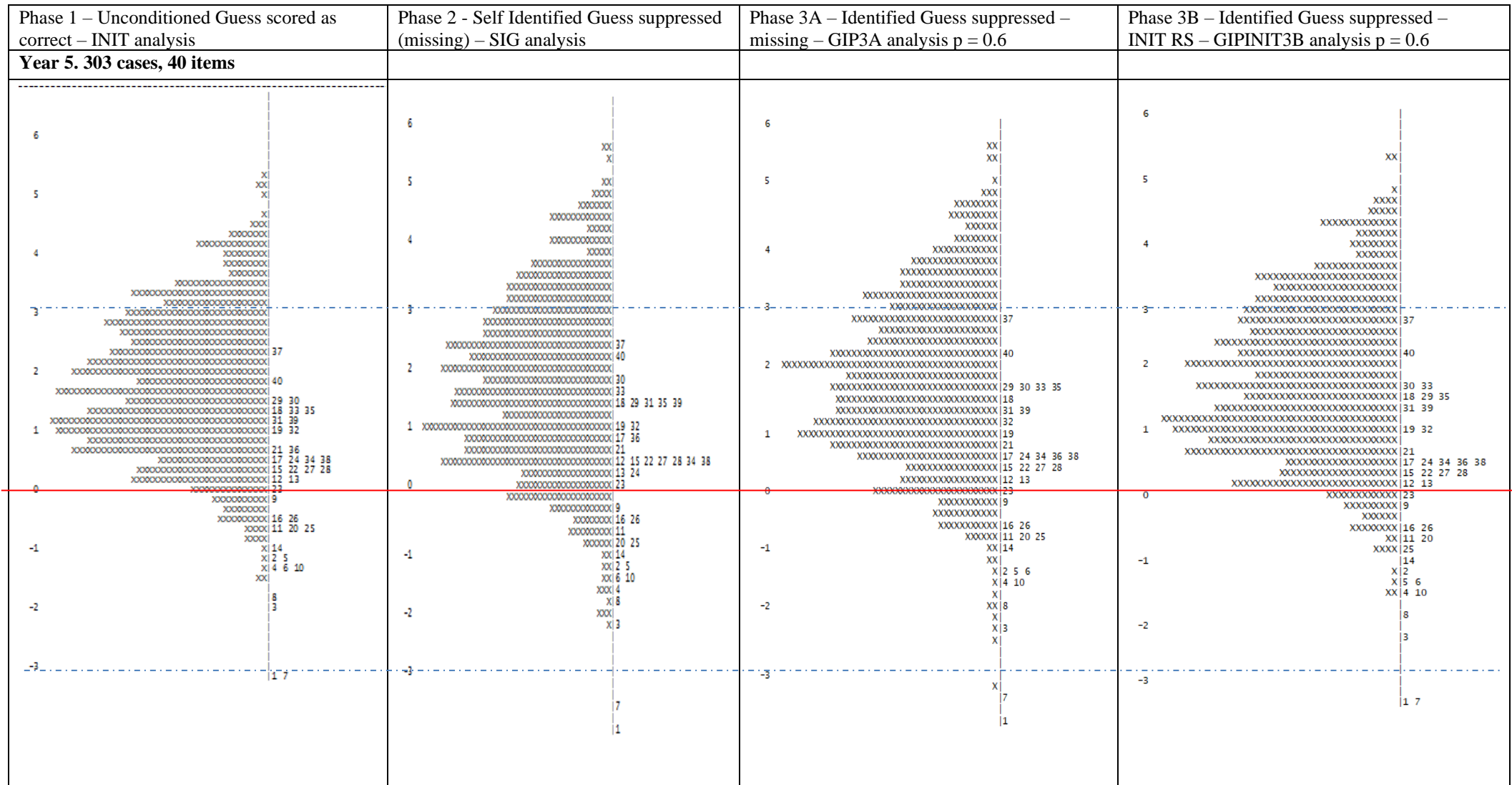
Year 7 Item Summary Statistics by Analysis Phase

ITEMS								
n = 44	Phase 1 - INIT		Phase 2 - SIG		Phase 3A (GIP)		Phase 3B (GIPINIT)	
	Location INIT	Fit Residual INIT	Location SIG	Fit Residual SIG	Location GIP3A p=0.6	Fit Residual GIP3A p=0.6	Location GIPINIT p=0.6	Fit Residual GIPINIT p=0.6
Mean	0	-0.135	0	-0.170	0	-0.135	0	0.678
SD	1.391	1.410	1.504	1.438	1.839	1.047	1.839	1.604
Skewness		1.242		0.553		0.849		1.323
Kurtosis		1.636		0.126		1.042		1.488
Maximum	2.39		2.86		4.44		4.44	
Minimum	-4.15		-3.09		-5.09		-5.09	

The outcomes from each of the Years 5 and 7 analyses are shown in Figures 6.7 and 6.8, respectively. Figure 6.8 also shows the impact of items that were too easy for the cohort, with items 1, 2, 4 and 8 functioning as below the relative ability of the lowest ability students in the Year 7 test. These items contribute little to the scale and highlight the issue of mis-targeting item difficulty to the participating cohort.

Figure 6.7

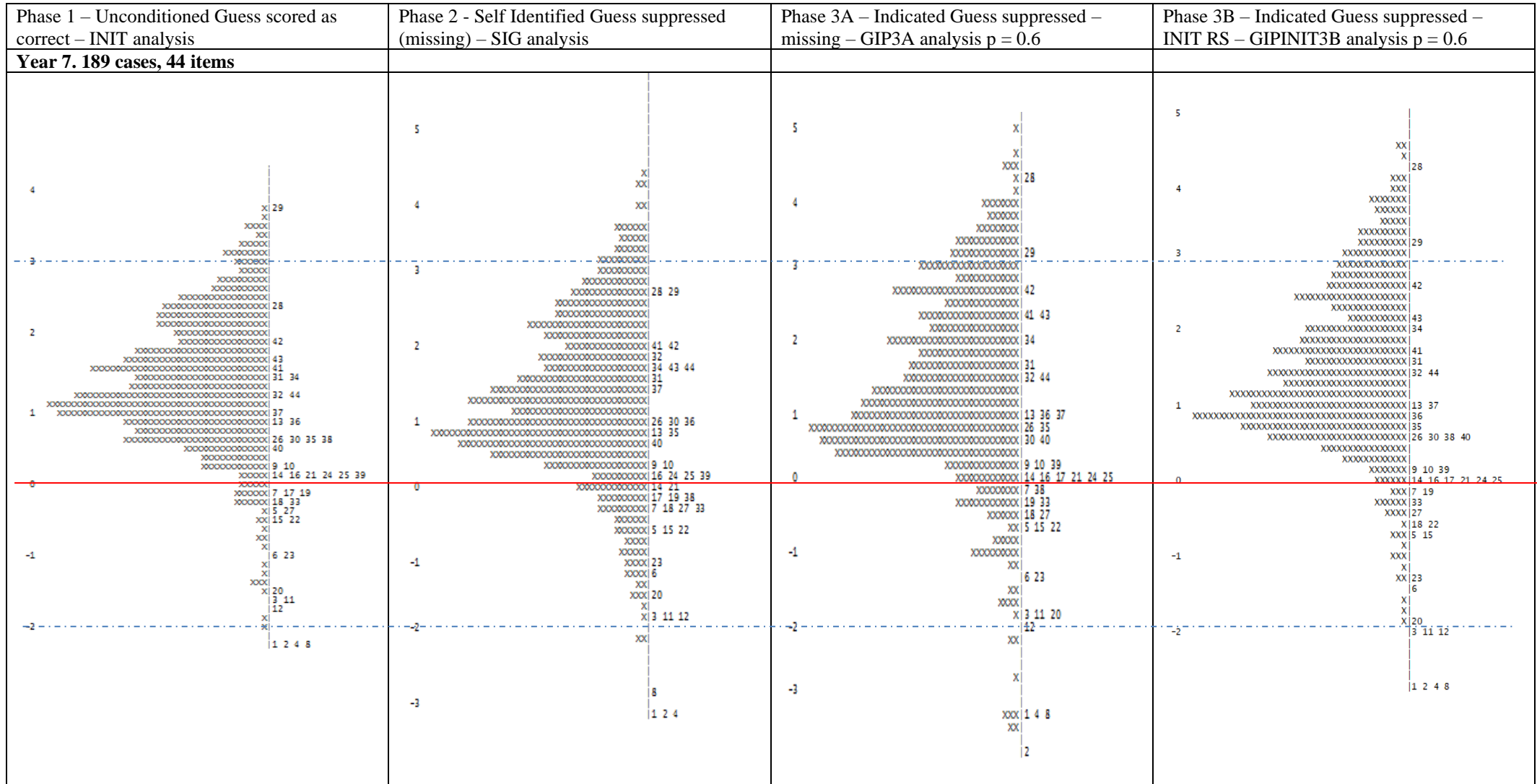
Field Trial Data Item Student Maps – Year 5 English Version Analyses



Note. Figure 6.7 was “aligned” to centre each of the distributions to highlight the shift in the student abilities and the relative differences in the item difficulties for each analysis.

Figure 6.8

Field Trial Data Item-Student Maps – Year 7 English Version Analyses



Note. Figure 6.8 was “aligned” to centre each of the distributions to highlight the shift in the student abilities and the relative differences in the item difficulties for each analysis.

6.3.4 GIP Indication Rates by Quartile

In order to evaluate the efficacy of the $p=0.6$ criterion, the GIP procedure was investigated for both $p=0.5$ and $p=0.6$ to confirm the implications of the analyses of the simulation data. Comparison of the proportion of guesses suppressed by each of the analyses (Table 6.8) shows that the overall mean GIP statistics of the $p=0.5$ recoding were very low (17.6% in Year 5 and 12.7% in Year 7). The process of defining the probability of a correct response at $p=0.6$ had a significant impact on the number of items indicated as a probable guess, with a corresponding increase in the rate of agreement between GIP indicated and the self-indicated guesses. Furthermore, the protocol was comparatively more successful for the lower-ability students, as experienced with the simulated data. For both Year 5 and Year 7 the recovery rate exceeded 70% for the lower-ability group when the $p=0.6$ process was implemented, compared to approximately 30% for the $p=0.5$ analysis (Table 6.8).

An interesting observation in Table 6.8 is the proportion of correct SIG responses for the higher-ability students. In Quartile 4 (most able) of the Year 5 sample, the success rate of the SIG items was above 65%, which would suggest that these were more informed responses than random guesses. Only in the least able quartile of Year 5 SIG statistic did the success rate approximate the expected rate of a random guess for a four-distractor MC item (25%). This confirms the earlier suggestions that in items that proved too easy some students possibly indicated a guess even though the response was supported by knowledge.

Table 6.8

Proportion of Responses Indicated as a Probable Guess by the GIP3A Procedure by Ability Group

Cohort	Group	Proportion of SIG items correct	Count of SIG items	Count Correct SIG items	Count of Guesses recovered by GIP3A $p = 0.5$	Recovery rate Self-Identified GIP3A 0.5 (SIG vs GIP)	Count of Guesses recovered by GIP3A $p = 0.6$	Recovery rate Self-Identified GIP3A 0.6 (SIG vs GIP)
Year 5 Mathematics, 303 cases	Q4, most able	65.2%	69	45	0	0.0%	0	0.0%
	Q3, able	37.9%	116	44	0	0.0%	0	0.0%
	Q2, less able	37.3%	418	156	8	5.1%	22	14.1%
	Q1, least able	25.0%	956	239	77	32.2%	171	71.5%
	Overall	31.0%	1559	484	85	17.6%	193	39.9%
Year 7 Mathematics, 189 cases	Q4, most able	67.3%	104	70	0	0.0%	0	0.0%
	Q3, able	44.5%	245	109	2	1.8%	29	26.6%
	Q2, less able	43.6%	431	188	21	11.2%	62	33.0%
	Q1, least able	42.1%	361	152	43	28.3%	131	86.2%
	Overall	45.5%	1141	519	66	12.7%	222	42.8%

6.3.5 Analysis of Fit

Following the methodology of Andrich et al. (2015), an analysis of the mean square statistic (see Eqn 5.6) was conducted using the outcomes from the Rasch INIT, SIG, and GIP _{$p=0.6$} 3A and GIPINIT3B analyses. The mean square statistic of each analysis was very similar (Table 6.9). In relation to the commentary by Andrich et al. (2015), this indicates that the SIG and each of the GIP analyses displayed an effectively unchanged fit of the data to the model.

These statistics reflect the similarities in the analysis outcomes, with the differences at the extremes compensated by constant estimates about the central tendencies and the minimal impact of the data conditioning due to the relatively few instances of GIP-indicated items. The implication of this outcome is that the GIP process was challenged in improving the scale when there was significant mis-targeting of the test items to the participating cohort to the degree observed in these data.

Table 6.9

Comparison of Mean Square Statistics – INIT and GIP Analyses for Year 5 and Year 7

Analysis	INIT			SIG			GIP3A $p = 0.6$			GIPINIT3B $p = 0.6$		
	Total Chi-Square	d.f.	Mean Square	Total Chi-Square	d.f.	Mean Square	Total Chi-Square	d.f.	Mean Square	Total Chi-Square	d.f.	Mean Square
Year 5 English Mathematics	355	160	2.2	415	160	2.6	414	160	2.6	421	160	2.6
Year 7 English Mathematics	231	88	2.6	247	88	2.8	236	88	2.7	740	88	8.4

6.4 Guttman Analysis of GIP-Indicated Items

Given the assumption of this research study that there exists an inverse relationship between item difficulty and student ability in regard to the propensity to guess, an analysis of the GIP-indicated items was conducted to confirm this relationship with these data. The data support this relationship (Table 6.10); that is, as ability decreased the number and percentage of GIP-indicated guesses increased, although these proportions were very low due to the relative low difficulty of the test item compared to the ability of the students.

Table 6.10

Proportions of GIP-Identified Items by Year and Ability Quartile

Group	Year 5		Year 7	
	Count	Percent	Count	Percent
Quartile 4	0	0.0%	0	0.0%
Quartile 3	0	0.0%	31	1.5%
Quartile 2	22	0.7%	52	2.6%
Quartile 1	123	4.0%	134	6.6%
Total	145	1.2%	217	2.6%

Another assumption of this research study is that the pattern of GIP-indicated interactions should be Guttman-like, as observed in the simulation data. The results of the analysis of the outcomes of the GIP-indicated probable guesses show that, although there were anomalies in some cells of the matrix, overall the pattern was Guttman-like, which largely confirms the expected outcomes of the GIP procedure. The distributions shown in Tables 6.11 and 6.12 confirm the expectations of the study that, as difficulty increased, the lower-ability students were likely to guess more often.

Table 6.12 shows similar patterns to those observed for Year 5 and displays the pattern expected of GIP-identified items in accord with the hypotheses; that is, as student ability decreased, the number of GIP-indicated items tended to increase, although the overall proportions were low. This was interpreted as a function of the tests being too easy for the students and the relatively small distribution of item locations for most items relative to the student ability estimates. As shown in Table 6.12 the high proportion of omitted responses indicated by the percentage of non-attempts had an adverse impact of the number of GIP identified responses.

6.5 Discussion – Implications of the Outcomes of the English Tests

Figures 6.7 and 6.8 demonstrate the mis-targeting of the English versions of the tests for the sample students, with the tests proving to be too easy. This characteristic of the tests impacted on the degree to which guessing was present, with the difference between the mean ability estimates being more than 1.4 logits above the mean item locations in the case of both Year 5 and Year 7 (see Tables 6.6 and 6.7). This degree of mis-targeting reduced the amount of guessing in the tests overall and thus constrained the capacity of the GIP procedure to indicate probable guessed items.

6.5.1 The SIG Outcomes

The initial observations that there were fewer occurrences of self-identified guessing for easier items and an increase in the amount of guessing in more difficult items (see Figures 6.1 and 6.3) support the assumptions that underpin the study and thereby provide at least partial support for use of the GIP procedure. For many of the items in Year 5 test, the item facility was greater than 80%, with the proportion of SIG items less than 10% (Figure 6.1). In the simulated data, typically about 90% of the explained variation in the comparable statistic (GS) was a function of the overall facility of the item. In the Year 5 data, the variation explained by the SIG was closer to 28%. In the Year 7 data the explained variation in the SIG was approximately 13%. Given the amount of incorrect, “known” answers, there was inconsistency in the SIG data that militated against the use of the data as a baseline for comparison with the GIP procedures.

It was noted that students in both Years who randomly guessed items (indicated by the “P” for pink marker) in general had probability estimates less than 0.25 for each indicated item and an item-student residual greater than 1.75 (see Tables 6.1 to 6.3). However, given the inconsistencies observed, it was difficult to confidently comment on the source of the variability in the SIG data. In the simulated data, the incidence of guessing was defined. In these “real-life data” the self-identified guessing was far less “determined”. Considering the intended methodology that involved the students’ self-identifying guesses (SIG), the results proved inconclusive and consequently were sub-optimal to evaluate the GIP process.

Table 6.11

Year 5 Number of GIP-Identified Items by Item Location

δ	-4.23	-3.12	-1.97	-1.50	-1.46	-1.32	-1.13	-1.11	-1.07	-0.95	-0.66	-0.66	-0.65	-0.47	-0.45	-0.05	0.08	0.20	0.26	0.36	0.40	0.40	0.40	0.45	0.50	0.53	0.53	0.62	0.66	0.93	1.00	1.08	1.26	1.28	1.32	1.36	1.42	1.47	1.94	2.37		
item	Q01	Q07	Q03	Q10	Q08	Q04	Q05	Q02	Q06	Q14	Q11	Q25	Q20	Q26	Q16	Q09	Q23	Q12	Q13	Q22	Q15	Q27	Q28	Q24	Q34	Q38	Q17	Q36	Q21	Q19	Q32	Q39	Q31	Q35	Q33	Q18	Q29	Q30	Q40	Q37		
Q4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Q3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	9
Q1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3	11	4	1	3	2	0	1	0	0	0	2	9	11	7	4	20	2	16	0	4	1	13	8		
Total	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3	11	4	1	3	2	0	1	0	0	2	9	11	7	4	20	2	16	0	17	1	13	17			

Table 6.12

Year 7 Number of GIP-Identified Items by Item Location

δ	-4.15	-2.94	-2.76	-2.41	-1.60	-1.50	-1.35	-1.34	-1.18	-1.08	-0.49	-0.45	-0.40	-0.37	-0.26	-0.23	-0.10	-0.03	0.01	0.06	0.08	0.10	0.11	0.12	0.28	0.32	0.32	0.49	0.64	0.67	0.69	0.70	0.83	0.95	1.07	1.26	1.29	1.46	1.47	1.51	1.71	1.97	2.15	2.39	
item	Q02	Q08	Q04	Q01	Q11	Q12	Q20	Q03	Q06	Q23	Q15	Q05	Q27	Q22	Q18	Q33	Q07	Q17	Q19	Q14	Q24	Q21	Q16	Q39	Q10	Q25	Q09	Q40	Q38	Q26	Q35	Q30	Q13	Q36	Q37	Q44	Q32	Q34	Q31	Q41	Q43	Q42	Q28	Q29	
Q4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	3	4	10
Q2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	0	10	0	0	0	0	0	0	6	0	5	8	11	
Q1	0	0	0	0	2	0	9	0	2	3	0	0	0	0	3	1	0	1	4	2	0	0	5	5	12	7	16	5	2	4	3	7	2	6	3	2	0	1	3	6	5	3	0	15	
Total	0	0	0	0	2	0	9	0	2	3	0	0	0	0	3	1	0	1	4	2	0	0	5	5	12	7	16	5	2	4	7	10	2	16	3	2	0	1	3	26	5	11	12	36	
Non_Att	6	7	4	0	0	2	2	0	13	2	2	2	4	4	2	8	11	4	5	14	7	19	0	6	0	20	4	77	69	27	29	39	44	45	55	55	58	66	65	33	76	83	79	25	
percent	3%	4%	2%	0%	0%	1%	1%	0%	7%	1%	1%	1%	2%	2%	1%	4%	6%	2%	3%	7%	4%	10%	0%	3%	0%	11%	2%	41%	37%	14%	15%	21%	23%	24%	29%	29%	31%	35%	34%	17%	40%	44%	42%	13%	

6.5.2 *The GIP outcomes*

The results for Simulation 4, a test too easy for the cohort, suggest that the GIP procedure was more challenged in identifying guessed responses for higher-ability students. This happened because the general over-estimation of all students prior to the conditioning of the data with the GIP process resulted in generally inflated ability estimates. The inflated ability estimates reduced the difference between student estimates and item difficulties, with a consequent reduction in the ability of the GIP process to identify guessing, especially in the higher-ability students.

The GIP process results were similarly influenced by the relative ease of these tests for the sample students, thus highlighting the relationship between item difficulty and test targeting when the students were attempting to identify probable guessing in the student responses. In these data, the mean ability of the participating students was significantly above the mean location of the items (1.7 and 1.4 respectively). Given that the critical difference to have a probability of a successful response of less than 25% (0.25) is a difference of 1.1 logits, the simple numeric capacity of GIP to operate effectively was severely reduced.

Despite that constraint, in both Years 5 and 7 there were marginal increases in the distribution of item locations between the INIT and the GIP3A analyses (see Table 6.5 and Figure 6.8), with the relative rank order of items remaining constant across all analyses. This is a consequence of how the GIP procedure identifies more instances of probable guessing as item difficulty increases. Hence the location order remains constant but the difference between locations increases. The relatively small number of GIP-indicated re-coded items resulted in little change in item difficulty and consequent ability estimates in the GIP results, when compared to the INIT outcomes.

In the results of the GIP3A analyses for both Years, the distance between the ability estimates increased in the more difficult item ranges, which indicates that the GIP was accounting for guessing in an increasing proportion. This is shown in Figure 6.7, where the range of each distribution was reasonably similar, although the GIP_{p=0.6} analysis had a greater range that was influenced predominately in the lower ability regions. This reflects the GIP procedure impacting the lower-ability students more effectively, with less impact on the higher-ability students.

The Separation Index of the GIPINIT3B improved compared to the INIT analysis, which reflects the outcomes observed in the simulation studies. Given that the lower-ability groups were more likely to guess (Figures 6.7 and 6.8), the results are encouraging as they indicate the GIP_{p=0.6} procedure had a positive impact on the accuracy of the estimates of the lower-ability group. This outcome is not provided by Andrich et al.'s (2015) methodology.

As much as these outcomes were sub-optimal in terms of the effectiveness of the GIP procedure, they don't detract from the outcome that the GIP procedure produced an improvement in the estimates of item difficulties and in the reliability of the student ability estimates. These results generally confirm the outcomes of the relevant simulation study and provide preliminary evidence that an improved measurement was obtained by procedures that account for guessing in student response patterns, despite the fact that the tests were too easy for the cohort.

6.6 Study 2, Stage 2, Arabic Versions of the Tests

6.6.1 Introduction: The Arabic Versions Study

The intent of having students complete an Arabic version of the same Mathematics achievement test was to gain insight into students' guessing patterns. The Arabic version was administered to the students prior to the administration of the English version to ensure responses could not be informed by recall of the English versions of the items. The tests were designed to provide commonality in the items that were not language dependent to enable analysis of the correlation between student ability and propensity to guess with the language dependent items. For instance, there were items with clues that allowed the higher ability mathematical students to engage with the items using strategies beyond pure guessing. There were also items that were simple algorithms or English format equivalents, which were language independent, given that the presentation was identical and in a format familiar to students in each cohort. Yet each of the Arabic tests also included some items that were language dependent. Given that Arabic was not the first language of the students, it was anticipated that administration of this Arabic assessment would compel at least some random guessing by all students.

It was noted that for the SIG and GIP phases for the Arabic versions of the tests there were some items for which no student attained the "correct answer" without guessing. In a Rasch analysis methodology these items would be removed from the data, which would impact and limit the comparability of results as the ability estimates are based on a lower potential raw score.

6.7 PfA of the Arabic Versions

The PfA of the Arabic versions of the tests followed the same basic plan as for the English versions of the test.

6.7.1 Comparisons of English Version and Arabic Version of Selected Items

Initial comparisons were undertaken to provide evidence that the participating students had made genuine attempts to respond to the English and Arabic versions of the tests, and thereby legitimate the planned analyses that would follow. A sample of items has been included (Figures 6.9 to 6.12) in this section to show the relative differences between items that were language dependent compared to those that were language independent. The statistics for language-independent items suggest that the students did not generally randomly guess the answer when there were clues to aid familiarity. However, the differences in the facility rates of the English and Arabic versions of these items indicated that random guessing was in fact occurring. The facility rates for all items are presented in Appendix C.

6.7.1.1 Year 5 Items

The items displayed in Figure 6.9 were selected as samples of language-dependent and language-independent items to show the relative facility rates (percent correct) of the cohorts in the Arabic versions of the test session (which were completed first in all cases) and the subsequent English versions of the test. The purpose of showing these items is to demonstrate the relationship between the capacity of the students to attempt the Arabic items and to gauge the impact of language in their ability to interpret the content and correctly answer the English version of the item. The results suggest that the students did make a genuine attempt in the tests, thus providing some insight into the observations regarding tests that proved “too difficult” for a cohort.

As an example of a language-independent item, Math5Q03 had a context that would be familiar to students, in that recognisable digits and equations were presented. In support of this item as language independent, the language demand of the Arabic version did not impede the students’ capacity to respond correctly to the item (96% and 96% for English and Arabic forms, respectively, in the INIT data). Response patterns showed that in this and in similar items that were not fully language dependent, the facility rate was due mainly to “known responses” or “informed guess” responses. This supports the conclusion that students made genuine attempts at each of the items encountered in the Arabic version of the test.

Figure 6.9

Year 5 Math5Q03

Year 5 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
3	Math5Q03	96%	96%	96%	96%	97%	96%

ما هو الرقم الناقص؟ 3

$$3 \times 8 = \boxed{?}$$

(A) 11 (B) 16 (C) 24 (D) 32

By comparison, Figure 6.10 shows an item (Math5Q08) in which the responses were significantly impacted by the language barrier. There was a pronounced difference between English and Arabic facility rates; the English version had a facility rate of 95%, while the Arabic version had a 30% success rate (which is about the expected value of a random guess in the INIT and GIP analyses). When the self-identified guesses were excluded in the SIG analysis, only 2% of the students answered the item correctly. Six students indicated that they knew the answer. All six had some Arabic language knowledge. Although this knowledge afforded an advantage to these students, it was considered inconsequential to the analysis overall due to the proportion of the sample they represented. These outcomes also provide confidence that the analyses conducted on these data were grounded in genuine attempts to complete the test, not simply spurious random data.

Figure 6.10

Year 5 Math5Q08

Year 5 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
8	Math5Q08	95%	95%	95%	30%	2%	31%

<p>8 The chart shows the number of visitors to a Sports Centre during four months.</p> <p>Which month had the most visitors?</p> <table border="1"> <thead> <tr> <th>Month</th> <th>Number of visitors</th> </tr> </thead> <tbody> <tr> <td>January</td> <td>6055</td> </tr> <tr> <td>February</td> <td>6505</td> </tr> <tr> <td>March</td> <td>6500</td> </tr> <tr> <td>April</td> <td>6550</td> </tr> </tbody> </table> <p> <input type="radio"/> January <input type="radio"/> February <input type="radio"/> March <input type="radio"/> April </p>	Month	Number of visitors	January	6055	February	6505	March	6500	April	6550	<p>8 بيّن الجدول أدناه عدد الزائرين إلى مركز رياضي في خلال أربعة أشهر.</p> <p>في أي شهر جاء العدد الأكبر من الزائرين؟</p> <table border="1"> <thead> <tr> <th>عدد الزائرين</th> <th>الشهر</th> </tr> </thead> <tbody> <tr> <td>6055</td> <td>يناير</td> </tr> <tr> <td>6505</td> <td>فبراير</td> </tr> <tr> <td>6500</td> <td>مارس</td> </tr> <tr> <td>6550</td> <td>أبريل</td> </tr> </tbody> </table> <p> <input type="radio"/> يناير <input type="radio"/> فبراير <input type="radio"/> مارس <input type="radio"/> أبريل </p>	عدد الزائرين	الشهر	6055	يناير	6505	فبراير	6500	مارس	6550	أبريل
Month	Number of visitors																				
January	6055																				
February	6505																				
March	6500																				
April	6550																				
عدد الزائرين	الشهر																				
6055	يناير																				
6505	فبراير																				
6500	مارس																				
6550	أبريل																				

6.7.1.2 Year 7 Items

As observed in the Year 5 examples, the facility rates and response modes support the contention that the Year 7 students made serious attempts at answering the items during both versions of the tests. As with the Year 5 selection, a few items have been provided to demonstrate various response patterns by the cohort to items of varying language dependency (Figures 6.11 and 6.12).

In the English version of item Math7Q12 (Figure 6.11) only two students omitted the question, with 94% of the remaining 187 students responding correctly. In the Arabic version 23% “guessed” correctly (the approximately expected chance). Only four students omitted the item in the Arabic version, which may be an indicator of the propensity to guess by students in an assessment regime with no downside for an incorrect guess. No student answered this item correctly without a self-indicated guess in the Arabic version of the SIG analysis. Consequently, the item was removed from the SIG analysis as a result of the re-coding of all the guessed answers to have “missing” values. In the Rasch analysis model, an item with no correct responses is “extreme” and omitted from the analysis. Hence there are no statistics for this item in the SIG analysis.

The data displayed in Figures 5.11 and 5.12 initially appear appropriate for evaluating the GIP, given they appear to be genuine attempts and authentic indications of guessing behaviours.

Figure 6.11

Math7Q12

Year 7 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
12	Math7Q12	94%	94%	94%	23%	N/A	6%



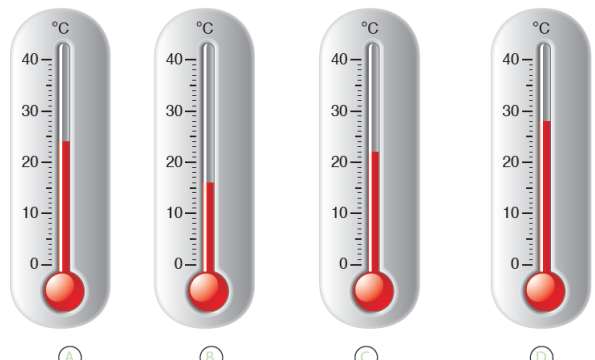
<p>12 Latifa has these shapes cut out of card.</p>  <p>She uses them to make a three-dimensional model. What is the name of the model she makes?</p> <p>(A) cone (B) cylinder (C) rectangular prism (D) sphere</p>	<p>12 لدى لطيفة الأشكال التالية المقطوعة من بطاقة.</p>  <p>استخدمت هذه الأشكال لتحصّل على نموذج ثلاثي الأبعاد. ما هو اسم النموذج الذي حصلت عليه؟</p> <p>(A) مخروط (B) اسطوانة (C) منشور رباعي (D) كرة</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 6.12 shows the relatively minor impact of language in item (Math7Q01) that is common in its presentation, and where the task demands of the item are conveyed independently of language. The minor increase in the facility rate observed by the SIG analysis reflects the reduction in the number of students attempting the item when self-identified guesses were suppressed in the SIG analysis.

Figure 6.12

Math7Q01– Arabic Version

Year 7 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
1	Math7Q01	98%	98%	98%	91%	89%	91%

<p>1 يخطط حمدان للذهاب في نزهة يوم غد. تبين التوقعات الجوية أن درجة الحرارة غدا ستكون 24°C. أي مما يلي تبين درجة حرارة 24°C ؟</p>  <p>(A) (B) (C) (D)</p>

6.7.1.3 Pfa Step 1, Guesses Indicated by the GIP Procedure

An analysis of the proportion of GIP-indicated guesses by quartile (similar to Table 6.10) was produced for the Arabic data, based on the INIT analysis (Table 6.13). This table shows relatively consistent percentages of GIP-indicated probable guesses irrespective of the ability of the students, which was inconsistent with expectations and the observed results in the simulations and the English tests. Instead, these results reflect the randomness of the data and show that they are not Guttman-like. This pattern suggests that the outcomes of the Rasch analyses, which assumes a general Guttman relationship, should be considered with caution.

Table 6.13

Year 5 and Year 7 Proportion of GIP identified items by INIT ability quartile

Group	Year 5		Year 7	
	Count	Percent identified	Count	Percent identified
Quartile 4	47	1.5%	53	2.6%
Quartile 3	72	2.4%	40	2.0%
Quartile 2	107	3.5%	50	2.4%
Quartile 1	65	2.1%	146	7.1%
Total	291	2.4%	289	3.5%

Note. The table shows the percentage of items indicated by the GIP procedure. The percentage reported is the ratio of the number of GIP-identified items in the quartile compared to the total number of items for which there was a response in the quartile.

The distributions of GIP-indicated guesses (Table 6.14) show little relationship to the Guttman-like patterns observed in the simulated data and the English versions of the test (see Tables 6.11 and 6.12) and also confirm the random nature of the responses and the randomness of the interaction of item location with student ability estimates.

Table 6.14

Year 5 Distribution of GIP-Identified Items by Quartile and Item Location

<i>s</i>	-3.86	-1.97	-1.83	-1.34	-1.29	-1.00	-0.94	-0.92	-0.90	-0.73	-0.67	-0.49	-0.28	-0.20	-0.18	-0.13	-0.09	-0.01	0.04	0.08	0.11	0.16	0.17	0.18	0.20	0.23	0.23	0.36	0.39	0.43	0.68	0.87	1.17	1.27	1.39	1.66	1.66	1.73	1.94	2.10	
Item	Q03	Q01	Q06	Q24	Q1F	Q1*	Q3*	Q1*	Q1F	Q16	Q29	Q09	Q26	Q11	Q40	Q31	Q1*	Q14	Q38	Q2*	Q33	Q2*	Q10	Q39	Q0*	Q30	Q08	Q28	Q23	Q04	Q18	Q13	Q19	Q3*	Q2*	Q20	Q0*	Q21	Q34	Q16	
Q4	0	0	0	0	0	0	9	0	0	0	0	0	0	0	7	0	0	0	0	0	1	3	4	0	0	0	10	0	0	0	0	0	0	0	0	0	9	0	0	4	0
Q3	0	0	0	15	0	5	5	0	0	1	0	2	0	0	7	0	0	0	0	1	2	1	0	0	11	0	5	0	1	7	0	3	0	0	0	2	0	1	1	1	0
Q2	0	0	0	17	0	0	3	0	0	7	0	7	2	7	4	7	0	0	0	2	0	0	0	3	10	0	6	0	7	5	0	4	4	0	1	3	0	0	2	4	
Q1	0	0	0	5	0	0	3	0	0	5	1	2	2	1	1	2	1	5	0	3	0	0	0	5	3	0	0	2	5	2	0	1	5	0	0	2	0	0	2	5	
Totals	0	0	0	37	0	5	20	0	0	13	1	11	4	8	19	9	1	5	9	6	3	4	4	8	24	9	21	2	13	14	9	8	9	0	1	16	0	1	9	9	

Note. The INIT item locations are sorted from easiest to most difficult and the number of GIP-identified items are noted in each cell of the matrix.

The distribution of GIP-indicated interactions in the Year 7 test (Table 6.15) are also quite random and do not follow the Guttman-like pattern. This outcome indicates similar cause for concern in the Rasch analyses that were conducted with these data.

Table 6.15

Year 7 Distribution of GIP-Identified Items by Quartile and Item Location

θ	-2.67	-2.47	-2.38	-2.31	-1.90	-1.70	-1.33	-1.27	-1.26	-1.22	-1.10	-0.87	-0.58	-0.37	-0.25	-0.08	-0.07	0.01	0.02	0.05	0.08	0.18	0.22	0.33	0.34	0.41	0.44	0.45	0.57	0.62	0.70	0.80	0.82	0.98	0.98	0.99	1.16	1.20	1.28	1.48	1.54	1.72	2.01	2.47
Item	Q01	Q02	Q08	Q04	Q11	Q05	Q25	Q03	Q27	Q23	Q20	Q24	Q10	Q30	Q18	Q36	Q16	Q44	Q37	Q32	Q13	Q19	Q06	Q34	Q14	Q40	Q07	Q09	Q41	Q38	Q29	Q43	Q17	Q12	Q28	Q26	Q39	Q15	Q22	Q31	Q33	Q21	Q35	Q42
Q4	0	0	0	0	0	0	0	0	0	0	0	6	0	0	6	0	1	0	0	0	3	6	0	0	0	3	0	2	0	0	6	0	9	0	2	0	0	0	7	0	0	2	0	0
Q3	0	0	0	0	0	0	2	0	5	0	0	5	0	0	4	0	12	0	0	0	6	3	0	0	0	14	0	12	7	0	11	0	5	0	3	0	0	1	8	0	9	2	17	0
Q2	0	0	0	0	0	12	18	0	18	0	0	14	5	13	12	0	10	0	10	0	5	8	0	0	0	9	0	9	17	0	5	8	5	9	2	0	0	8	11	14	16	6	6	0
Q1	0	0	1	1	1	16	8	1	15	6	1	7	16	15	8	15	10	3	11	2	4	9	1	3	1	6	0	9	6	2	6	13	6	12	7	12	11	12	6	11	10	1	11	19
Total	0	0	1	1	1	28	28	1	38	6	1	32	21	28	30	15	33	3	21	2	18	26	1	3	1	32	0	32	30	2	28	21	25	21	14	12	11	21	32	25	35	11	34	19

Note. The INIT item locations are sorted from easiest to most difficult and the number of GIP-identified items are noted in each cell of the matrix.

Despite these concerns, a Rasch analysis of the data and supplementary analyses were conducted for completeness and to determine if further inferences or observations relative to the aims of the study could be obtained from these data.

6.8 PfA Step 2, Rasch Analyses of the Arabic Response Data

6.8.1 Year 5 Item Statistics

The results of the Rasch analyses indicate that there was a significant shift in the standard deviation of the item locations, with the SIG analysis having more than twice the spread of the INIT analysis (Table 6.16). This was expected, as it was associated with the suppression of the items that were language dependent and/or influenced by the language component of items that contained clues for the mathematically more able students. The INIT analysis has item statistics that are close to a normal distribution ($SD = 1.17$). By comparison, the SIG analysis shows a considerable positive skew, and the distribution of the item locations is non-normal, as indicated by the large kurtosis statistic as a result of the suppression of the student/item interactions. The shift in these statistics highlights the degree to which the SIG data are fundamentally different from the INIT data, which is a cause for caution in drawing inferences from the SIG data analyses.

Table 6.16

Year 5 Arabic INIT, SIG, GIP, and GIPINIT Summary of Analyses – Items

n = 40	Phase 1		Phase 2		Phase 3A		Phase 3B	
	Location INIT	Fit Residual INIT	Location SIG	Fit Residual SIG	Location GIP3A $p = 0.6$	Fit Residual GIP $p = 0.6$	Location GIPINIT $3B_p = 0.6$	Fit Residual GIPINIT $p = 0.6$
Mean	0	0.042	0	-0.383	0	0.042	0	0.096
SD	1.173	1.414	2.657	1.129	1.206	1.555	1.206	1.468
Skewness		-0.302		2.814		-0.330		-0.384
Kurtosis		-0.469		11.471		-0.517		-0.518
Correlation		0.415		-0.154		0.407		0.384

6.8.2 Year 5 Student Ability Statistics

The INIT analysis indicates a homogenous group with the mean ability estimate of -0.628 and a standard deviation of 0.651, which is a small distribution (Table 6.17). The Separation Index (the index of score correlations) of 0.663 indicates a test that did not discriminate between the students, and consequently suggests a test of low reliability. The SIG analysis in Table 6.17 shows a decrease in the mean ability estimate (as a result of the decreased raw score) of more than two logits, and a doubling of the spread of results when comparing the standard deviations. Although this shift in the SIG outcomes was expected, due to the impediments of the language barrier, the degree to which the observed change in the variable was actuated means that these results have not contributed to a meaningful interpretation of the outcomes.

The GIP3A and GIPINIT3B results are very similar to the INIT statistics. This may be a function of the extreme randomness of the data, due to language barriers mitigating the impact of ability, which influenced the range of item locations and student ability estimates in these analyses. In these data the homogeneity of ability estimates resulted in relatively few student/item interactions indicated as probable guessing (see Table 6.15). Consequently, these data provided little information for an evaluation of the GIP procedure.

Table 6.17

Year 5 Arabic INIT, SIG, GIP, and GIPINIT Summary of Analyses – Students

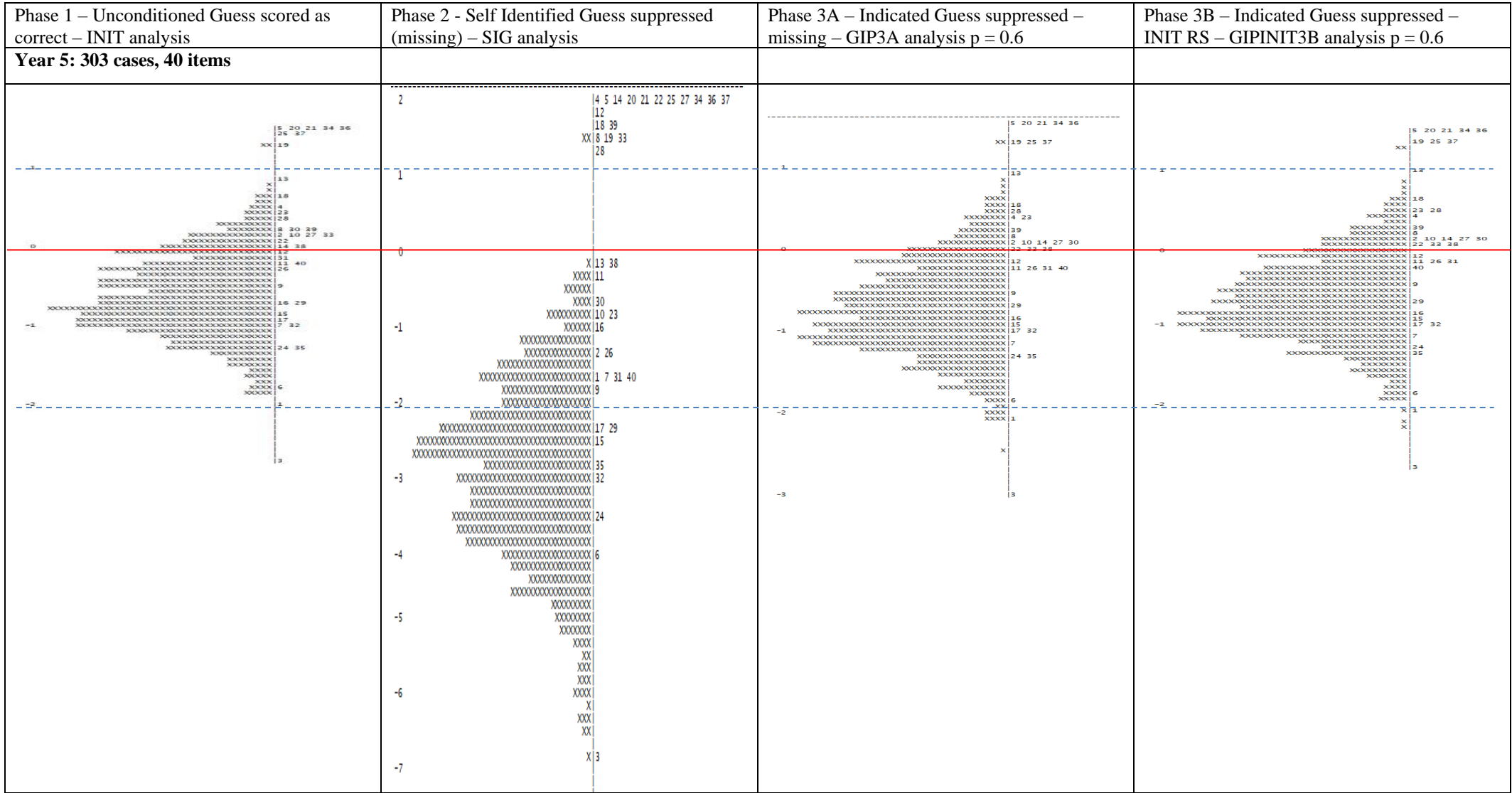
n = 303	Phase 1		Phase 2		Phase 3A		Phase 3B	
	Mean Ability estimate INIT	Fit Residual INIT	Mean Ability estimate SIG	Fit Residual SIG	Location GIP3A p = 0.6	Fit Residual GIP p = 0.6	Location GIPINIT 3B p = 0.6	Fit Residual GIPINIT p = 0.6
Mean	-0.628	-0.121	-2.975	-0.235	-0.684	-0.101	-0.628	-0.086
SD	0.651	0.963	1.396	0.313	0.685	0.934	0.655	0.961
Skewness		0.607		2.210		0.538		0.666
Kurtosis		0.266		13.237		0.084		0.421
Correlation		-0.307		-0.158		-0.326		-0.303
Separation Index	0.663		0.719		0.675		0.665	

The SIG item/student map (Figure 6.13) displays a larger number of items with locations considerably higher than ability levels compared to the INIT analysis. The SIG figure also shows a long negative “tail” for the students, with the majority achieving ability estimates below -2.0 logits, which was the lower extreme of the INIT analysis.

The range of item locations in the INIT analysis is effectively -1.5 to +1.0 (Figure 6.13) if the very difficult items (top right of Figure 6.13) from the INIT analysis are ignored. This means there was little discrimination between the 32 items in the effective range, with a cluster of items located between 0 and +1.0. By comparison, student ability estimates ranged between -2.0 and +1.0, with the vast majority falling between 0 and -1.0. Again, this means there was little discrimination between relative abilities of the students generated by the interaction of the items with the student responses.

Figure 6.13

Item/Student Maps for Year 5 Arabic INIT and SIG Analyses



This result indicates that the data did not fit the RM, and consequently could not contribute to the study aims apart from highlighting the issues in relation to targeting of test items to the ability of the group and the challenges when there is considerable mis-targeting of the test to the cohort. Consequently, these results regarding the analysis of the Year 5 data are provided for reference and to document the completion of the Plan for Analysis, rather than to provide insights and comparisons that would have significant caveats.

6.8.2.1 Year 7 Item Statistics

The removal of many items in the SIG and GIP analyses, due to facility rates of zero, obfuscated comparison of the results (Table 6.18). Again, it is apparent that the test was too “hard” for the participants, and this impaired the capacity of the GIP process to indicate probable guessing. The high degree of guessing in the responses to the Arabic items generated less reliable indicators of ability in the trait of interest. This is supported by the low reliability indices shown in Table 6.19.

The item statistics in Table 6.18 display a pattern of outcomes similar to those recovered for the simulation data for a test that was too hard for the cohort (SIM5). However, these data represent an extreme case of a test which is too hard. It is significant that eight items were removed from the SIG analysis due to the students universally nominating these items as “guessed”. These eight items were language dependent. Four items were removed from the GIP analysis, all of which were also language dependent. Relative to the INIT analysis, the standard deviation of the items in the GIP_{p=0.6} analysis increased, although this was influenced by the removal of items in the GIP3A and GIPINIT3B phases.

Table 6.18

Year 7 Arabic Summary Analysis for Items by Analysis Phase

n = 184	ITEMS							
	Phase 1		Phase 2		Phase 3A		Phase 3B	
			36 items – 8 removed		40 items converged – 4 removed			
	Location INIT	Fit Residual INIT	Location n SIG	Fit Residual SIG	Location GIP3A p = 0.6	Fit Residual GIP3A p = 0.6	Location GIPINIT 3B p = 0.6	Fit Residual GIPINIT 3B p = 0.6
Mean	0	0.017	0	-0.603	0	-0.450	0	1.948
SD	1.258	1.545	2.055	1.218	2.084	1.153	2.084	1.790
Skewness		-0.086		0.881		0.501		-0.975
Kurtosis		0.011		1.642		0.578		1.557
Correlation		0.372		0.009		-0.403		0.359

6.8.2.2 Year 7 Student Ability Statistics

In considering the interaction between the item difficulties and the student ability estimates, it is noted that the ability estimates fell within the range +1 to -2 logits, with the mean of -0.628 logits in the INIT analysis. This lack of differentiation restricted the capacity of the GIP process to function at an optimal level. The INIT and GIPINIT3B analyses identified several items that had difficulty locations in excess of the general calculated ability of the participating student sample; 18 items in the case of the SIG analysis and nine in the INIT analysis (see Figure 6.14). These results confirm that the Arabic language versions of the test were “too difficult” for the cohort and, as per the Year 5 data, detract from the capacity to compare outcomes (see Figure 6.13). Specifically, the relatively small distribution of item locations and student ability estimates in the INIT analysis resulted in a low proportion of student/item interactions indicated as probable guesses. Despite the reduced number of re-coded items, there was a reduction in the mean ability estimate and an increase in the standard deviation compared to the INIT values as predicted by the simulations. However, these comparisons are made with caution as the GIP outcomes are based on 40 items compared to the INIT analysis of 44 items.

The re-introduction of the INIT raw scores to the GIPINIT3B analysis effectively negated the impact of the GIP on the mean ability estimates, although increased discrimination among lower-ability students was observed. Given the low reliability of these results, they must be considered with caution.

Table 6.19

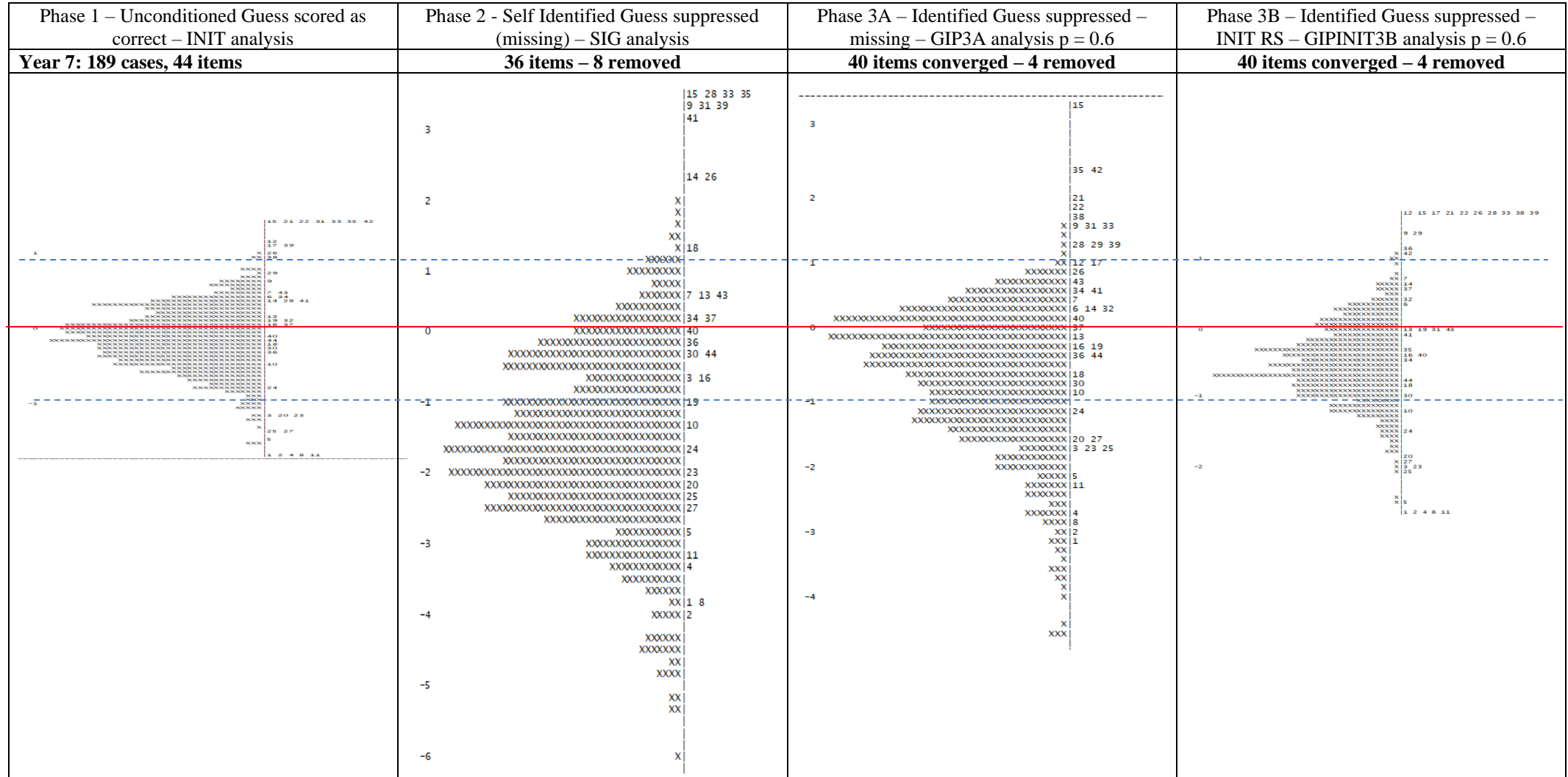
Year 7 Arabic Summary Analysis for Students by Analysis Phase

STUDENTS								
n = 184	Phase 1		Phase 2		Phase 3A		Phase 3B	
			36 items – 8 removed		40 items converged – 4 removed			
	Mean Student ability INIT	Fit Residual INIT	Mean Student ability SIG	Fit Residual SIG	Mean Ability estimate GIP3A p = 0.6	Fit Residual GIP p = 0.6	Mean Ability estimate GIPINIT 3B p = 0.6	Fit Residual GIPINIT 3B p = 0.6
Mean	-0.310	-0.218	-1.369	-0.471	-0.948	-0.354	-0.321	0.986
SD	0.555	1.222	1.291	0.824	0.991	0.667	0.722	1.355
Skewness		0.323		1.244		0.939		0.310
Kurtosis		-0.489		2.333		1.342		-1.028
Correlation		-0.346		-0.089		0.119		-0.313
Separation Index	0.568		0.781		0.780		0.667	

These results reflect the opposite situation compared to the English version of the tests. In these data there were again extreme differences between student ability and item difficulty, but in this case the reverse of the English test situation. However, the ultimate outcome was the same: a constraint of the capacity of the GIP procedure to confirm self-indicated guesses.

Figure 6.14

Item-Student Maps for Each of the Analysis Phases Performed on the Year 7 Data



6.9 Discussion and Implications of the Arabic Test Outcomes

Given the outcomes of the analysis and the random nature of these data, it is inappropriate to draw conclusions from this section of the study, apart from recognising the impact on the GIP procedure of a data set in which the test was too hard for the cohort. These observations may still be valuable in informing the utility of the process in sub-optimal targeting situations.

6.10 Summary

The purpose of this chapter was to report the outcomes of a study that was intended to compare the outcomes of student-identified guesses with the GIP-indicated guesses and, through this comparison, evaluate the effectiveness of the application of the GIP procedures with a small-scale sample. The results show that targeting aspects of the two tests impaired the ability to evaluate the effectiveness of the proposed GIP procedure against these data. Nevertheless, the results provide some insight into the constraints imposed on the protocol by poor targeting.

The results of the English versions of the tests had similar outcomes to those observed in SIM4 (a set of data approximating a test too easy for students), although the field study data were even easier than the respective simulation data. There was also similarity in the outcomes of the Arabic tests and SIM5 (which was simulated to be too hard for the cohort). However, the degree to which the Arabic items in the test were inaccessible and hence the test proved “too hard” for students exceeded the difficulty in SIM5.

In each case the GIP procedure was constrained by the degree of mis-targeting of the tests to the sample populations. However, in the case of the English tests, the procedure produced outcomes generally in accord with the expectations of the hypotheses and findings of Study 1, the simulation study. In contrast, the Arabic tests proved too difficult to make any consistent observations with respect to the overall way the GIP procedure functioned in this data environment.

Despite these short-comings, Study 2 has provided an opportunity to evaluate the GIP procedure with another small sample of live data, and it has allowed some insight into the how the GIP procedure functioned with poorer targeted tests. While the GIP proved to be sub-optimal under the test-targeting conditions of the current study, that it functioned similarly in authentic data as in simulation data suggests the integrity of the assumptions and procedures embedded within the GIP. The next chapter extends these findings by evaluating the GIP with sets of large-scale data generated in cohort-wide system-sponsored tests.

Chapter 7

Study 3: Results of the Application of the Proposed GIP Model with Large-Scale Authentic Data

7.1 Introduction

The simulations and field data described in Chapters 5 and 6, respectively, were used to determine a set of parameters that were consistent in mapping “likely guessing” in student responses that were essentially “Guttman-like”. The purpose of this chapter is to describe the outcomes of the application of the GIP parameters in a large-scale authentic assessment setting. As detailed in Chapter 4, these data were collected from cohort tests given to Grades 4 and 8 (the TIMSS cohorts) for the purpose of estimating how well these cohorts were prepared for the TIMSS assessments.

Although these cohort tests were constructed with a combination of multiple choice (MC) and constructed response items, only the MC items were included in these analyses. The MC items were of the typical structure, with four distractors, one of which was the correct key. The content and contextual validity (Messick, 1989) of the test were evidenced by the processes implemented in the development of the tests. That is, each test was linked to the curriculum by detailing the outcome assessed in each item. The tests were constructed to a predefined and evaluated test specification written within an overall assessment framework. The test structure covered three cognitive domains and a number of proposed item difficulties that extended the range of the anticipated ability of the target cohorts (Thurstone, 1928). Prior to implementation of the test, the items and overall test structures were reviewed by curriculum and assessment experts of the governing educational authority independently of the test developers. The test instruments were administered in an online environment with the data collated centrally via cloud technology. Three data sets were investigated: Grade 4 Mathematics cohort ($N = 26,280$ students); Grade 4 Science cohort ($N = 26,070$); and Grade 8 Mathematics cohort ($N = 21,003$ students). De-identified student responses were collected along with the response time taken for each item.

The test development processes just described are consistent with Rasch’s concept of a “model of intent”. These tests were unidimensional, and their development accounted for the target cohorts’ stages in their educational journey.

7.2 Plan for Analysis (PfA)

Since these are live authentic data collected under typical examination conditions, there are no indicators of which items were responded to with a guess, as there were in the simulated data (Study 1) and in the small-scale field study (Study 2). Consequently, the PfA was a two-step process, with only the INIT and GIP analyses conducted. As these data were “authentic”, there are cases within each set where a null response was observed.

These were coded as “9” in the original data. Consequently, to enable distinctions between student-defined omissions (missing data) and GIP-implemented conditioning of the data, the items indicated as guessed by the GIP process were recoded as “7” in the conditioned data.

Using the ability estimates and item locations extracted from the INIT analysis, the probability of a correct response for each item/student interaction was calculated, together with an item/student residual coefficient for each interaction taking account of the response correctness. The calculation algorithms for these processes are described in Chapter 4 and shown in Eqn 4.5. A number of supplementary analyses were also conducted to determine fit statistics and the effect sizes of the implementation of the protocol with these large-scale data.

7.1.1 Application of the GIP Procedure

Following from the results reported in Chapters 5 and 6, and the final parameters they supported, the GIP procedure applied to these authentic data was defined as having:

- the p value for ability estimates for a correct response at 0.6;
- a probability of a correct response at, or less than, 0.25 ($\Pr(1) < 0.25$); and
- the item student residual at a critical level of 1.75 ($r \geq 1.75$), based on the estimates from the INIT analysis.

At the individual item/student response level, a failure of the GIP criteria was indicative of a probable guess for the item, which was re-coded to “missing data” (“7”). The subsequent analysis of this set of conditioned data produced a new set of item parameters that were impacted by the suppression of the item/student responses indicated by GIP as guessing.

Two sets of salient parameters were generated from these analyses:

1. GIP3A: An item location estimate based on the data, in which the GIP-indicated items were re-coded as missing – a reduction in the student raw score compared to the INIT analysis together with revised item locations that have accounted for probable guessing; and
2. GIPINIT3B: A student estimate and scaled score based on the INIT raw score for the set data, with the item locations derived in GIP3A anchored in this analysis.

Following these Rasch analyses, supplementary analyses of fit, effect size, and comparisons of the GIP-identified items (with the expected Guttman-like structure) were conducted to assess the efficacy of the procedure with a large-scale data set. The next section reports on the results of each of these analyses. The chapter concludes with a brief discussion of the outcomes of applying the GIP with these large-scale data.

7.3 PFA Step 1, Rasch Analysis of the Large-Scale data

7.3.1 Summary Item Statistics

As also found in the research of Andrich et al. (2012), there were significant increases in the distributions of the item locations when the likely guessed items had been accounted for. The comparative ranges of item difficulty locations shown in Tables 7.1 to 7.3 show the relative differences in the item parameters generated by the application of the GIP procedures to the data, compared to the INIT parameters.

The distribution of item locations increased by a factor of almost 2 across the three data sets. These results are consistent in direction with the outcomes of the simulations of Study 1, which shows that the GIP procedure functions consistently with large-scale data sets as indicated by the simulations.

Table 7.1

Comparison of Rasch Analysis G4 Mathematics – Item Parameters INIT, GIP, and GIPINIT

Item Statistics	G4 Mathematics		
	Items INIT data	Items data GIP3A _{p=0.6}	Items GIPINIT3B _{p=0.6} data
	Location	Location	Location
Mean	0	0	0
SD	0.796	1.542	1.542
Min δ	-1.565	-2.769	-2.769
Max δ	0.955	2.398	2.398
Range δ	2.520	5.167	5.167

Table 7.2

Comparison of Rasch Analysis G4 Science – Item Parameters INIT, GIP, and GIPINIT

Item Statistics	G4 Science		
	Items INIT data	Items data GIP3A _{p=0.6}	Items GIPINIT3B _{p=0.6} data
	Location	Location	Location
Mean	0	0	0
SD	0.617	1.169	1.169
Min δ	-1.736	-2.750	-2.750
Max δ	0.956	2.272	2.272
Range δ	2.692	5.022	5.022

Table 7.3

Comparison of Rasch Analysis G8 Mathematics – Item Parameters INIT, GIP, and GIPINIT

Item Statistics	G8 Mathematics		
	Items INIT data	Items data GIP3A _{p=0.6}	Items GIPINIT3B _{p=0.6} data
	Location	Location	Location
Mean	0	0	0
SD	0.492	1.186	1.186
Min δ	-0.971	-1.903	-1.903
Max δ	0.941	2.675	2.675
Range δ	1.912	4.578	4.578

Note. The item parameters for the GIPINIT3B analysis are identical to the GIP parameters, by definition, in the anchored analysis.

7.3.2 Summary Student Ability Statistics

The impact of the GIP process on the summary statistics of the ability estimates of the students was considerable compared to the INIT values (Tables 7.4 to 7.6). As hypothesised, the relative ability estimate of the higher-ability students had increased, as shown in the comparison of the *Max β* results in each analysis, between the INIT data compared to the GIP3A_{p=0.6} (indicated guess scored as missing) outcomes and in the GIPINIT3B analysis. The mean ability estimates of the GIP3A and GIPINIT3B analyses were universally lower than the INIT analysis in all analyses. The range of the distributions of ability estimates, as shown by the comparison of the standard deviations of each analysis, also increased compared to the INIT analysis.

These results are consistent with the outcomes of the simulated data for a large cohort (SIM3) and indicate that the GIP procedures for each data set functioned uniformly with the data, as predicted by Study 1 and Study 2.

Table 7.4

Comparison of Rasch Analysis G4 Mathematics Ability Estimates – INIT, GIP, and GIPINIT

Statistics	G4 Mathematics		
	Student INIT data	Student Pr(Guess) GIP3A _{p=0.6}	Student GIPINIT3B _{p=0.6} data
	Estimate	Estimate	Estimate
Mean	-0.398	-1.144	-0.512
SD	1.096	1.890	1.395
SEm	0.590	0.800	0.650
Min β	-2.744	-4.139	-3.391
Max β	3.536	4.210	4.210
Range β	6.280	8.349	7.601

Table 7.5

Comparison of Rasch Analysis G8 Mathematics Ability Estimates – INIT, GIP, and GIPINIT

Statistics	G8 Mathematics		
	Student INIT data	Student Pr(Guess) GIP3A _{p=0.6}	Student GIPINIT3B _{p=0.6} data
	Estimate	Estimate	Estimate
Mean	-0.438	-1.260	-0.559
SD	0.936	1.739	1.118
SEm	0.50	0.78	0.54
Min β	-2.744	-3.999	-3.999
Max β	3.536	4.231	4.231
Range β	6.280	8.230	8.230

Table 7.6

Comparison of Rasch Analysis G4 Science Ability Estimates – INIT, GI,P and GIPINIT

Statistics	G4 Science		
	Student INIT data	Student Pr(Guess) GIP3A _{p=0.6}	Student GIPINIT3B _{p=0.6} data
	Estimate	Estimate	Estimate
Mean	0.034	-0.508	0.033
SD	1.112	1.888	1.291
SEm	0.51	0.69	0.54
Min β	-3.738	-4.055	-3.738
Max β	3.673	4.157	4.226
Range β	7.411	8.262	7.974

7.3.3 Standard Errors

The Standard Error of the measure (SEm) has also been reported in Tables 7.4 to 7.6. It is calculated as:

$$SEm = S\sqrt{1 - r_{xx}} \quad \text{Eqn 7.1}$$

where r_{xx} is the reliability index – the Separation Index in the case of incomplete data
S is the standard deviation about the mean.

In each analysis, the value of the SEm for the GIP analysis was higher than the relative INIT analysis. The higher values of the SEm of the GIP result compared to the INIT results are due to the higher standard deviations observed in the GIP analyses, and the higher reliability index. Together these reduce the value within the square root sign and with the increased standard deviation increase the SEm. A similar impact is observed when comparing the GIPINIT statistics with the INIT and GIP statistics. An increase in SEm usually means that the confidence in the precision of the measure was reduced. However, in this case, the statistic for the GIPINIT3B analyses was plausibly interpreted as a marginally higher range of scores in the

confidence level about the true score, but it relates to a better scale of the trait of interest. The SEM is greater, but the variable is more precise.

7.3.4 Reliability Indices

In each of the analyses of the data sets, the reliability coefficient improved following the implementation of the GIP3A and GIPINIT3B procedures compared to the INIT analysis. Table 7.7 indicates that there was a better fit of the GIP-conditioned data to the RM, as well as a higher reliability in the outcomes achieved as a result of implementation of the GIP3A and GIPINIT3B procedures for each of the data sets analysed.

Table 7.7

Comparison of Reliability Indices for Each Analysis INIT vs GIP Conditioned Data Phases

Student Statistics	G4 Mathematics		
	Student INIT data	Student GIP3A $p=0.6$	Student GIPINIT3B $p=0.6$ data
Separation Index	0.714	0.822	0.782
Student Statistics	G4 Science		
	Student INIT data	Student GIP3A $p=0.6$	Student GIPINIT3B $p=0.6$ data
Separation Index	0.791	0.868	0.823
Student Statistics	G8 Mathematics		
	Student INIT data	Student GIP3A $p=0.6$	Student GIPINIT3B $p=0.6$ data
Separation Index	0.710	0.801	0.764

7.4 Display of Distributions – Student/Item Maps for Test

7.4.1 Graphic Representation of the Rasch Outcomes

Figures 7.1 to 7.3 show the distributions of items and student abilities of the INIT analyses and the GIP analyses for each of the three tests investigated. The GIP analyses shown include both phases – the GIP3A and the GIPINIT3B analysis. Each of Figures 7.1 to 7.3 displays the compression of the scales generated by the INIT analyses compared to the GIP analyses, with the range of student ability estimates and item locations in the GIP analyses greater than those in the INIT analyses. To show this feature explicitly, each analysis has been re-sized to centre the item difficulties at zero and to approximate the relative scale of the range of parameters displayed.

The increase in the range of item difficulties and consequent discrimination between student ability estimates (Figures 7.3 to 7.5) has been shown to be a direct result of the identification and suppression of the probable guessing in the GIP analyses of the student response data in the outcomes of Study 1 and Study 2.

In Grade 4 Mathematics the effective range of item locations in the INIT analysis was 2.4 logits (-1.2 to +1.2). In Grade 8 Mathematics the comparable range was just 2.0 logits (-1.0 to +1.0), and in Grade 4 Science, excluding the outlier, it was also 2.0 logits (-1.0 to +1.0). The impact of this low range of discrimination in the difficulty locations was to constrict the estimation of ability estimates as demonstrated previously. In relation to the GIP procedure, the homogeneity of these locations means the difference between the estimated ability and items location that defines the indication of a guess is reduced, which has a negative impact on the capacity of the protocol to indicate probable guesses.

The critical difference between item location and student ability estimate is 1.1 logits to produce a probability of a correct response of less than 25%. With item locations compacted, and consequent ability estimates constrained, this difference is observed relatively infrequently in the mid-range interactions and is restricted in the upper ability regions. However, it was far more common in the lower-ability groups where the guessing is more likely to occur.

Figure 7.1 shows that the order of the items was relatively constant; however, the distance between the item difficulty locations had increased in the GIP analysis to produce a wider scale of parameters. Although the mid-range items had very similar difficulty locations in both the INIT and the GIP analyses (e.g., Items 2 and 10 near zero), there was an increasing difference apparent at the upper and lower extremes of the scale for the same item in the GIP analysis relative to the INIT analysis. Item 1 was located at approximately -1.7 in the INIT analysis and at about -3.1 in the GIP analysis scale, whilst Item 15 was located at about 3.0 in the INIT analysis scale and at about 4.0 in the GIP analysis scale. Overall, the distributions were similar, with a negative skew evident in both cases. Figure 7.1 also shows the changes in the range of ability estimates generated by the GIP procedure with Grade 4 Mathematics data, with the mid-range candidates showing the depression of the mean statistic, the increased ability estimates of higher-ability students, and the removal of the overestimation of the lower ability Grade 4 students in the INIT analysis.

Figure 7.2 shows similar outcomes for the Grade 8 students' Mathematics data, with the overall test being hard for the cohort, but with similar outcomes in relation to the shift in the estimates of the higher ability, mid-range ability, and lower-ability students relative to the Grade 4 outcomes (Figure 7.1). As shown in Figure 7.3, the GIP3A process on Grade 4 Science data generated a significant difference in the distributions compared to the INIT analysis, but the re-introduction of the INIT raw scores in the GIPINIT3B analysis returned the student ability estimates to values very similar to the INIT analysis. Despite the reintroduction of the INIT data, the higher-ability students' abilities were better recognised by the GIPINIT3B procedure.

Figure 7.1

Comparison of Analysis Distributions – Grade 4 Mathematics INIT vs GIP Conditioned Data

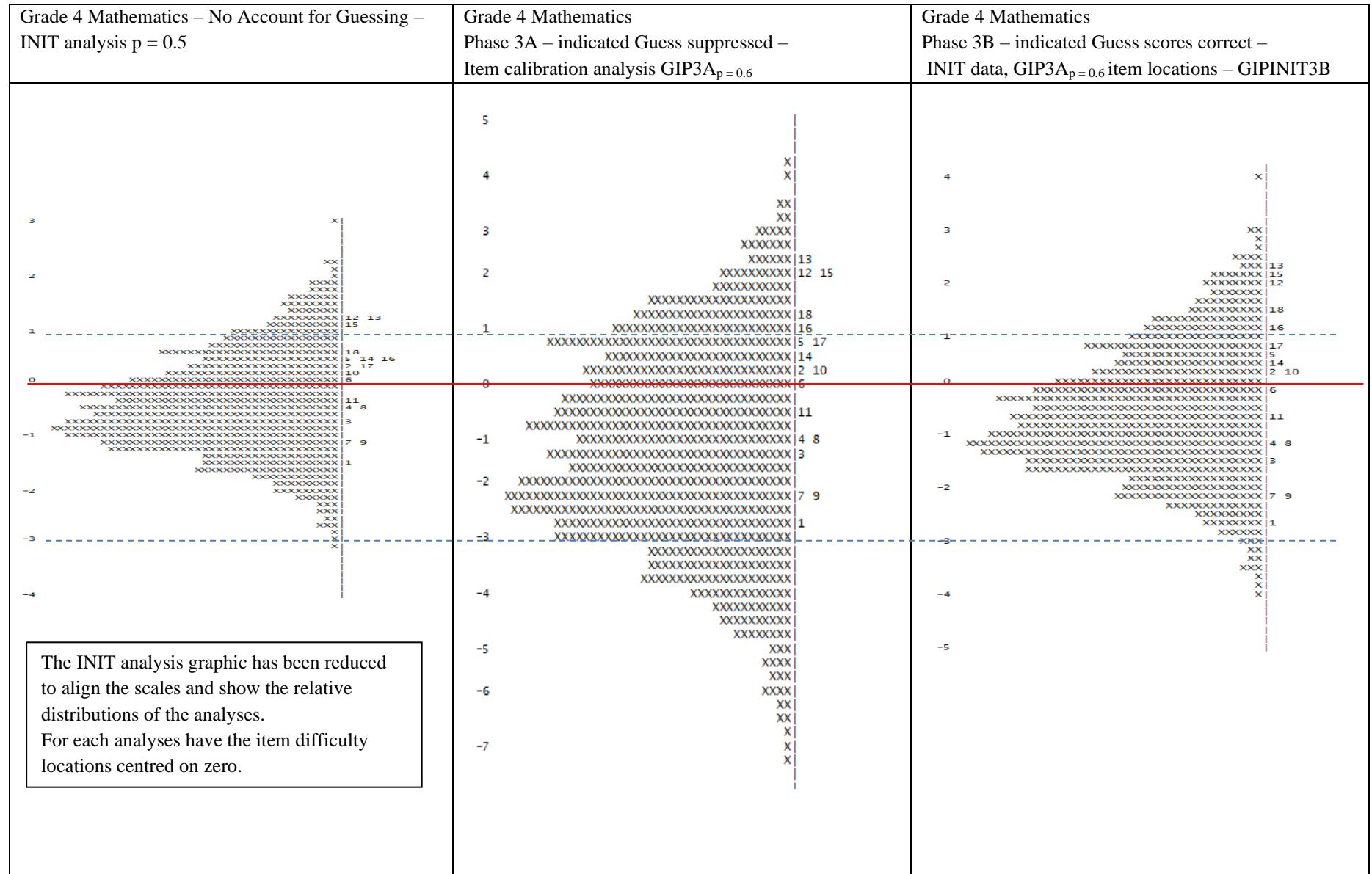


Figure 7.2

Comparison of Analysis Distributions – Grade 8 Mathematics INIT vs GIP Conditioned Data

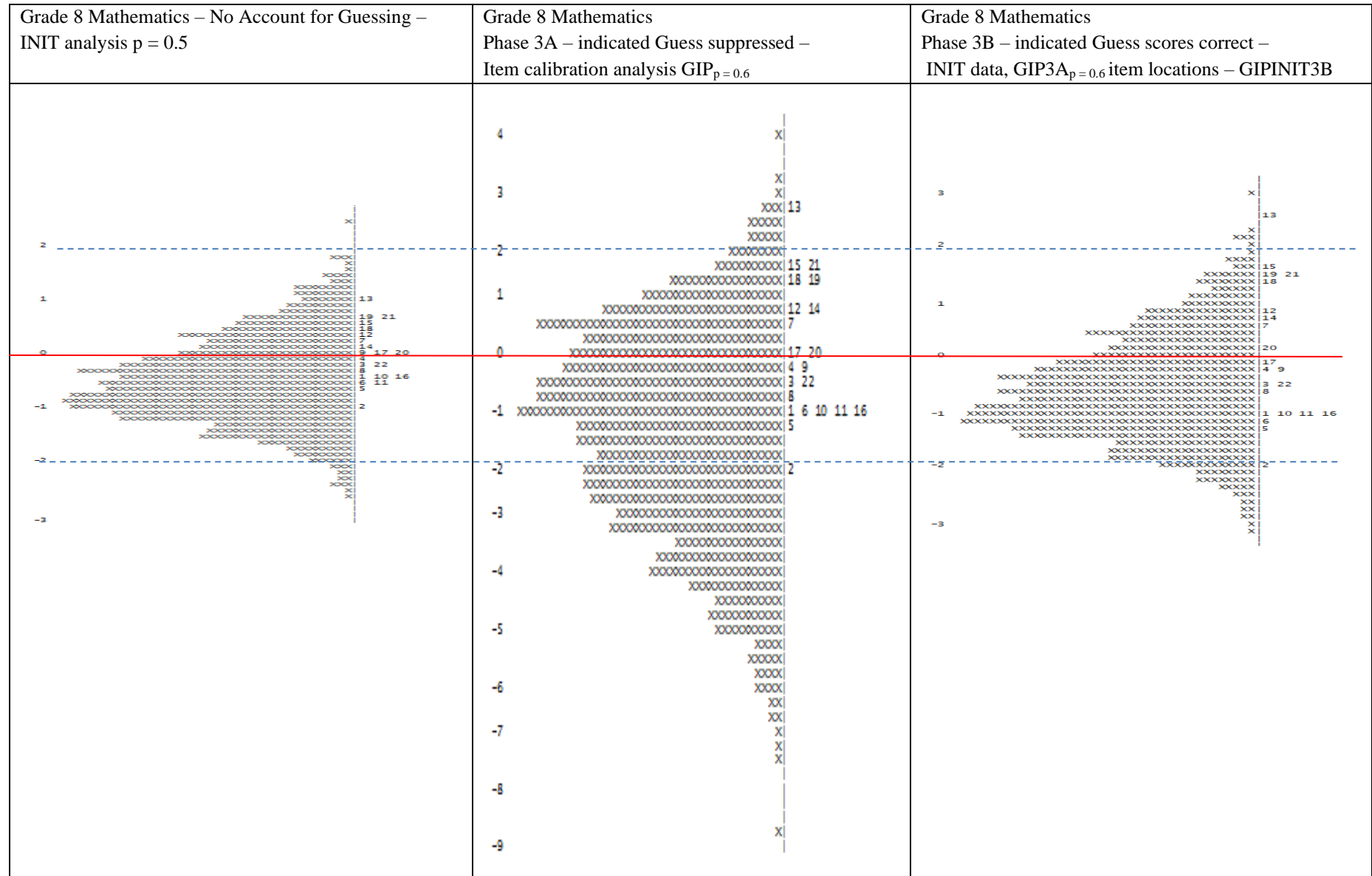
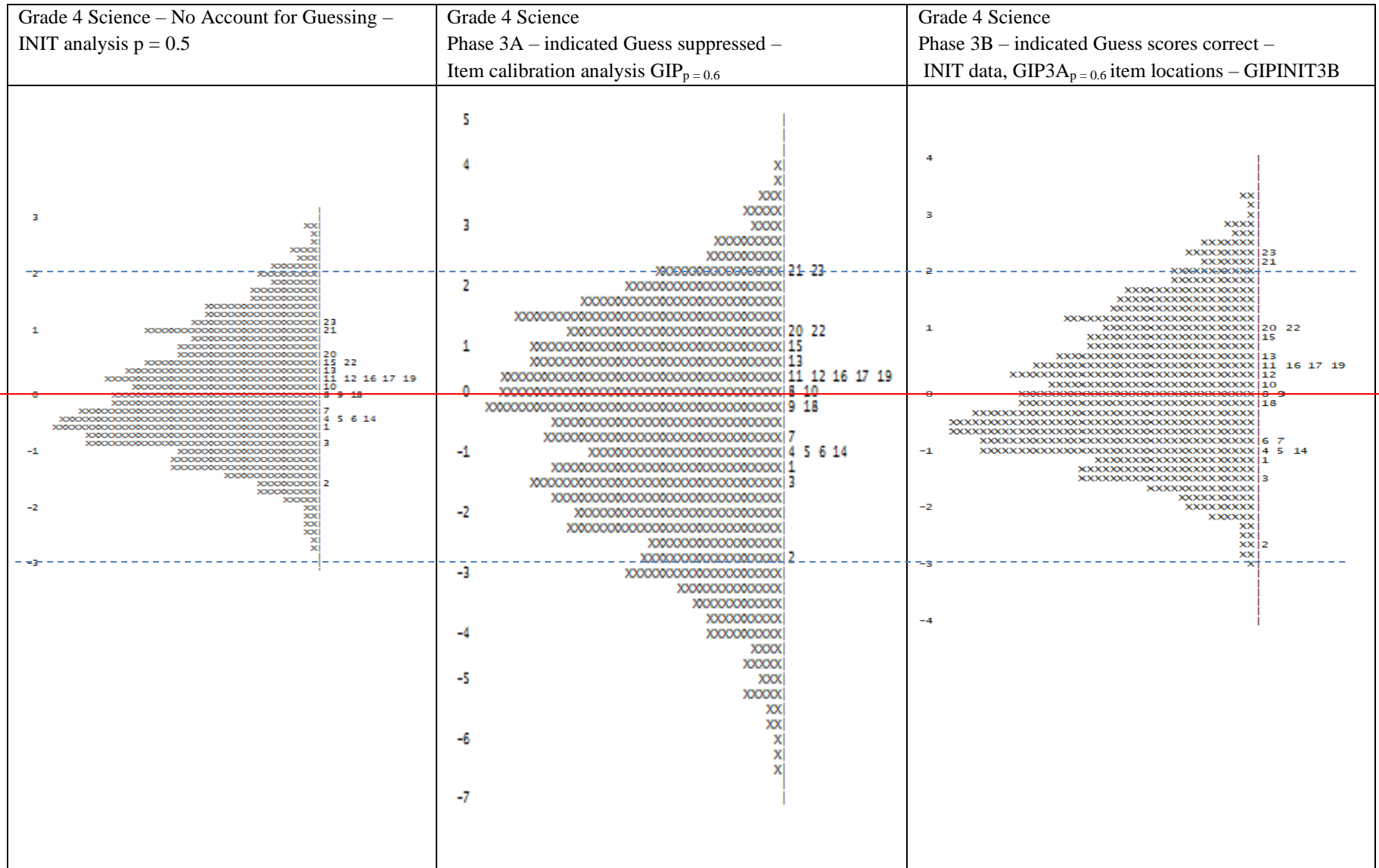


Figure 7.3

Comparison of Analysis Distributions – Grade 4 Science INIT vs GIP Conditioned Data



In each test outcome for the GIPINIT3B analysis, the ability estimates of the higher-ability students increased (see Tables 7.4 to Table 7.6), the ability estimate of the lower-ability students decreased, and the reliability indices improved over the INIT values. The outcomes presented in Tables 7.4 to 7.6 demonstrate improvements in the measurement scale produced by the GIP procedure and a better distribution of item and student ability estimates in the GIPINIT3B phase of the analysis. These results are consistent with the expectations indicated by the simulations, and they also support the logic that underpinned the hypotheses of the study, namely, it is possible to improve the scale by reducing the source of the systematic error due to the noise in the data resulting from correct guesses.

7.5 Shifts in Distributions of the GIP Analyses Compared to the INIT Analysis

This section details the analysis outcomes and comparisons that confirm that the GIP process functioned uniformly, as predicted by the simulated data and as expected in respect of the hypothesis and assumptions that directed the research question.

7.5.1 Item Distributions

Table 7.8 shows the distribution statistics for each analysis. In the Mathematics test, the item distributions of the GIP3A analysis tended to be more normal, with a higher skew value and a reduced kurtosis for both Grades 4 and 8. The impact of the accounting for guessing achieved by the GIP process is intended to remove the underestimation of the item difficulty, which not only generates a wider distribution of item locations but makes the distribution more normal rather than constrained by the bias caused by the guessing in the INIT data. In each of the large-scale cohort Mathematics tests in Study3, the INIT analysis revealed that the tests were marginally hard for the respective cohorts (consistent with SIM 5 – the simulation designed to be ‘too hard’) as indicated by Figures 7.1 to 7.3 and the skew of each GIP distribution (Table 7.8).

Table 7.8

Item Distribution Statistics for Each Analysis

Grade 4			
Mathematics	INIT	GIP3A	GIPINIT3B
Skewness	0.104	0.667	-0.062
Kurtosis	-1.411	-0.247	-1.183

Grade 8			
Mathematics	INIT	GIP3A	GIPINIT3B
Skewness	0.188	1.128	0.045
Kurtosis	-1.599	0.57	-0.393

Grade 4			
Science	INIT	GIP3A	GIPINIT3B
Skewness	0.507	0.202	1.14
Kurtosis	-0.458	-0.084	0.212

7.5.1.2 Student Ability Distributions

By comparison, the statistics shown in Tables 7.9 to 7.11 of the GIP3A analysis are consistent with an increase in the positive skew of the student ability estimates, which resulted from the increase in the higher-ability students' estimates. However, the re-introduction of the raw scores of the INIT data in the GIPINIT3B analysis effectively returns the distribution range to the INIT distributions values for the Mathematics tests. Although the scale of the GIP analysis was improved, the increase in the estimates of the lower-ability students in the GIPINIT3B analysis reduced the efficacy of the process for this group.

Table 7.9

Student Ability Estimate Distribution Statistics for Each Analysis

Grade 4			
Mathematics	INIT	GIP3A	GIPINIT3B
Skewness	0.363	0.967	0.038
Kurtosis	1.823	1.247	-0.643
Grade 8			
Mathematics	INIT	GIP3A	GIPINIT3B
Skewness	0.051	0.926	0.045
Kurtosis	-0.399	1.094	-0.393
Grade 4			
Science	INIT	GIP3A	GIPINIT3B
Skewness	0.465	0.784	0.327
Kurtosis	0.063	0.927	-0.614

7.5.2.2 Distributions of Ability Estimates

In each of the large-scale data sets displayed in Figures 7.1 to 7.3 there was an increased discrimination between the ability estimates of the students across the distribution in each GIP analyses. In all cases, the student estimates in each GIP analysis were in excess of the maximum achievement reported in the INIT analysis, which reflected the increased estimate of some item difficulties and the consequent higher ability estimates for the higher-ability students. By contrast, the distribution of ability estimates for the lower-ability students was “stretched” to reveal a significant proportion of students whose INIT ability estimates had been advantaged by not only their inflated raw score as a consequence of correct guessing, but also by the inflated estimates due to the guessing not being accounted for in the INIT Rasch analysis. This result is consistent with the expectations indicated by the simulations and the logic that underpinned the hypotheses of the study.

Tables 7.10 to 7.15 show the net result of the impact of the GIP procedures, by ability group in each cohort. The compositions of the groups were as follows:

- A higher-ability group comprised of students whose ability estimate was in decile 9 and/or decile 10 (more than one standard deviation above the mean);
- a middle-ability group comprised of students in deciles 3 through 8 (approximately one standard deviation about the mean); and
- a lower-ability group, comprised of students in decile 1 and decile 2 (more than one standard deviation below the mean).

Table 7.10 shows the relative mean ability estimate for each decile. In Grade 4 Mathematics, the size of each decile was about 2,600 students. The table shows the increase in the distribution of estimates, with the range of these averages being about 4 logits in the INIT analysis and about 5.5 logits in the GIPINIT3B analysis. The impact of the GIPINIT3B analysis in Grade 4 Mathematics was to reduce the mean of the ability estimates of students in deciles 1 through 7 with a break-even point (where the INIT and GIPINIT3B values intersect) located in Decile 8. The mean of the students in Deciles 9 and 10 was higher after implementation of the GIP procedure. Thus, the GIP procedure produced mean ability estimates that were higher in the higher deciles and lower in the lower deciles. This is consistent with the expectation of applying the GIP procedure relative to the nature of the data set to which it was applied. Table 7.10 also shows the degree to which estimates were overestimated for the lower-ability students and, to a lesser extent, the underestimation of the higher-ability students. In the middle-range ability there was some variation in the distribution of abilities, which is reflected in Figure 7.4; however, as observed in the simulated data, the change in the middle-range ability students was marginal. In these data the upper two deciles were disadvantaged by a simple Rasch analysis that did not account for guessing, while the lower 5 deciles were all advantaged by a process that did not account for probable guessing (Table 7.11).

Table 7.10

Grade 4 Mathematics Comparison of Mean Ability Estimate by Decile

Decile	Mean INIT ability estimate by Decile (logits)	Mean GIP3A_{p=0.6} ability estimate by Decile (logits)	Mean GIPINIT3B_{p=0.6} ability estimate by Decile (logits)
10	1.679	2.126	2.126
9	0.757	0.994	0.994
8	0.297	0.284	0.401
7	-0.089	-0.340	-0.104
6	-0.381	-1.129	-0.490
5	-0.645	-1.793	-0.838
4	-0.903	-2.244	-1.175
3	-1.165	-2.697	-1.511
2	-1.488	-3.246	-1.918
1	-2.163	-3.954	-2.720

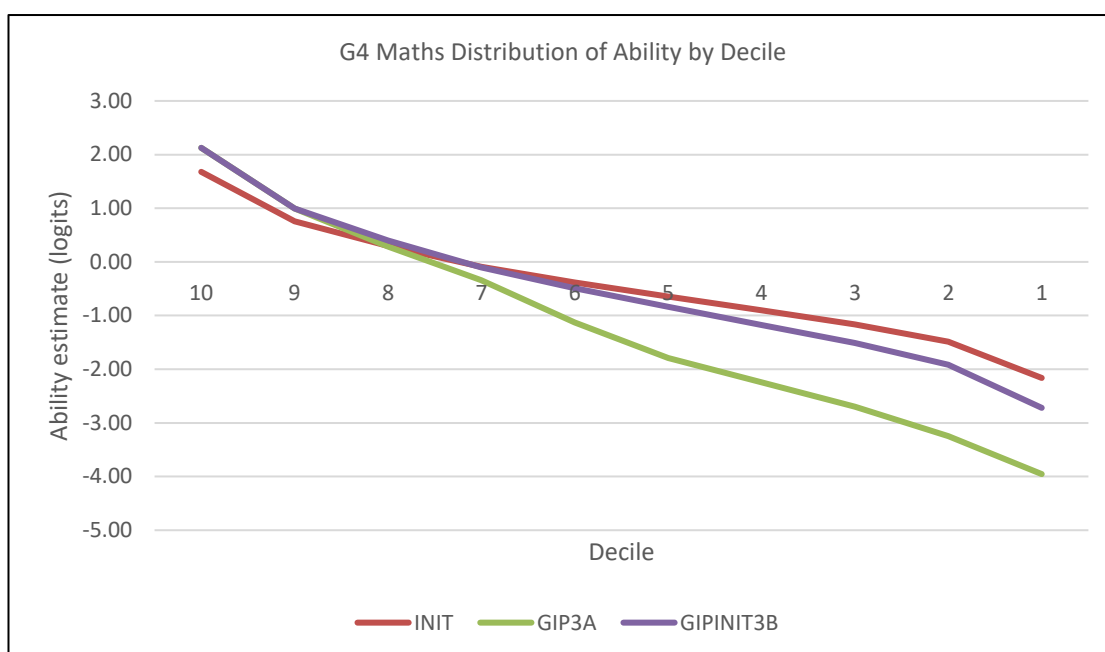
Table 7.11

Grade 4 Mathematics Comparison of Mean Ability Estimates by Ability Groupings

Group	Mean INIT ability estimate by Group (logits)	Mean GIP3A _{p=0.6} ability estimate by Group (logits)	Mean GIPINIT3B _{p=0.6} ability estimate by Group (logits)	Δ INIT vs GIPINIT3B
Higher-ability	1.218	1.560	1.560	+0.342
Middle-ability	-0.481	-2.436	-0.620	-0.139
Lower-ability	-1.825	-3.599	-2.318	-1.493

Figure 7.4

Grade 4 Mathematics Break-Even Mean Ability Estimates by Decile



Tables 7.12 and 7.13 show a similar pattern for the ability distribution of Grade 8 Mathematics as for the Grade 4 cohort. The lowest decile was overestimated by about .3 of a logit, which is approximately one third of a standard deviation in this test, or approximately four months of learning (Choppin, 1983). The break-even point again fell in Decile 8, with the upper two deciles being under-estimated by the INIT analysis, which took no account of probable guessing.

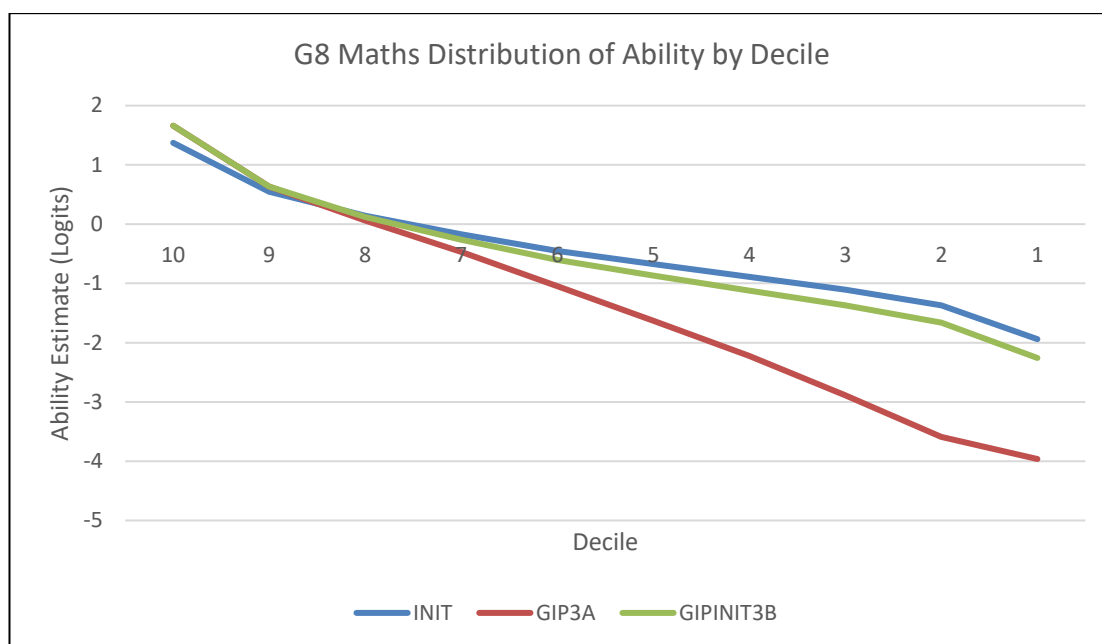
As shown in Table 7.12, there was little variation in the middle-level ability students, and similar mean values of variation for the overestimation of the lower-ability students and the underestimation of the higher-ability students after implementation of the GIP procedure. The reduced range in these variations may be a function of the lack of discrimination evidenced in the INIT analysis, with most of the cohort performing within the -1 to 2 logit range on the analysed scale and the relatively few MC items in each analysis.

The points at which accounting for guessing (break-even point) impacts students' ability estimates are made explicit in Figures 7.5 and Table 7.13. These data follow a similar pattern to those displayed in the Grade 4 Mathematics data. The Grade 4 Science test also displayed a relatively compressed scale for most students. The effective ability range of the estimates of students' ability was within the range -2 to +2.5 logits on the INIT-analysed scale. The Grade 4 test was relatively equally distributed about the Decile 5 break-even point, with relatively small variations in ability estimated revealed by the GIP procedure implementation.

Table 7.12

Grade 8 Mathematics Comparison of Ability Estimate by Decile

Decile	Mean INIT ability estimate by Decile (logits)	Mean GIP3Ap=0.6 ability estimate by Decile (logits)	Mean GIPINIT3B p=0.6 ability estimate by Decile (logits)
10	1.372	1.661	1.661
9	0.548	0.633	0.633
8	0.141	0.059	0.120
7	-0.170	-0.468	-0.263
6	-0.453	-1.045	-0.605
5	-0.675	-1.631	-0.868
4	-0.891	-2.226	-1.119
3	-1.108	-2.887	-1.368
2	-1.371	-3.588	-1.664
1	-1.941	-3.963	-2.258

Figure 7.5*Grade 8 Mathematics Break-Even Mean Ability Estimates by Decile***Table 7.13***Grade 8 Mathematics Comparison of Mean Ability Estimates by Ability Groupings*

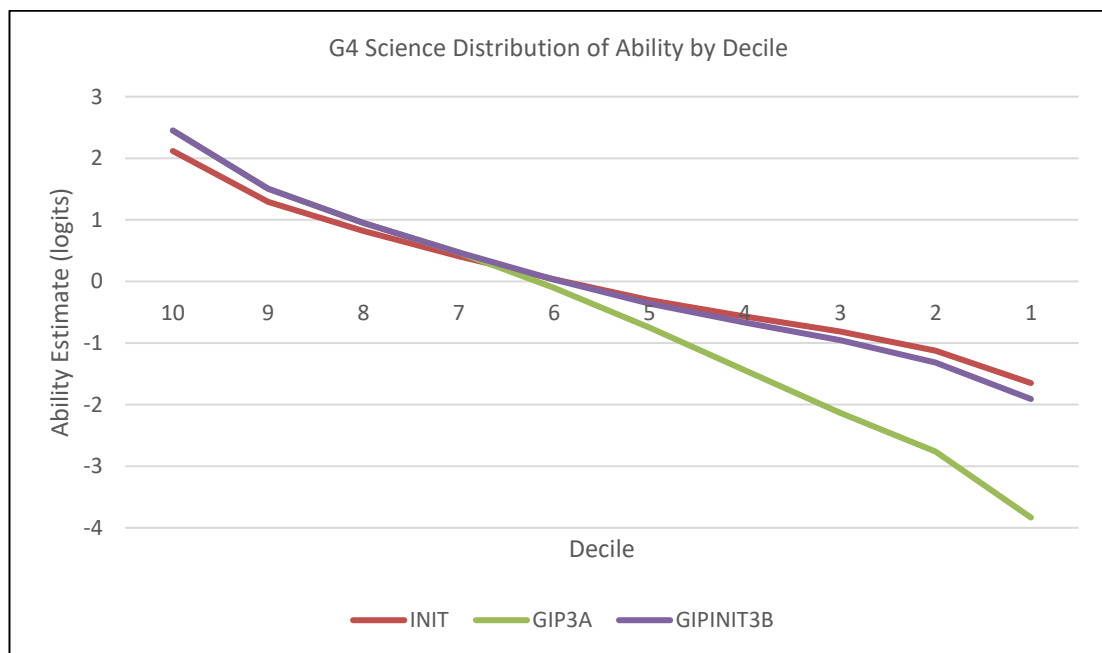
Group	Mean INIT ability estimate by Group (logits)	Mean GIP3Ap=0.6 ability estimate by Group (logits)	Mean GIPINIT3B p=0.6 ability estimate by Group (logits)	Δ INIT vs GIPINIT3B
Higher-ability	.951	1.135	1.135	+0.184
Middle-ability	-0.519	-1.351	-0.676	-0.157
Lower-ability	-1.468	-3.693	-1.771	-0.303

Although the differences in ability estimates between the INIT and GIPINIT values were relatively small (Table 7.15 and Figure 7.6), it is worth noting that for this more targeted test the break-even point was more central in the Decile5/Decile 6 region, all students above that point were advantaged by the GIPINIT process, and all the students below that point had ability estimates revised downwards (Table 7.14). This is consistent with the expected outcomes articulated in the hypothesis.

These outcomes are similar to the SIM 3 outcomes, which indicated that the GIP process functioned uniformly across the full distribution when tests are well targeted.

Table 7.14*Grade 4 Science Comparison of Mean Ability Estimates by Decile*

Decile	Mean INIT ability estimate by Decile (logits)	Mean GIP3Ap=0.6 ability estimate by Decile (logits)	Mean GIPINT3B p=0.6 ability estimate by Decile (logits)
10	2.118	2.450	2.450
9	1.292	1.504	1.504
8	0.818	0.951	0.951
7	0.410	0.472	0.472
6	0.033	-0.106	0.031
5	-0.304	-0.749	-0.363
4	-0.568	-1.442	-0.670
3	-0.813	-2.133	-0.954
2	-1.126	-2.761	-1.315
1	-1.650	-3.834	-1.909

Figure 7.6*Grade 4 Science Break-Even Mean Ability Estimates by Decile***Table 7.15***Grade 4 Science Comparison of Mean Ability Estimates by Ability Groupings*

Group	Mean INIT ability estimate by Group (logits)	Mean GIP3Ap=0.6 ability estimate by Group (logits)	Mean GIP3Bp=0.6 ability estimate by Group (logits)	Δ INIT vs GIPINT3B
Higher-ability	1.705	1.977	1.977	+0.272
Middle-ability	-0.071	-0.501	-0.089	-0.018
Lower-ability	-1.196	-3.297	-1.393	-0.137

7.7 Summary of Indicated Guesses of the GIP Procedure

7.7.1 Grade 4 Mathematics

In Grade 4 Mathematics, 36,802 responses (~8%) were suppressed by the GIP procedure. The highest proportions of suppressed responses were indicated in the lower and mid-ability groups, but interestingly the highest proportion of suppressed items was not in the lowest ability group. This presumably was a function of the lower-ability group having the highest count of non-attempted items, which reduced the proportion of overall items the GIP had an opportunity to condition.

Table 7.16 summarises the comparison of the INIT data responses for the Grade 4 Mathematics test when the GIP procedure has been implemented to identify probable guesses.

Table 7.16

Comparison of Number of Items Re-Coded by GIP Procedure by Ability Group – Grade 4 Mathematics

Cohort	Group	Items not attempted	Count of Guesses indicated GIP3A $p=0.6$	% of responses suppressed by GIP3A $p=0.6$
Grade 4 Mathematics:				
n = 26,279 items: 18 Decile group size: 2,680	Top Decile 10	20	0	0.0%
	Decile 9	37	0	0.0%
	Decile 8	65	958	2.0%
	Decile 7	105	1878	4.0%
	Decile 6	113	4628	9.8%
	Decile 5	88	6349	13.4%
	Decile 4	149	6420	13.6%
	Decile 3	250	6270	13.3%
	Decile 2	317	5985	12.7%
	Decile 1	1740	4314	9.1%
	Overall	2884	36802	7.8%

Note. The effect of the GIP recoding has been disaggregated by ability groups based on the INIT analysis outcomes. For comparison purposes the students were ordered by ability estimates and grouped by decile. The table shows the number of items not attempted in the original data set and the number of guesses indicated by the implementation of the GIP procedure within each decile.

Table 7.17 makes explicit the relationships between the item difficulty, the student ability, and the probable guesses indicated by the GIP procedures. It shows that at the higher ability deciles (9 and 10), no instances of possible guessing were indicated. The overall pattern of GIP identifications (shown in Table 7.2 to be very Guttman-like in its structure) supports the initial assumption that lower-ability students tend to guess the harder items more often than higher-ability students. The lower values of GIP items identified with the lower-ability students in the most difficult items are explained by the increased omission rate by this group for these items.

Table 7.18 follows the same format as Table 7.16 for the Grade 8 Mathematics cohort. The table shows similar patterns with relatively few items not attempted by the cohort. As also shown in Table 7.16, a higher proportion of items were not attempted by the lower ability decile. Most items re-coded by the GIP procedure were in the middle-range and lower ability deciles. These outcomes were largely as expected, with the unexpected anomaly of the lowest ability decile having the highest incidence of non-attempted items, influencing the proportion of GIP-indicated guessed items.

Wright and Stone (1989) predicted that lower-ability groups would discontinue the test and tend to display a higher proportion of non-attempts for items of higher difficulty. This is consistent with patterns of each of the test cohorts reported in these data, with the highest number of omitted items observed in the lower-ability groups.

Table 7.17

Grade 4 Mathematics Count of GIP Implementations by Item Difficulty (Logits)

δ	-1.879	-1.484	-1.416	-0.984	-0.828	-0.81	-0.572	-0.155	-0.028	0.008	0.058	0.12	0.152	0.182	0.272	0.719	0.85	0.922
Cognitive Domain	Know	Apply	Apply	Know	Apply	Know	Know	Know	Apply	Apply	Reason	Apply	Apply	Reason	Apply	Apply	Apply	Reason
Decile	G4MQ01	G4MQ09	G4MQ07	G4MQ03	G4MQ08	G4MQ04	G4MQ11	G4MQ06	G4MQ10	G4MQ02	G4MQ23	G4MQ05	G4MQ22	G4MQ16	G4MQ24	G4MQ19	G4MQ14	G4MQ13
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	77	429	452
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	261	618	524	475
6	0	0	0	0	0	0	0	0	0	448	531	432	565	406	723	596	531	396
5	0	0	0	0	0	0	0	373	414	603	780	645	861	546	738	567	474	348
4	0	0	0	0	0	0	521	454	653	479	729	582	731	426	656	475	415	299
3	0	0	0	0	289	313	747	363	542	380	650	468	622	299	598	403	341	255
2	0	0	0	273	569	629	634	265	418	279	510	407	507	201	509	295	267	222
1	51	216	431	415	358	363	353	131	258	152	272	210	276	93	282	192	134	127

Table 7.18
Comparison of Number of Items Re-coded by GIP Procedure by Ability Group – Grade 8 Mathematics

Cohort	Group	Items not attempted	Count of Guesses indicated by GIP3A	% of responses suppressed by GIP3A
Grade 8 Mathematics n = 21004 items: 22 Group Size: 2100	Top decile	10	0	0.0%
	Decile 9	52	0	0.0%
	Decile 8	45	513	1.1%
	Decile 7	64	1857	3.8%
	Decile 6	113	3728	7.7%
	Decile 5	100	5775	12.0%
	Decile 4	143	7276	15.1%
	Decile 3	109	8486	17.6%
	Decile 2	261	8553	17.7%
	Decile 1	1904	5791	12.0%
	Overall		2801	41979

Table 7.19
Grade 8 Mathematics Count of GIP Implementations by Item Difficulty

δ	-										0.05	0.06	0.16		0.58		0.68					
	0.971	0.575	0.515	0.492	0.446	0.434	0.382	0.224	-0.2	0.106	0.049	3	9	0.129	4	0.2	0.363	0.482	5	6	0.72	0.941
Cog	Reaso										Reaso		Reaso		Reaso		Reaso					
	Know	Reason	Apply	Apply	Reason	Apply	Know	Know	n	Know	Know	Apply	Apply	n	Apply	Know	n	n	Apply	Apply	n	n
Decile	MQ03	MQ06	MQ08	MQ13	MQ19	MQ12	MQ01	MQ10	MQ25	MQ04	MQ05	MQ11	MQ23	MQ20	MQ16	MQ09	MQ14	MQ21	MQ17	MQ24	MQ22	MQ15
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	205	308
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	82	50	434	313	563	415
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	266	268	628	558	615	430	504	459
5	0	0	0	0	0	0	0	0	0	0	314	290	369	268	714	649	592	642	572	432	451	482
4	0	0	0	0	0	0	0	326	358	292	549	420	649	437	662	633	587	571	544	424	372	452
3	0	0	332	304	302	274	225	454	528	407	460	331	556	364	649	578	497	529	496	430	339	431
2	0	238	670	577	419	504	360	359	417	292	389	239	412	295	612	481	369	501	381	421	263	354
1	233	244	460	389	188	304	238	226	252	180	313	144	261	193	360	346	222	308	239	335	151	205

Table 7.20 shows the impact of the GIP process on the response data for a cohort of Grade 4 Science students. It shows a similar pattern of responses and GIP impact for Science as for Mathematics in both grades. The uniformity in the increasing number of re-coded items using the GIP protocol follows the anticipated pattern, with increasing identification of probable guessing as student ability decreased. The alignment of these outcomes with the outcomes predicted in the hypothesis may relate to the fact that the Science test was better targeted to the cohort than each of the Mathematics test outcomes, as shown in Figure 7.3, although relatively small distributions of the range of item difficulties were observed, which may be a function of the homogeneity of the MC item locations with little discrimination between the relative difficulties of the items.

Table 7.20

Comparison of Number of Items Re-coded by GIP Procedure by Ability Group – Grade 4 Science

Cohort	Group	Items not attempted	Count of Guesses indicated by GIP3A	% of responses suppressed by GIP3A
Grade 4 Science: n = 26070 items: 23 Group size: 2607	Top decile	3	0	0.0%
	Decile 9	25	0	0.0%
	Decile 8	55	0	0.0%
	Decile 7	129	8	0.0%
	Decile 6	149	1638	2.7%
	Decile 5	106	4259	7.1%
	Decile 4	115	7674	12.8%
	Decile 3	171	10075	16.8%
	Decile 2	350	10003	16.7%
	Decile 1	2074	9118	15.2%
	Overall		3177	42775

The Science ability estimates for the Grade 4 cohort may have been impacted by concepts, language, and interpretation of questions expressed at a level that were challenging for some students. A review of the overall response patterns did not reveal any time-limit effects that may have contributed to the very different distribution of non-attempted items across the ability groups shown in Table 7.20 when compared to Tables 7.16 and 7.18.

It is suggested that the Guttman-like pattern in the relationship between the ability groups and the number of items indicated by the GIP process provides some evidence of the effectiveness of the process in these authentic data. It is not unreasonable, given the evidence regarding the difference between item difficulty and the student propensity to guess, to propose that there could be high levels of guessing present in the student responses for these items and that the GIP provided a reasonable approach to identify these guesses (based on hypotheses related to the conditions under which guesses are more/less likely). Table 7.21 displays a similar pattern to that observed in Tables 7.17 and 7.19.

Table 7.21 *Grade 4 Science Count of GIP Implementations by Item Difficulty*

Cog Dom	Know	Know	Know	Know	Apply	Know	Know	Know	Apply	Know	Apply	Apply	Apply	Apply	Apply	Reason	Apply	Reason	Reason	Apply	Apply	Apply	Reason
δ	-1.736	-0.852	-0.576	-0.507	-0.492	-0.462	-0.398	-0.361	-0.09	0.009	0.038	0.164	0.194	0.229	0.247	0.307	0.317	0.351	0.462	0.577	0.676	0.945	0.956
Decile	G4SQ02	G4SQ03	G4SQ01	G4SQ05	G4SQ15	G4SQ04	G4SQ06	G4SQ07	G4SQ19	G4SQ08	G4SQ09	G4SQ20	G4SQ11	G4SQ13	G4SQ17	G4SQ12	G4SQ18	G4SQ14	G4SQ24	G4SQ16	G4SQ21	G4SQ22	G4SQ25
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	249	667	722
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	208	248	242	767	731	551	741	771
4	0	0	0	0	0	0	0	0	0	0	0	559	462	492	472	581	683	860	861	652	530	774	748
3	0	0	0	0	0	0	0	6	768	720	654	919	606	671	757	516	579	684	788	524	523	663	697
2	0	0	146	165	212	166	119	664	727	735	619	755	428	531	571	453	413	579	697	404	446	563	610
1	2	300	404	502	766	400	357	444	521	540	469	574	246	337	431	309	258	348	492	264	327	411	416

7.8 PFA Step 2, Supplementary Analyses

7.8.1 Analysis of Fit

Following the methodology of Andrich et al. (2015, p. 430), an investigation of the mean chi-square was conducted on each of the analyses. Table 7.22 displays the outcomes (see Eqn 5.7). The reduction in the mean chi-square statistic was consistent across all analyses, indicating that the GIP3A procedure produced a better fit of the data to the model than the original unconditioned INIT data. Hence, and consistent with the previous results, the GIP scale appeared to yield a more refined variable and a better measure of the trait than the INIT analysis. This was a result of the removal of the “noise” in the measure caused by the guessed responses.

Table 7.22

Comparison of Reliability Indices for Each Rasch Analysis INIT, GIP3A, and GIPINIT3B

Analysis	INIT			GIP3A $p = 0.6$			GIPINIT3B $p = 0.6$		
	Total Chi-Square	d.f.	Mean Square	Total Chi-Square	d.f.	Mean Square	Total Chi-Square	d.f.	Mean Square
Grade 4 Mathematics	11447	162	70.7	10128	162	62.5	56088	162	346.2
Grade 4 Science	18956	207	91.6	14917	207	72.1	60938	207	294.4
Grade 8 Mathematics	12860	198	64.9	10124	198	51.1	63143	198	318.9

The misfit of the GIPINIT3B analysis was expected. This is a consequence of the introduction of the guessed responses to the more refined GIP scale, which caused the degree to which the guessed items misfitted the refined model to be increased. Hence the total chi-square and consequent mean square statistics were considerably greater in the GIPINIT3B analysis than in the INIT analysis.

7.8.2 Statistical Significance of the GIP Interventions

The t-test for the significance of the difference of the means and the Effect Size analysis was undertaken on these data for completeness. The results are as expected, with the GIP3A analysis displaying significant differences in the means between the GIP3A and the INIT values because the GIP3A process reduced the scores of the students by the number of items indicated as probable guesses. By comparison, the results of these analyses of the GIPINIT3B data are more variable. The reintroduction of the INIT raw scores as a result of crediting students with the indicated probable guesses tends to return the overall scores to the original mean. However, the fact that the scale was revised in the GIP3A process introduces some variation in the degree to which the revised scale scores are changed.

Throughout this research the significant feature of the GIP process is that the GIPINIT3B process does not impact the mean greatly. However, both the GIP3A and GIPINIT3B processes do impact the distribution and uncover higher estimates for the higher-ability students and lower estimates for the lower-ability students. As noted, the lower-ability students are these whose estimates are most biased by not accounting for guessing in the analysis of results using the RM.

Tables 7.23 to 7.25 report *t*-tests that show the difference between the Rasch means from the INIT ability estimates, GIP3A (conditioned data) estimates, and the GIPINIT3B analysis. In the comparison of the mean ability estimates of both Grade 4 and Grade 8 Mathematics, the two-tailed *t*-tests reported in Tables 7.23 and 7.24 indicate that there were significant differences in the mean estimates for GIP3A and the GIPINIT3B compared to the INIT mean statistic.

Table 7.23

Comparison of Means – Grade 4 Mathematics Student INIT Ability Estimate and GIP and GIPINIT Ability Estimates

One-Sample t-Test Grade 4 Mathematics						
Test Value = 0						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
G4INITab	-119.13	26278	0.000	-0.8110	-0.8239	-0.7972
G4GIP3Aab	-100.81	26278	0.000	-1.1990	-1.2223	-1.1757
One-Sample Test Grade 4 Mathematics INIT ability estimate vs GIPINIT ability estimate statistics						
Test Value = 0						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
G4INITab	-119.13	26278	0.000	-0.8106	-0.8239	-0.7972
G4GIPINIT3Bab	-60.46	26278	0.000	-0.5229	-0.5398	-0.5059

Table 7.24

Comparison of Means – Grade 8 Mathematics Student INIT Ability Estimate With GIP and GIPINIT Ability Estimates

One-Sample t-Test Grade 8 Mathematics						
Test Value = 0						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
G8INITab	-130.81	21006	0.000	-0.8511	-0.8638	-0.8383
G8GIP3Aab	-107.47	21006	0.000	-1.3373	-1.3617	-1.3129
One-Sample Test Grade 8 Mathematics INIT ability estimate vs GIPINIT ability estimate statistics						
Test Value = 0						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
G8INITab	-130.81	21006	0.000	-0.8511	-0.8638	-0.8383
G8_GIPINIT3Bab	-73.42	21006	0.000	-0.5704	-0.5857	-0.5552

Table 7.25 shows the outcomes of the *t*-tests of the comparison of the mean statistics of each analysis of Grade 4 Science. There is a significant difference between the means of the INIT analyses compared to the relative GIP3A analyses. This indicates the difference in the more refined variable of the GIP3A analysis compared to the INIT analysis. However, in comparing the GIPINIT3B mean to the INIT mean, the difference is not statistically significant, as shown by the *p*-value of 0.015.

Table 7.25

Comparison of Means – Grade 4 Science Student INIT Ability Estimate With GIP and GIPINIT Ability Estimates

One-Sample t-Test Grade 4 Science						
Test Value = 0						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
G4ScINITab	-54.76	26064	0.000	-0.3803	-0.3939	-0.3667
G4GIP3Aab	-47.04	26064	0.000	-0.5653	-0.5888	-0.5417

One-Sample Test Grade 4 Science INIT ability estimate vs GIPINIT ability estimate statistics						
Test Value = 0						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
G4ScINITab	-54.76	26064	0.000	-0.3803	-0.3939	-0.3667
G4SciGIPINIT3B	2.42	26064	0.015	0.0195	0.0037	0.0353

The non-significance of the comparison of the means of the INIT and GIPINIT3B analyses was expected, given the degree to which the means differed (.001 logits) (Table 7.12). This may be related to the small discrimination in item locations in these data with most locations falling within the range -1.0 to +1.0. It was apparent that the small range in the item locations and consequent student ability estimates impacted on the capacity of the GIP to indicate probable guessed items, due to the interaction between item locations and ability estimates in the RM.

7.8.3 Effect Size

Given the sample size of each of these cohorts, the statistical significance in the variation of the mean statistics of the analyses were significant for most comparisons. An analysis of the Effect Size was conducted to further obtain an estimate of the magnitude of the difference between the mean ability estimates of the GIP and GIPINIT analyses, compared to the INIT analysis for each data set.

7.8.4 Cohen's *d*

Cohen's *d* (Cohen, 1988, 1992) calculates a standardised mean difference between two groups by subtracting the mean of one group from the mean of the other ($M1 - M2$) and dividing the result by the standard deviation (SD) of the population from which the groups were sampled (Eqn 7.2).

$$\text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{Standard Deviation of the group (pooled)}} \quad \text{Eqn 7.2}$$

In this case, the control group is the INIT analysis statistic, and the experimental group is the GIP analysis statistic, or the GIPINIT statistic, as appropriate (Table 7.26).

Table 7.26*Cohen's Interpretation of Effect Size Statistic*

Relative Size	Effect Size 'd'	% of control group below the mean of the experimental group
	0.0	50%
Small	0.2	58%
Medium	0.5	69%
Large	0.8	79%
	1.4	92%

The observed changes in the summary statistics are shown to be statistically significant using the Effect Size calculations.

Using Cohen's d as the indicator the following values are returned for the GIP3A means compared to the INIT means:

Grade 4 Mathematics: $(-1.660 - -0.936) / 1.637 = 0.483$

Grade 4 Science: $(-0.517 - 0.034) / 1.555 = 0.350$

Grade 8 Mathematics: $(-1.279 - -0.557) / 1.428 = 0.589$

The differences in Grade 4 Mathematics and Grade 8 Mathematics for the GIP3A analysis were of the order of about half of a standard deviation change in the mean statistic, which represents a 13% to 19% reduction in the proportion of students who would be reported as having achieved the INIT analysis mean standard. The effect size of the student ability statistic for the Grade 4 Science GIP3A compared to the INIT mean was less significant than the mathematics values, however, the Effect Size statistic approached the medium range in Cohen's matrix (Table 7.26) and confirmed a significant difference in the mean statistics. The Effect Sizes of the GIPINIT3B means compared to the INIT analysis were universally in the small range using the Cohen's *d* scale indicators.

Using Cohen's d as the indicator the following values are returned for the GIPINIT3B means compared to the INIT means:

Grade 4 Mathematics: $(-1.660 - -0.936) / 1.637 = 0.090$

Grade 4 Science: $(-0.517 - 0.034) / 1.555 = 0.001$

Grade 8 Mathematics: $(-1.279 - -0.557) / 1.428 = 0.117$

The Effect Size statistics for the GIPINIT3B analyses were expected to be non-significant, as re-introduction of the INIT raw score re-established the original responses that included the probable guessed items scored as correct. This proved to be the case.

7.9 Discussion

This investigation focused on the distributions of abilities and, although statistical means are important, this research study has highlighted the impact on the range of abilities reported by the analyses performed. It is important to keep in mind the key issues while interpreting these analyses. The statistics indicate that there was a difference in the mean of the GIPINIT3B analysis compared to the mean of the INIT analysis. In comparing the INIT statistics and the GIPINIT3B statistics for the analyses, although the respective means have generally changed only marginally, the distribution of ability estimates uniformly increased in the GIPINIT3B analysis, which was grounded on a more refined variable that was a better measure of the trait.

It was expected that the GIPINIT3B mean would vary marginally compared to the that of the INIT analysis, as suggested by the simulated data analyses. These results confirm the expected outcomes of implementation of the GIPINIT3B processes.

7.9.1 *Significance of Test Targeting and Item Difficulty Distributions*

An observation suggested by the analysis of the simulated data and also evident in these large-scale authentic data relates to the capacity of the GIP procedure to indicate probable guessing when applied to tests of few items with low distributions of item location and consequent student ability estimate distributions.

The impact of a well-targeted or a poorly targeted test was shown to be important in relation to the GIP. In relation to this research, the better targeted the test, the more effective was the GIP procedure for identifying guessing, provided there was a reasonable distribution of item locations. A well-targeted test does not negate the presence of harder and easier items in the test structure. This is a significant issue for test developers and item writers. Findings from the analysis of simulated data were also evident in these large-scale authentic data; that is, the capacity of the GIP procedure to indicate probable guessing differed with different item location and student ability distributions. Figures 7.4 to 7.6 show that the Grade 4 and 8 Mathematics tests were verging on being too difficult for the cohorts tested. By comparison, the Grade 4 Science test was relatively well targeted.

Thurstone (1928) intimated the requirement that items used to measure cohort ability should be of the appropriate general difficulty relative to the cohort. The current research has revealed that it was possible to identify student guessing, and to do this to a better extent when tests were well targeted relative to the participating cohort and constructed with a wide range of item locations that extended across the full range of student ability.

7.10.2 Significance of Accounting for Guessing in These Data

According to Choppin (1983),

A change in achievement of one logit represents a considerable amount of learning. Studies in various parts of the world indicate that in a given subject area, the typical child's achievement level would rise by rather less than half a logit in a typical school year. (p. 4)

The indicated degree to which ability estimates have been changed by the GIP process shown in Tables 7.10, 7.12, and 7.14 suggests that although the changes in terms of logit values may be relatively small they are significant in terms of learning progress.

7.11 Summary

Study 3 had two aims:

1. to confirm the efficacy of the GIP procedure for indicating likely guessed items in student response patterns in large-scale authentic data; and
2. to provide a measure of the degree to which student estimates are biased by procedures that do not account for guessing, the premise being that stakeholders are guessing when using biased information for “data driven” decision making.

The statistics regarding the GIP3A analyses show a better fit to the model compared to the INIT analyses for each data set. This confirms the hypothesis that removing probable guessing from the data improves the quality of the measurement scale of the trait. For both the GIP3A and GIPINIT3B analyses, the reliability index (Separation Index) showed the improved reliability of the test compared to the INIT analysis. The re-introduction of the initial data in the GIPINIT3B analysis resulted in considerable misfit between these data and the model, as shown in the SEM and the mean square statistics. Although the GIPINIT fit statistics are a poorer resultant if considered as individual fit-statistic, they are derived from a better fit of the data to the model based on the GIP analysis.

The benefit achieved by the application of the GIP procedure, compared to previous research outcomes that attempted to account for guessing (Andrich et al., 2015), was the impact on lower-ability students, for whom the estimates of the INIT analysis were inflated. This was a different result compared to Andrich et al.'s (2015) anchored procedure in that the GIP procedure indicated and accounted for the overestimation of the lower-ability students in each of the analysed data sets.

The analyses presented in this chapter, particularly the consistency of improving the estimates of both the upper and lower-ability groups in a large-scale assessment, indicate that the procedure should be applied uniformly to all assessments of these scale that use the RM for analysis and reporting student achievement. Although the targeting of the tests may limit the capacity of the GIP procedure to operate uniformly across the full distribution of students, the outcomes of a GIP procedure are an improvement over a process that takes no account of probable guessing in the tests.

This chapter has demonstrated that the differences between the outcomes of an analysis that did not account for guessing (INIT analyses) and the proposed protocol (GIP procedure) were significant in an authentic large-scale assessment context. Given the current trend towards reporting student ability and achievement outcomes in terms of Bands or Proficiency Bands (e.g., Gonski 2.0; NSW HSC), Chapter 9 presents the outcomes of these analyses in relation to reported scores and achievements in proficiency bands. It also quantifies the degree to which the reclassification of student abilities and achievement levels was apparent in each of the data sets analysed for these tests.

Chapter 8

Response Time as a Factor in Indicating Guessing

Investigation of Student/Item Response Time as a Potential Parameter in Identifying or Confirming Guessed Items

8.1 Introduction

8.1.1 Rationale

The GIP procedure is based on a single piece of data, the response to an item by a student. Calculations involving the Rasch model derive the probability of a correct response and the degree of misfit if a correct response to an item is observed for a student of a given ability.

In the case of the authentic data referenced in Chapter 7 a second piece of datum, item response time, was collected for each student/item interaction. This Chapter is devoted to investigating whether this second data source can be leveraged to improve the consistency and reliability of the GIP procedure in indicating guessed responses.

An advantage of an online testing environment, such as that analysed in Chapter 7, was that the time students took to respond to each item (response time) could be collected. Researchers have examined the interaction of student/item response time with possible guessing and/or test effort (Guo et al., 2016; Lee & Yue, 2014; Michaelides et al., 2020; Setzer et al., 2013). In some assessment programs, such as the MAP Growth Northwest Evaluation Association assessment (<https://www.nwea.org/map-growth>), student reports noted the percentage of items recorded with rapid-response behaviour as an indicator of student effort on a test. This chapter investigates the potential of using student/item response time to support the GIP procedure in the indication of guessing, leveraging the data collected as part of Study 3 (see Chapter 7).

8.2 Plan for Analysis (PfA)

The investigation of item response time as an indicator of guessing had two steps.

Step 1 involved a simple review of the total engagement times of students of different ability, as indicated by the students' resective raw scores, with the tests overall. This relationship was investigated to ascertain if there was an initial indication of a relationship between student ability and item/student response time.

Step 2 involved an analysis of the response times of groups of students of similar ability estimates for items of varying difficulty by reviewing individual response times with INIT item difficulties and student ability statistics. This relationship was investigated to ascertain if there was a relationship between item difficulty, student ability and item/student response time that could be leveraged to improve the GIP procedure. In this step the GIP indications from Chapter 7 were used to observe if there were any relationships between the time parameter and the GIP processes.

The response time of each student/item interaction was compared to an “average” time to evaluate a set of plausible relationships that can be determined to be an indication of a guess.

Consistent with the method of grouping shown in Chapter 7, each cohort was divided into the following three sub-groups for comparison purposes:

Higher-ability group: the students whose INIT analysis estimate located them in deciles 9 and 10;

Mid-range ability group: the students whose INIT analysis estimate located them in deciles 3 to 8; and

Lower-ability group: the students whose INIT analysis outcome located them in deciles 1 and 2.

In deriving this average time, it was observed that the higher-ability students (top two deciles) tended to spend considerably more time on items than the lower-ability students (bottom two deciles) (see Table 8.1 below). Consequently, the mean of the mid-range ability group was used as an indicator of the time taken for a typical student/item interaction and as a reference to determine if an individual response student/item time would be considered “rapid”.

Rule 8.1 was imposed for indicating if a student/item rapid response would be a potential indicator of a guess.

if $\{r_{GIPij} = 1 \text{ AND } T_{ip} < avT\}$; $r_{GIPij} = 6$, else $r_{GIPi} = 1$

Rule 8.1

Three potential time (avT) constraints were considered: when the response item time was half ($T/2$) the defined rapid-response time, when the response item time was one-third ($T/3$), and when the response item time was one-quarter ($T/4$) of the average time of the mid-range ability group.

In general, the addition of an extra constraint on any of the characteristics that define a variable reduces the occurrence of instances that meet the composite’s requirements. As an analogy, if an aim was to determine the number of females in a population, a survey would determine ‘ n ’ cases that meet this criterion; however, a constraint that the females need to be taller than 1.8m would reduce the original frequency of ‘ n ’ to a smaller subsample. In the current study, a reduction in the item/student interactions that would be indicated as a guess was an expected outcome of including time as an additional parameter in the GIP equation.

8.3 PfA Step 1 Results: Interaction of Response Time and Raw Scores

In reviewing the student/item response times in the tests overall, a relationship between the total test response times of students and their ability, as indicated by the raw score, was observed. The lower-ability students tended to spend less time completing the test, whilst generally more time was taken by mid-range and higher-ability students. Tables 8.1 to 8.3 relate to the completion of the MC items in each test.

Table 8.1*Year 4 Mathematics Comparison Between Observed Raw Score and Time Taken on the Test*

Av raw score (/25)	% of cohort	Time Taken on Test
5.7	6.1%	less than 5 minutes
5.9	17.5%	from 5 minutes and less than 10 minutes
7.6	19.6%	from 10 minutes and less than 15 minutes
8.4	19.5%	from 15 minutes and less than 20 minutes
10.3	15.3%	from 20 minutes and less than 25 minutes
10.7	9.9%	from 25 minutes and less than 30 minutes
11.2	12.1%	30 minutes or more

Table 8.2*Year 8 Mathematics Comparison Between Observed Raw Score and Time Taken on the Test*

Av raw score (/25)	% of cohort	Time Taken on Test
6.2	16.0%	less than 5 minutes
7.3	13.1%	from 5 minutes and less than 10 minutes
8.6	12.2%	from 10 minutes and less than 15 minutes
10.0	13.7%	from 15 minutes and less than 20 minutes
11.2	13.8%	from 20 minutes and less than 25 minutes
11.8	11.1%	from 25 minutes and less than 30 minutes
12.5	8.1%	from 30 minutes and less than 35 minutes
13.1	12.0%	35 minutes or more

Table 8.3*Year 4 Science Comparison Between Observed Raw Score and Time Taken on the Test*

Av raw score (/25)	% of cohort	Time Taken on Test
7.0	18.8%	less than 5 minutes
9.4	23.4%	from 5 minutes and less than 10 minutes
13.4	21.1%	from 10 minutes and less than 15 minutes
15.0	17.1%	from 15 minutes and less than 20 minutes
15.3	9.5%	from 20 minutes and less than 25 minutes
15.2	4.9%	from 25 minutes and less than 30 minutes
14.8	5.2%	30 minutes or more

The issue that was not clear from this initial observation was the degree of causality between time taken to respond to the test and the raw score achieved (the indicator of relative ability). Specifically, did the higher-ability students take longer overall because they were able to apply the skills, knowledge and demands of the items that caused them to take longer to complete, or did they simply take longer on the items due to a different level of engagement than the lower-ability students?

However, as an initial indicator, the consistency between ability (raw score) and response time suggested that these relationships functioned similarly and may have complemented each other. Consequently, investigation of how response time might assist in the indication of guessed responses was undertaken to determine if the time parameter could enhance the GIP procedure in improving the quality of the scale and increasing the accuracy of the protocol in indicating probable guessed responses.

8.4 PfA Step 2 Results: Relationship Between Item Location and Response Time by Ability Group

Approximately a quarter of cohort completed the mathematics tests in less than 10 minutes – a quarter of the time available for the test – and these students achieved raw scores of less than 30% of the possible score. Since there was no penalty for guessing, these students may have simply answered some questions quickly, which tends to obfuscate any systematic analysis of student/item response times. In the Grade 4 Mathematics assessment, the lower-ability students took an average of six minutes to complete the MC items in the test, whereas the mid-range ability group took nine minutes, and the higher-ability group took over 12 minutes (approximately double the time of the lower-ability students).

Table 8.4 shows the relationships between the average time taken to respond to each MC item and the difficulty location of the items sorted from least difficult to most difficult by the defined ability groups for each Grade. The lower-ability students (those in deciles 1 and 2) spent a shorter time on more items than the mid-range ability students (those in deciles 3 to 8) and the higher-ability students (those in deciles 9 and 10). It is important to note that in most cases, regardless of ability, the students completed all items.

In the Year 8 Mathematics test, the lower-ability students were again, on average, the quickest to complete the individual items. However, there was only a marginal difference between the mid-range and higher-ability groups in relation to individual item and overall test response time. In Year 8, the lower-ability students generally completed the test in about two-thirds of the time that the mid-range and higher-ability students took. Although there was some evidence of non-attempted items, relatively few students omitted items in this test.

The Grade 4 Science test showed similar ratios of response time across ability levels. However, the lower-ability group completed the test in less than half the time of the mid-range ability group, and it was in the latter items (more difficult by assessment design) that this lack of engagement was most apparent. In the case of Science, the tendency towards rapid responses may be a function of multiple variables such as item difficulty, item context, language complexity, or simple disengagement with the test as it became too difficult.

Table 8.4

Average Time Taken per MC Item by Ability Groups for Grades 4 and 8 Mathematics and Grade 4 Science Students

Grade 4 Mathematics

Group	INIT_SS	Blanks	M4Q01	M4Q02	M4Q03	M4Q04	M4Q05	M4Q06	M4Q07	M4Q08	M4Q09	M4Q10	M4Q11	M4Q13	M4Q14	M4Q16	M4Q19	M4Q22	M4Q23	M4Q24	Total
Upper-ability	16.4	0.2	0:16	0:56	0:32	0:43	0:48	0:35	0:24	0:56	0:23	0:35	0:25	1:00	0:28	0:35	0:36	0:52	0:53	1:34	12:40
Mid-ability	8.0	0.3	0:19	0:41	0:25	0:36	0:39	0:29	0:22	0:45	0:27	0:26	0:22	0:38	0:21	0:28	0:27	0:29	0:28	0:41	9:14
Lower-ability	3.0	0.9	0:18	0:26	0:18	0:24	0:26	0:20	0:19	0:28	0:21	0:17	0:18	0:21	0:16	0:22	0:20	0:19	0:17	0:22	6:21
ALL	8.6	0.4	0:18	0:42	0:25	0:35	0:39	0:29	0:22	0:44	0:25	0:26	0:22	0:40	0:22	0:29	0:28	0:33	0:32	0:50	9:30

Grade 8 Mathematics

Group	INIT_SS	Blanks	M8Q01	M8Q03	M8Q04	M8Q05	M8Q06	M8Q08	M8Q09	M8Q10	M8Q11	M8Q12	M8Q13	M8Q14	M8Q15	M8Q16	M8Q17	M8Q19	M8Q20	M8Q21	M8Q22	M8Q23	M8Q24	M8Q25	Total
Upper-ability	17.5	0.5	0:35	0:24	0:57	0:24	0:34	0:26	0:27	0:31	0:41	0:39	0:38	0:32	0:34	0:36	0:38	0:35	0:39	0:38	0:44	0:29	0:46	0:27	15:09
Mid-ability	9.2	0.5	0:39	0:23	1:00	0:25	0:34	0:27	0:29	0:32	0:44	0:39	0:38	0:34	0:38	0:40	0:43	0:40	0:45	0:39	0:50	0:32	0:51	0:29	16:08
Lower-ability	3.8	0.0	0:35	0:20	0:25	0:09	0:14	0:16	1:03	1:06	0:09	0:12	0:15	0:08	0:13	0:29	0:16	0:02	0:07	0:13	0:08	0:10	0:10	0:23	10:14
ALL	9.8	0.5	0:38	0:23	1:00	0:25	0:35	0:27	0:28	0:32	0:44	0:40	0:38	0:34	0:38	0:40	0:43	0:39	0:46	0:40	0:50	0:32	0:51	0:29	16:14

Grade 4 Science

Group	INIT_SS	Blanks	S4Q01	S4Q02	S4Q03	S4Q04	S4Q05	S4Q06	S4Q07	S4Q08	S4Q09	S4Q11	S4Q12	S4Q13	S4Q14	S4Q15	S4Q16	S4Q17	S4Q18	S4Q19	S4Q20	S4Q21	S4Q22	S4Q24	S4Q25	Total
Upper-ability	20.2	0.1	0:12	0:09	0:13	0:21	0:18	0:25	0:27	0:25	0:36	0:33	0:29	0:44	0:42	0:30	0:18	0:34	0:28	0:35	0:58	1:05	0:52	1:00	0:52	12:53
Mid-ability	11.6	0.2	0:14	0:13	0:17	0:22	0:19	0:26	0:26	0:23	0:24	0:25	0:28	0:30	0:35	0:24	0:17	0:23	0:25	0:24	0:32	0:39	0:28	0:33	0:28	9:42
Lower-ability	4.9	0.7	0:10	0:10	0:14	0:13	0:11	0:15	0:13	0:11	0:10	0:10	0:14	0:12	0:19	0:13	0:11	0:10	0:11	0:10	0:13	0:16	0:10	0:14	0:12	4:48
ALL	11.9	0.3	0:13	0:11	0:16	0:20	0:18	0:24	0:24	0:21	0:24	0:24	0:25	0:29	0:33	0:23	0:16	0:23	0:23	0:23	0:34	0:39	0:29	0:34	0:30	9:21

Note 1. Blanks records the average number of non-responses/omits to items in the test for each group.

Note 2. The pattern for Grade 4 was consistent between Mathematics and Science, with the lower-ability students recording more non-responses, on average, than the higher-ability students. However, given the rotated item presentation design, there was no consistent pattern of omission of the later items in any student's test, although there was some evidence of harder items, which were presented later in the test, being omitted.

8.5 Investigation of Response Time as a Unique Indicative Guessing Parameter

8.5.1 Frequency of Rapid Responses by Criterion Investigated

A large number of student responses would have been considered a probable guess if the defined student/item response time constraint had been applied as the unique indicator to indicate a guess (Table 8.5). The information in this sub-section demonstrates the variability that would have been introduced if the item response time had been used as a single variable to indicate guessing in the large-scale assessments of this study.

As expected, when the differing time constraints were applied to defining a rapid response (T/2, half the average time; T/3, one-third of the average time; T/4, one-quarter of the average time), fewer cases were indicated as possible complements of the GIP process. The T/4 condition indicated fewer cases than either the T/2 or the T/3 conditions.

In Year 4 Mathematics, the impact of applying the constraint T/2 resulted in 187,486 items (40% of the total responses (see Table 8.5)), were indicated as possibly guessed, irrespective of the student's success on the item. Of these rapid responses, approximately one-third were correct responses that could be considered probable guesses due to the rapid-response criteria of T/2. When the constraint T/3 was applied, the number of items identified as a rapid response reduced to 139,200 (30% of the total responses) and 46,810 (33.6% of rapid responses). In each case these responses had been correctly answered by the student.

As indicated in Table 8.5, there were a number of cases in which a student/item interaction was identified as a rapid response under the criteria investigated. The application of this variable as the sole indicator was likely to result a significant number of "false positive" outcomes when, for example, a higher-ability student might respond correctly quickly to an item within their knowledge/skill region.

Table 8.5

Count of Rapid Responses Potentially Re-Coded as Guesses by the Defined Time Constraint

Data Set	Count of items for defined for each time constraint only					
	TIME /2 (T/2)		TIME /3 (T/3)		TIME /4 (T/4)	
	Items identified	Correct responses identified	Items identified	Correct responses identified	Items identified	Correct responses identified
Grade 4 Mathematics	187,486 (40%)	63,676 (14%)	139,200 (30%)	46,810 (10%)	114,740 (25%)	38,790 (8%)
Grade 8 Mathematics	196,107 (42%)	97,115 (21%)	162,740 (35%)	72,575 (16%)	143,804 (31%)	68,411 (15%)
Grade 4 Science	281,228 (47%)	126,190 (21%)	209,352 (35%)	90,385 (15%)	167,841 (28%)	71,590 (12%)

Note. Table 8.5 shows the number of identified rapid-response items and the number of correctly answered rapid-response items (indicated guesses) when the response time was compared to the mean item response time of the mid-range ability group for each respective sample.

In Table 8.5, if the response time was a single indicator of a probable guess, almost half of those items identified as a rapid response with the T/2 constraint would be correct responses re-coded as a guess irrespective of the student ability. While response time alone was thus likely to be an ineffective parameter to identify guesses, the response times were investigated in conjunction with the GIP outcomes of Chapter 7 to determine if they would prove useful as an additional parameter to increase the confidence in the indication of the guess using the GIP process.

8.5.2 Consideration of the Relationship Between Item Difficulty and Rapid Responses

The information so far presented in this chapter relates to the observed response times of groups of students on the test overall and in an aggregated form by item (see Table 8.4). However, the issue of concern is not simply the count of rapid responses by student, but the degree to which the observation of a rapid response is an indicator of a probable guess. In other words, are the observed relationships between the variables of interest, a rapid-response time and a probable guess, causal?

Table 8.6 shows the occurrence of rapid responses with the T/2 criterion for Grade 4 Mathematics disaggregated by decile group. Overall, 39.6% of responses would be considered rapid. The average number of rapid responses was in the range of 31% to 45%, but the number of observed rapid responses was relatively independent of item difficulty. In the higher ability deciles (deciles 9 and 10), the easiest items have the most frequent rapid responses, and as items become more difficult the rates of rapid responses stabilise at about 20%. However, in the lower ability ranges (deciles 1 and 2), the rates of rapid response stabilise at approximately 60% to 70%, irrespective of item difficulty. In the mid-range ability deciles (deciles 3 to 8), there is again no clear relationship between item difficulty and rapid responses.

Given that it is assumed that items of higher difficulty would require more effort and typically require a longer time to respond, the data in Table 8.6 suggest that this is not the case with difficulty having little bearing on the rate of rapid responses observed within decile groups. These observations appear to negate the hypothesis that rapid response time is necessarily a function of item difficulty.

Table 8.6
Grade 4 Mathematics Rapid Response Proportions by Decile

Item	Decile	Q01	Q09	Q07	Q03	Q08	Q04	Q11	Q06	Q10	Q02	Q23	Q05	Q22	Q16	Q24	Q19	Q14	Q13
δ		-1.57	-1.18	-1.14	-0.7	-0.53	-0.52	-0.3	0.13	0.24	0.29	0.31	0.38	0.4	0.46	0.52	0.96	1.09	1.17
n(T/2)	Decile1	1250	1361	1213	1633	1688	1696	1291	1542	1701	1745	1922	1558	1885	1184	2056	1498	1691	1920
% T/2	%Dec1	47.60%	51.80%	46.20%	62.10%	64.20%	64.50%	49.10%	58.70%	64.70%	66.40%	73.10%	59.30%	71.70%	45.10%	78.20%	57.00%	64.30%	73.10%
n(T/2)	Decile2	1056	1224	1016	1420	1520	1594	1142	1357	1526	1545	1847	1380	1767	1094	1950	1330	1503	1808
% T/2	%Dec2	40.20%	46.60%	38.70%	54.00%	57.80%	60.70%	43.50%	51.60%	58.10%	58.80%	70.30%	52.50%	67.20%	41.60%	74.20%	50.60%	57.20%	68.80%
n(T/2)	Decile3	1045	1156	1023	1390	1511	1500	1120	1315	1494	1502	1764	1372	1714	1068	1867	1313	1425	1699
% T/2	%Dec3	39.80%	44.00%	38.90%	52.90%	57.50%	57.10%	42.60%	50.00%	56.80%	57.20%	67.10%	52.20%	65.20%	40.60%	71.00%	50.00%	54.20%	64.60%
n(T/2)	Decile4	896	1043	948	1281	1360	1404	1034	1185	1314	1381	1635	1283	1602	1004	1784	1239	1276	1523
% T/2	%Dec4	34.10%	39.70%	36.10%	48.70%	51.80%	53.40%	39.30%	45.10%	50.00%	52.50%	62.20%	48.80%	61.00%	38.20%	67.90%	47.10%	48.60%	58.00%
n(T/2)	Decile5	817	926	861	1106	1180	1242	965	1030	1161	1199	1456	1112	1454	919	1650	1140	1109	1357
% T/2	%Dec5	31.10%	35.20%	32.80%	42.10%	44.90%	47.30%	36.70%	39.20%	44.20%	45.60%	55.40%	42.30%	55.30%	35.00%	62.80%	43.40%	42.20%	51.60%
n(T/2)	Decile6	708	742	733	910	953	965	824	789	972	939	1169	869	1131	758	1378	898	872	1053
% T/2	%Dec6	26.90%	28.20%	27.90%	34.60%	36.30%	36.70%	31.40%	30.00%	37.00%	35.70%	44.50%	33.10%	43.00%	28.80%	52.40%	34.20%	33.20%	40.10%
n(T/2)	Decile7	696	704	630	725	762	732	751	659	798	764	915	701	915	577	1084	702	667	866
% T/2	%Dec7	26.50%	26.80%	24.00%	27.60%	29.00%	27.90%	28.60%	25.10%	30.40%	29.10%	34.80%	26.70%	34.80%	22.00%	41.20%	26.70%	25.40%	33.00%
n(T/2)	Decile8	674	678	564	587	544	551	640	488	648	500	625	512	638	469	831	531	531	641
% T/2	%Dec8	25.60%	25.80%	21.50%	22.30%	20.70%	21.00%	24.40%	18.60%	24.70%	19.00%	23.80%	19.50%	24.30%	17.80%	31.60%	20.20%	20.20%	24.40%
n(T/2)	Decile9	842	835	605	529	516	514	676	379	617	465	464	527	497	401	632	445	477	531
% T/2	%Dec9	32.00%	31.80%	23.00%	20.10%	19.60%	19.60%	25.70%	14.40%	23.50%	17.70%	17.70%	20.10%	18.90%	15.30%	24.00%	16.90%	18.20%	20.20%
n(T/2)	Decile10	1153	1185	740	738	663	644	862	560	822	612	502	737	522	502	612	551	637	641
% T/2	%Dec10	43.90%	45.10%	28.20%	28.10%	25.20%	24.50%	32.80%	21.30%	31.30%	23.30%	19.10%	28.00%	19.90%	19.10%	23.30%	21.00%	24.20%	24.40%
n(Total)	187476	9140	9857	8336	10323	10701	10846	9308	9307	11057	10656	12303	10055	12129	7979	13849	9650	10192	12043
% Overall	39.60%	34.80%	37.50%	31.70%	39.30%	40.70%	41.30%	35.40%	35.40%	42.10%	40.60%	46.80%	38.30%	46.20%	30.40%	52.70%	36.70%	38.80%	45.80%

Note. Items have been sorted from least to most difficult based on INIT analysis. Each quartile consists of 2628 students.

8.5.3 Consideration of the Relationship Between Rapid Responses and GIP Indicated Items

The lack of a causal relationship between responses and response time is further demonstrated in Table 8.7. The table shows combinations of correct responses that are (a) rapid responses, and (b) incorrect responses that are rapid responses, and (c) instances of GIP-indicated responses that are rapid responses. However, there is no consistent pattern. This lack of a pattern is also obfuscated by the number of non-rapid correct and incorrect responses in the sample of student data provided.

Table 8.7 has items sorted horizontally by INIT item difficulty from easiest to hardest. A random selection of students (Stud) from a sample of decile groups has been drawn to exemplify particular characteristics.

The defined average time for each item is noted in row 3. Two sets of statistics are provided for each student. The row 'Time' indicates whether the time for that student/item interaction was rapid, and the row 'Score' the result of the student/item interaction (0/1) conditioned in some cases by the GIP indicated items (annotated as code '7').

The student ability is noted together with the number of rapid responses observed and the number of GIP-indicated items indicated in the final three columns.

Student 10 is used to exemplify the structure of the Table. Student 10 has nine instances of rapid response when the student's response time is compared to the calculated T/2 average time. Student 10 has not recorded a response to items Q22, Q16, Q19 and Q14. However, given that the student has attempted Q23 it is assumed that the non-response was a 'considered' action, not a case of student 10 being unable to access all the items in the test.

In inspecting the student response pattern, Q01 was a correct response that was not a rapid response. Q07 was a correct answer that was a rapid response. Q08 was a correct response that was indicated by GIP as a probable guess (7) but was not a rapid response. Whilst Q23 was a correct response that was also indicated as a probable guess (7) by the GIP procedure and was also a rapid response. Hence Student 10 provides examples of all possible combinations in the matrix of correct responses with rapid responses.

Of note is Student 51, who has all 18 items considered as rapid responses. In two instances the responses are not indicated as probable guesses (GIP) and in three instances the responses are GIP indicated. By comparison, Student 10 has one instance of a non-rapid response indicated as a probable guess (GIP) and a second instance of a rapid response that is a GIP-indicated probable guess. These inconsistencies mitigate against the inclusion of time as a complementary indicator of likely guesses in the GIP procedure.

Table 8.7

Extract of Response Time (T/2), Response, and GIP-Indicated Items for a Random Selection of Students

	Item	Q01	Q09	Q07	Q03	Q08	Q04	Q11	Q06	Q10	Q02	Q23	Q05	Q22	Q16	Q24	Q19	Q14	Q13	Maths INITab	GIP change	T/2 Rapid
	δ	-1.565	-1.18	-1.135	-0.7	-0.533	-0.523	-0.296	0.129	0.237	0.287	0.309	0.379	0.395	0.461	0.519	0.955	1.09	1.17			
Av. Time	Stud	0:18	0:25	0:22	0:25	0:44	0:35	0:22	0:29	0:26	0:42	0:32	0:39	0:33	0:29	0:50	0:28	0:22	0:40			
Time	8	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	-2.78	1
Score	8	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	-2.78	1	
Time	10	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	-2.35	9
Score	10	1	0	1	0	7	0	0	0	0	0	7	0	#NULL!	#NULL!	0	#NULL!	#NULL!	0	-2.35	2	
Time	13	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	-1.72	3
Score	13	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	7	0	0	-1.72	1	
Time	51	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	-1.45	18
Score	51	0	0	0	0	0	0	1	1	0	7	0	0	0	7	0	7	0	0	-1.45	3	
Time	9	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	-1.22	3
Score	9	1	1	1	0	0	0	1	1	0	1	0	0	1	0	0	0	7	0	-1.22	1	
Time	14	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	-0.99	18
Score	14	1	1	1	0	1	0	0	0	1	1	0	0	0	0	1	0	0	7	-0.99	1	
Time	2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	-0.78	3
Score	2	1	1	0	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0	-0.78	0	
Time	5	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	-0.38	2
Score	5	1	1	1	1	0	1	0	0	1	1	0	1	1	0	0	0	0	0	-0.38	0	

Note. An annotation of time as “false” indicates that this was not a rapid response for that item/student interaction. An annotation of “true” indicates that that student/item interaction was a rapid response.

8.6 Results of Analyses of Student Responses Times

The initial observations of the relationship between overall test response times and student ability detailed in Sections 8.3 and 8.4 indicated that there may be relationships between student ability, item difficulty and student/item response time that could be leveraged to develop a more reliable and accurate GIP. However, further investigation of these relationships revealed inconsistencies in that there were cases where a rapid response time was observed with a correct response for multiple students with higher ability estimates and relative slow response time for multiple students of lower ability to items which the GIP procedure suggested were likely guesses.

These inconsistencies were found in each subject and grade with specific examples of the issue made explicit in Tables 8.6 and 8.7. The lack of a consistent relationship between student ability, item difficulty and student/item response time made inclusion of response time as a variable in the GIP process problematic as discussed below.

8.7 Discussion of the Outcomes of PfA Step 1 and Step 2

In the review of item response rate, two relationships were apparent from the summary data (see Table 8.4):

1. Lower-ability students generally responded to each item in the test in the fastest time; and
2. Higher-ability students consistently responded to the MC items in the test in the longest time.

These observations are not considered causal, and may have been influenced by two opposing factors:

1. The relative inability of lower-ability students to interact with the items engendered guessing as the likely response action, with quick response time reflecting a ‘no-knowledge’ reaction; and
2. The time taken by the higher-ability students, for whom the item was within their range of knowledge, skill, and ability, reflected the time taken to read and decode the question and consider the best answer.

Considering that the lower-ability students had lower success rates on the more difficult items overall, the observation that they tended to respond more rapidly to those items that were too difficult for their ability level supports the assumption that these responses were likely to have been guessed.

However, it is worth noting that the lower-ability students generally took no longer to respond to the items requiring higher cognitive demand than to those requiring lower cognitive demand. There was increasing interaction time for the higher-ability students on the items as cognitive demand and item difficulty increased. In considering these interactions, these data support the principal hypothesis regarding the relationship between student ability, item difficulty, and guessing. However, the potential for a time variable in the GIP procedure is not supported, as there was no consistent not apparent causal relationship between item difficulty and the time taken to respond to items as difficulty increased across the full range of student ability estimates.

8.8 Conclusion Regarding the Student/Item Response Time as an Additional Parameter in the GIP Algorithm

This chapter has shown that, at least for some cases, there was a likely relationship between response time and possible guessing. Some researchers have linked this to test-taking effort (Guo et al., 2016, Michaelides et al., 2020; Setzer et al., 2013). However, this analysis illustrates that a rapid response may indicate knowledge of the item demand and quick identification of the correct option. Indeed, the inconsistency of the time-only constraint across all groups of students suggests that there was no consistent response time ratio that can provide a reliable indicator of guessing.

Given the inability of response time to be a reliable indicator of guessing and the diminishing returns in relation to the confirmation of the GIP indicators as a result of including time as an additional variable, the student/item response time parameter was ultimately not used in conjunction with the previously defined GIP process to indicate probable guessing.

Chapter 9

Reports of Student Achievement

Impact of the Guessing Indication Protocol (GIP) on Reported Student Performance

9.1 Introduction

The outcomes reported in each of Chapters 5, 6, and 7 indicate a significant difference between the individual ability estimates of the higher and lower-ability student groups in analyses that account for probable guessing in MC items using the defined Guessing Indication Protocol (GIP) procedures compared to the outcomes when no processes were implemented to account for probable guessing. This chapter discusses the potential impact of guessing on the reporting of student performance.

The increase in the standard deviations of the distribution of student ability estimates is considerable in each of the analyses of Chapter 7. It is appreciated that these increases, with marginal increases in the reliability coefficient, increased the measurement error in relation to each estimate by a significant factor. Consequently, the confidence interval about each student's ability estimate also increases in absolute terms. However, the Level in which a student's performance is reported, discussed below, is calculated on the point estimate of student ability and does not take the confidence interval into consideration.

Section 9.1 explains the construct of scaled scores and how they relate to the assignment of Levels in student achievement. The differences in scaled scores observed for the different analysis Phases is a fundamental aspect of the discussion regarding the final reports of student assessments.

Sections 9.2, 9.3 and 9.4 provide detail of the manner in which the GIP outcomes differ from the INIT outcomes in relation to the assignment of Levels for each of respective Studies provided in Chapters 5, 6 and 7. In particular it shows the degree to which the reported Levels change when the GS and GIP analyses are conducted.

Section 9.5 discusses the degree of misclassification of students in Levels when no account is taken for probable guessing in the Rasch analysis. The discussion highlights the potential scale of misinformation provided to stakeholders when a simple INIT analysis provides data upon which decisions are made.

9.1.1 Achievement Standards in Australia

In the Australian context, student achievement is guided by the aims and content of the Australian Curriculum (<https://australiancurriculum.edu.au>). The Curriculum articulates the levels of knowledge, understanding, and skills that are expected to be achieved at various stages of a student's educational journey. Figure 9.1 is an extract from the Australian Curriculum that provides a framework for the expected learning outcomes to be achieved by the end of Year 4 in Mathematics. It is against these Achievement Standards that summative reports are prepared to indicate progress of individual students against the curriculum expectations.

Figure 9.1

Extract From the Australian Curriculum Year 4 Mathematics Achievement Standards

Year 4 Achievement Standards

By the end of Year 4, students choose appropriate strategies for calculations involving multiplication and division. They recognise common equivalent fractions in familiar contexts and make connections between fraction and decimal notations up to two decimal places. Students solve simple purchasing problems. They identify and explain strategies for finding unknown quantities in number sentences. They describe number patterns resulting from multiplication. Students compare areas of regular and irregular shapes using informal units. They solve problems involving time duration. They interpret information contained in maps. Students identify dependent and independent events. They describe different methods for data collection and representation, and evaluate their effectiveness.

Students use the properties of odd and even numbers. They recall multiplication facts to 10 x 10 and related division facts. Students locate familiar fractions on a number line. They continue number sequences involving multiples of single digit numbers. Students use scaled instruments to measure temperatures, lengths, shapes and objects. They convert between units of time. Students create symmetrical shapes and patterns. They classify angles in relation to a right angle. Students list the probabilities of everyday events. They construct data displays from given or collected data.

Source. <https://www.australiancurriculum.edu.au/f-10-curriculum/mathematics/>

9.1.2 Reported Levels and the Link to Achievement Standards

This section provides examples of reports and standards statements to illustrate the context for the discussion of possible misinformation from data that have not accounted for guessing. Figure 9.2 is an extract of a Mathematics Standards Scale (Abu Dhabi Educational Council, EMSA scale, 2018) that shows the link between student scaled scores (the outcomes of the analyses shown in section 9.1.3) and the described achievement standards.

Figure 9.2

An Example of a Mathematics Standards Framework (Grade 3 to Grade 6)

Level Label	Scaled Score	Level Descriptor
Numeric Range		Learners at this level can typically: -
Cycle 1 Level 7 and above (max for G4 and G5)	Above 801	Use a probability tree diagram. Understand stem and leaf plots. Calculate averages. Solve simple permutation problems. Solve contextual percentage problems. Calculate values of investments. Solve simple rate problems. Identify properties of complex 2-D shapes. Solve problems involving volume, area and length. Find properties of graphs including domain, range, gradients, intercepts & continuity. Apply differentiation rules. Evaluate simple integrals. Identify a term in a geometric sequence.
Cycle 1 Level 6 (max for Grade 3)	701 – 800	Work with positive and negative integers. Solve problems involving multiplication, division and rounding. Round numbers to significant figures and convert from scientific to decimal notation. Solve measurement problems involving area and perimeter. Use Pythagoras' theorem and identify equivalent trigonometric values. Solve spatial reasoning problems involving visualisation and addition. Evaluate a formula to solve a problem. Transform an algebraic expression using standard procedures, including differentiating simple functions. Sum a finite arithmetic series.
Cycle 1 Level 5	601 - 700	Use four operations to solve simple problems and relate processes. Identify numbers and fractions on number lines and work with fractions and simple percentages. Continue sequences in number patterns, identify missing values in increasing patterns and relate shape patterns to value tables. Interpret graphical displays and relate frequency numbers in tabled data. Work with trigonometric ratios and angle properties. Identifies the properties of 3D objects and follow position language on grids. Evaluate a simple function of one variable for a given value.
Cycle 1 Level 4	501 - 600	Identify and use simple fractions, decimals and percentages. Solve addition, subtraction and multiplication and work with square and cubed roots. Identify common multiples of simple factors and understand 4-digit numbers. Select minute equivalent of part of hour. Identify combined shape from given shapes.
Cycle 1 Level 3	401 - 500	Understand place value to three digits and order numbers on number lines. Identify value equivalence of different coins and ordinal position. Continue simple shape patterns and number patterns. Match tabled data with graph representation, and match analogue and digital times. Identify cylindrical shapes and follow simple positional language.
Cycle 1 Level 2	301 - 400	Solve simple word problems and identify missing terms in ascending and descending patterns. Select appropriate instrument to measure length. Understand tally representation and interpret column graphs. Identify familiar 3D objects.
Cycle 1 Level 1	201 - 300	Perform simple operations including counting a small collection, adding, subtracting and multiplying single digits. Identify key times on an analogue clock and identify area in units.

Note 1. The Level for a student's performance in the assessments is operationalised by the assignment of a score that indicate achievement within each level. The ranges that relate to each level are defined in the column Numeric Range.

Note 2. The first column shows the possible levels that can be achieved by students who participate in the test. The second column represents the possible range of standardised scale scores that define a Level, and the third column describes the skills and contexts of the test for items within each range of difficulty.

Standards, expressed as Level descriptors, provide meaning to the numeric scaled score. The Level descriptors are developed by aggregating the task demands of items that fall within defined difficulty ranges into meaningful text. The text typically describes the skills demonstrated within the defined range. Figure 9.2 shows seven levels that transcend the learning outcomes expected of students in Grades 3 through 5 (termed “Cycle 1” in this UAE ²context). The purpose of this scale is to allow students to be located on a scale, constant over time, and provide a stable standard against which progress may be measured. The descriptions provided are derived from the types and contexts of items that are observed within the score ranges indicated in numeric range column. A student who falls within a particular level will have shown evidence of the skills in the test(s) from which this report is generated.

Overall, the scale attempts to capture the expected achievement of students of the target cohort and describe the continuum of the domain as learning advances through the Grades/Year levels. The report is divided into levels, which describe the observed achievement of the students as evidenced by their performance in the assessment. The detail articulated in Figure 9.2 captures the principle that, although the curriculum describes an achievement standard for the Year/Grade, not all students progress at the same rate. Hence, the report allows for a range of achievements, from students who are well below the expected level to students who are exceeding the expected outcomes of the Year/Grade. The Levels described in Figure 9.2 reflect the observed achievement and the skills typically demonstrated by a student who has achieved a specific level.

In Figure 9.2 the “width” of each Level is 100 scaled score points. The derivation of the level width for a particular domain could be considered “arbitrary” as the level³ width will vary for different domains depending upon several factors, including (but not limited to);

- the capacity to unambiguously describe the attributes that determine a level within the scale;
- the number of bands (levels) that can be unambiguously described;
- the range of Year/Grade levels that the scale is referring to (NAPLAN has 10 bands (levels) encompassing 7 years of learning, as shown in Figure 9.3; and
- the range of the distribution of item locations in the original implementation of the assessment from which the scale and descriptors have been derived.

As demonstrated in Figure 9.2, the alignment of student achievement, typically represented as a scaled score, with the standards statements provides a meaningful description of the student score as a manifestation of learning evidenced in the assessments. The information displayed in the Standards Scale provides a framework for the targeting of interventions and comparisons to future outcomes that evidence student progression towards increasing standards of learning in the domain of interest.

² The large-scale data were derived from the Abu Dhabi Education Council’s Semester 1 test of the 2019 Question a Day program

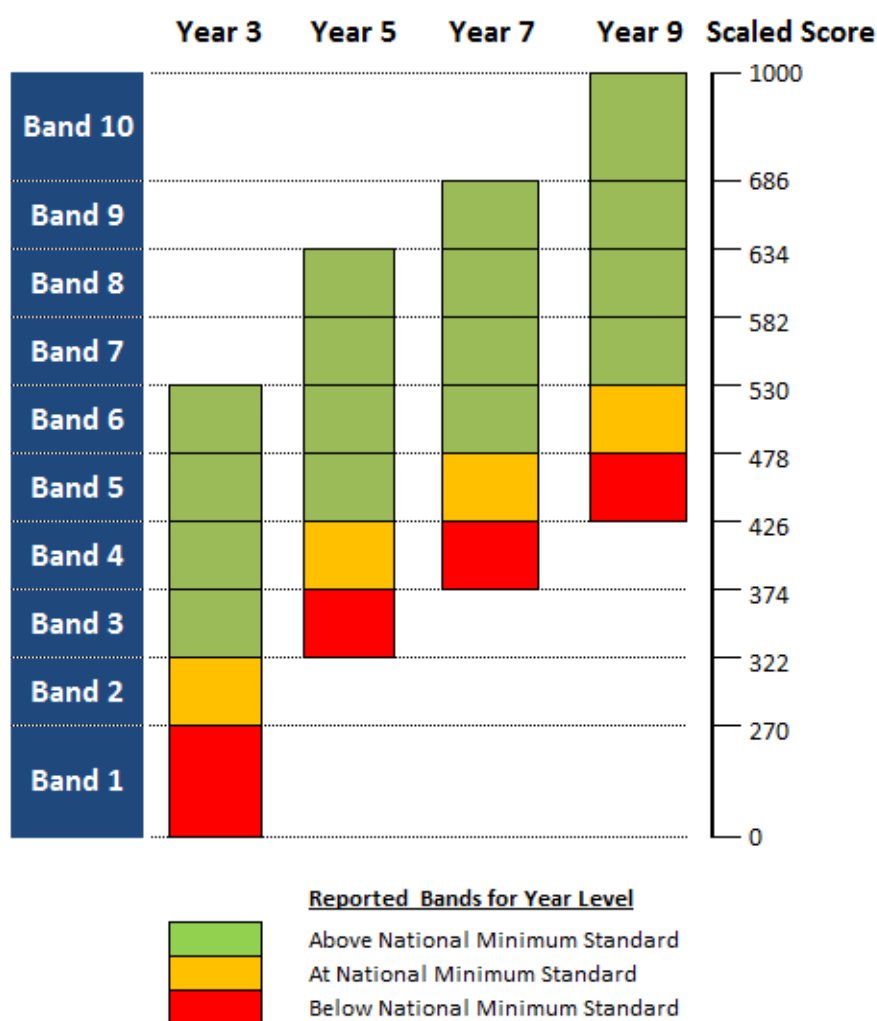
³ The terms “band” and “level” are equivalent; they are used alternatively by different jurisdictions.

9.1.3 Scaled Scores and the Link to Levels

The Victorian Curriculum and Assessment Authority (VCAA) have published explanatory notes (<http://usingassessmentdata.vcaa.vic.edu.au/naplan/tut1>) that relate the NAPLAN Reading scaled scores to the NAPLAN Band (Level) descriptors, as shown in Figure 9.3. The figure highlights the one-to-one relationship between scaled scores, the band cut scores, and the band descriptors. It also shows that the level width between successive bands for this subject is 52 scaled score points. This represents one standard deviation in the Reading Scale. The scaled score value of a level width varies for different subjects, as detailed in the NAPLAN Technical Report (2019).

Figure 9.3

NAPLAN Reported Bands (Reading) and Scaled Score Cut Scores



Source: http://usingassessmentdata.vcaa.vic.edu.au/naplan/tut1_1/mod1.aspx#

In the modern educational environment of large-scale standardised assessments, scaled scores are derived from analyses such as the Rasch to provide numeric values to the student outcome that are more user friendly than the Rasch logit value. Figure 9.2 shows a scale in which the standardisation process involved fixing the mean to a scaled score 500 and the standard deviation to a scaled score of 100.

The use of mathematical equating procedures between Years and across time (Andrich, 2008; Linn, 1993; Mislevy, 1992; NAPLAN Technical Reports, 2008 to 2019) enables comparison of equivalent cohorts to measure identifiable improvement in the system, and/or the tracking of individual students to measure growth in learning attainment. The process used to determine the standardised scaled scores is the common procedure defined in Eqn 9.1.

$$X_{SS} = \left\{ \frac{[X_j - \bar{X}]}{SD_j} \right\} \times SDD + \bar{X}d \quad \text{Eqn 9.1}$$

Where X_{SS} is the calculated Standardised Scaled Score;

x_j is the achieved score of person j ;

\bar{X} is the mean of the achieved scores of the cohort;

SD_j is the standard deviation of the achieved scores;

SDD is the defined standard deviation of the standardised scaled scores; and

$\bar{X}d$ is the defined mean of the standardised scaled scores.

The parameters used to standardise the ability estimates calculated for the INIT phase outcomes are shown in Table 9.1. Throughout this current research, the standardisation of the logit values (Chapters 5, 6, and 7) used a mean of 500 and a standard deviation of 100 to create a scaled score in order to facilitate comparisons.

Table 9.1

Parameters Determined for Calculation of INIT Standardised Scaled Scores

Data		Mean \bar{X}	SD_j	Source	$\bar{X}d$	SD_d
Study 1, Simulations	SIM 1	0.72	1.39	Table 5.27	500	100
	SIM 2	1.40	1.27	Table 5.27	500	100
	SIM 3	0.72	1.30	Table 5.27	500	100
	SIM 4	1.27	1.48	Table 5.27	500	100
	SIM 5	0.24	1.39	Table 5.27	500	100
Study 2, Small Scale	FT Y5	1.74	1.29	Table 6.15	500	100
	FT Y7	1.42	1.01	Table 6.16	500	100
Study 3, Large-scale	Maths G4	-0.40	1.10	Table 7.10	500	100
	Science G4	0.03	1.11	Table 7.11	500	100
	Maths G8	-0.44	0.94	Table 7.12	500	100

In this study, levels have been ascribed by reference to the standards described in Eqn 9.2

if $SS\{analysis\} > 700$ $\{analysis\}Level = 6.$

if $SS\{analysis\} < 701$ AND $SS\{analysis\} > 600$ $\{analysis\}Level = 5.$

if $SS\{analysis\} < 601$ AND $SS\{analysis\} > 500$ $\{analysis\}Level = 4.$

if $SS\{analysis\} < 501$ AND $SS\{analysis\} > 400$ $\{analysis\}Level = 3.$

if $SS\{analysis\} < 401$ AND $SS\{analysis\} > 300$ $\{analysis\}Level = 2.$

if $SS\{analysis\} < 301$ $SIM1Level = 1.$

Eqn 9.2

These Levels and the associated “cut scores” may be considered as arbitrary, but they are efficient as a basis for comparisons for determining the impact of the GIP conditioning of the response data for each analysis.

9.2 Study 1, Simulated Data – Reporting of Achievement Levels for Each Analysis Phase

Tables 9.2 to 9.6 provide comparisons of the relative “reassignment” of students to performance levels in the simulated data sets developed and analysed in this research. These tables present the three-phase calibrations of the response data for each of the five simulation data sets analysed. The purpose of each column will now be explained:

1. The column “{analysis} LevelINIT” describe the distribution outcomes of a simple Rasch analysis of the scored data without any accounting for guessing. The ability estimates are determined from Raw Score to Ability tables derived in the INIT analysis and then standardised applying the mean and standard deviation of the data for the particular data set (mean 500, st.d. 100).
2. The column “{analysis} LevelGS” describes the distribution outcomes of a Rasch analysis of the scored data with the defined guesses (or in the case of the field trial data, the self-nominated guesses) suppressed.
3. The column “{analysis} LevelGIP3A” describes the distribution outcomes of a Rasch analysis of the scored data conditioned to suppress responses that “fail” the GIP thresholds for indicating a guess. In these tables the GIP identified responses are scored as missing data. Hence the levels ascribed are based on a reduced raw score – the GIP3A raw score.
4. The column “{analysis} LevelGIPINIT3B” describes the distribution outcomes a Rasch analysis of the scored data conditioned to suppress responses that “fail” the GIP thresholds for indicating a guess. The item locations resolved by the GIP3A analysis are anchored in this analysis. The levels ascribed to students are based on the raw score determined in Phase 1 (the INIT Raw Score). This phase is the outcome of interest and termed the GIPINIT3B analysis in the discussion that follows.

The values shown in the column LevelINIT represent the relative percentages in each of the defined levels as a result of the Rasch analysis and the application of the standardisation process on the observed student ability estimates generated. By comparison, the column LevelGS shows the comparable percentages where the responses that have been predefined as guesses are suppressed and considered as “missing data” for both calibration and ability calculations for these data. In Tables 9.2 to 9.6, the lower ability levels are shaded.

9.2.1 Reported Achievement Levels for the SIM1 Data

The outcomes of the three phases of the data set Simulation 1 (SIM1) are displayed in Table 9.2. The differences in the scaled scores and levels for each phase reflect the re-calibration of item difficulties, the consequent raw score, and the ability estimate determined in each phase, as a result of the reclassification of the defined (GS) and GIP identified guesses as appropriate.

Table 9.2*Comparisons of Percentages in Each Level for Simulation 1 by Analysis Phase*

Level	SIM1 LevelINIT			SIM1 LevelGS			SIM1 LevelGIP3A _{p=0.6}			SIM1 LevelGIPINIT3B _{p=0.6}		
	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent
6	9	2.3	100.0	104	26.0	100.0	43	10.8	100.0	58	14.5	100.0
5	54	13.5	97.8	34	8.5	74.0	55	13.8	89.3	47	11.8	85.5
4	116	29.0	84.3	103	25.8	65.5	94	23.5	75.5	96	24.0	73.8
3	156	39.0	55.3	53	13.3	39.8	100	25.0	52.0	103	25.8	49.8
2	60	15.0	16.3	46	11.5	26.5	51	12.8	27.0	80	20.0	24.0
1	5	1.3	1.3	60	15.0	15.0	57	14.3	14.3	16	4.0	4.0
	400	100.0		400	100.0		400	100.0		400	100.0	

Comparing the INIT values to the GS values of SIM1, Table 9.2 shows that the percentage of students in the lowest ability groups (Levels 1 and 2) on this trait has been overestimated by about 10.2% (16.3% vs 26.5%, respectively) when the defined guessing has not been accounted for. This overestimation derives from the inflation in the raw scores achieved by the correct random guessing achieved by students. This percentage indicates that approximately 10% of the population who are performing below the expected achievement range would be reported as achieving in an acceptable range of achievement for the Grade. At the other extreme of the scale, the comparison of the frequencies between the INIT levels and the GS levels for the highest ability group (Level 6) were underestimated by almost 24%. The calibration effect of unaccounted-for guessed items was the reduction of the relative difficulty of harder items due to the impact of correct guesses from the lower-ability groups, with consequent impact on the ability determination for the set of items across the range of abilities. In the range between the extremes of the INIT and GS outcomes, there is variation in the observed proportion of students in each level. However, of interest is the degree to which students achieved the lowest two bands (INIT – 16.3% compared to GS – 26.5%) and the upper two bands (INIT – 15.8% compared to GS – 34.5%, respectively). These represent significant misclassifications of students for the purpose of both individual and cohort reporting.

9.2.2 Discussion – Achievement Levels for SIM1 Simulated Data

The focus of this thesis is to report what is feasible and practical to achieve with a process grounded in Rasch modelling that takes account of guessing. The analysis that satisfies these requirements is the GIPINIT3B outcomes. The application of the GIP provides a further comparison of the set of values shown in the column SIM1 LevelGIPINIT3B_{p=0.6} against the INIT analysis outcomes. As discussed in Chapter 5, the GIP process is less efficient in the identification of guessing in the responses of the higher-ability students due to the consequent positive probability of success on all but the most difficult items and the requirement that the item difficulty exceeds the ability estimate of the student by at least 1.1 logits. However, it has an important degree of efficiency in indicating probable guessing for the lower-ability students. This was evident by the comparison of the INIT (guessing not accounted for) outcomes compared to the GIPINIT3B outcomes for students achieving in the lower levels.

Although not as stark as the comparisons between the unconditioned data (column SIM1 LevelINIT) and the defined guessing data outcomes (column SIM1 LevelGS), the GIPINIT3B process resulted in the students in the lower two levels being reported at 24% compared to 16.3% in the LevelINIT, and the upper two levels at 26.3% for the LevelGIPINIT compared to 15.8% in the unconditioned INIT data. In large-scale data sets (25,000 candidates) this represents an overestimation of ability of around 1,900 students in the lower ability levels, who are the students most likely in need of greater support. In addition, there would be approximately 2,600 students at the higher end of the ability spectrum whose abilities, and consequent achievement levels, have been underestimated.

9.2.3 Reported Achievement Levels for the Simulated Data Sets

The outcomes of the three phases of the data set Simulation 2 (SIM2) are displayed in Table 9.3. In SIM2, the comparison of the GIPINIT3B analysis with the INIT outcomes shows relatively small variations in the percentages of students in each of the lower two levels. This may be an artefact of the small cohort sample and the relatively few items in this simulation, together with the lack of discrimination in item difficulty locations and student ability estimates. At the upper two levels a larger difference was observed.

Table 9.3

Comparisons of Percentages in Each Level by Analysis Phase

Level	SIM2 LevelINIT			SIM2 LevelGS			SIM2 LevelGIP3A $p=0.6$			SIM2 LevelGIPINIT3B $p=0.6$		
	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent
6	11	4.4	100.0	11	4.4	100.0	40	16.0	100.0	40	16.0	100.0
5	29	11.6	95.6	49	19.6	95.6	46	18.4	84.0	46	18.4	84.0
4	87	34.8	84.0	67	26.8	76.0	65	26.0	65.6	65	26.0	65.6
3	73	29.2	49.2	92	36.8	49.2	27	10.8	39.6	49	19.6	39.6
2	48	19.2	20.0	28	11.2	12.4	43	17.2	28.8	47	18.8	20.0
1	2	0.8	0.8	3	1.2	1.2	29	11.6	11.6	3	1.2	1.2
	250	100.0		250	100.0		250	100.0		250	100.0	

The outcomes of the three phases of the data set Simulation 3 (SIM3) are displayed in Table 9.4. By comparison with Table 9.3, Table 9.4, with both a larger cohort size, item numbers, and discrimination between item difficulty locations, displays substantial differences in the percentages of students in each level of the GIPINIT3B compared to the INIT analysis percentages. The changes reflect the increase in the distribution of outcomes when guessing in the data responses had been considered. When the GS outcomes are compared to the INIT outcomes, there is a considerable increase in the percentage of students in the lowest achievement level. This increase reflects the process of accounting for the defined random guessing in the lower-ability student groups. In the lower two levels there are almost twice the number of cases in the GS outcomes compared to the INIT outcomes. In the higher levels, the GS outcomes show about 25% in the highest two levels, while the INIT analysis shows about 16% in these levels. The GS outcomes reflect what would be the “true” situation if it were possible to identify all the guessed items.

In comparing the INIT outcomes with the GIPINIT3B outcomes there are marginal increases in the percentages of students in the lower two levels. The degree to which the GIPINIT3B outcomes are able to reflect the actual achievement of the students is impacted by the awarding of credit for those items indicated by the GIP process as guessed. This is made apparent when the GIP3A outcomes are observed. In comparing the GIP3A outcomes with the INIT outcomes there is an increase of approximately 12.6% in students indicated as performing at the lowest level when the GIP 3A outcomes are used as the indicator of achievement. At the higher ability end of the scale the GIP3A and GIPINIT3B analyses both report 23.6% in the highest two levels compared to 16.4% in the INIT analysis. This pattern of outcomes is consistent across all the simulation studies in the upper regions of achievement.

Table 9.4

Comparisons of Percentages in Each Level for Simulation 3 by Analysis Phase

Level	SIM3 LevelINIT			SIM3 LevelGS			SIM3 LevelGIP3A $p=0.6$			SIM3 LevelGIPINIT3B $p=0.6$		
	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent
6	18	1.8	100.0	167	16.7	100.0	71	7.1	100.0	71	7.1	100.0
5	146	14.6	98.2	88	8.8	83.3	165	16.5	92.9	165	16.5	92.9
4	273	27.3	83.6	79	7.9	74.5	202	20.2	76.4	254	25.4	76.4
3	402	40.2	56.3	193	19.3	66.6	270	27.0	56.2	332	33.2	51.0
2	145	14.5	16.1	198	19.8	47.3	150	15.0	29.2	144	14.4	17.8
1	16	1.6	1.6	275	27.5	27.5	142	14.2	14.2	34	3.4	3.4
	1000	100.0		1000	100.0		1000	100.0		1000	100.0	

The outcomes of the three phases of the data set Simulation 4 (SIM4) are displayed in Table 9.5. Table 9.5 shows a relatively high percentage of students classified in the middle two levels in the INIT analysis. This reflects the contraction of the distribution in the simple Rasch analysis of this simulation, in which the items are too easy for the population. By comparison, the GS analysis distributes the students relatively equally among the six levels. The GIPINIT3B analysis is similar in the percentages of students in each level when compared to the GS analysis, and they distribute students relatively equally among the levels, with the exception of Level 1. There is also a noticeable increase in the percentage of students in the higher two levels. This is an improvement over the Rasch analysis when considering that the GS analysis reflects the “true” values.

Table 9.5

Comparisons of Percentages in Each Level for Simulation 4 – Mismatched – Too Easy.

Level	SIM4 LevelINIT			SIM4 LevelGS			SIM4 LevelGIP.A $p=0.6$			SIM4 LevelGIPINIT3B $p=0.6$		
	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent
6	3	0.8	100.0	72	18.0	100.0	47	11.8	100.0	47	11.8	100.0
5	69	17.3	99.3	57	14.3	82.0	86	21.5	88.3	86	21.5	88.3
4	114	28.5	82.0	42	10.5	67.8	71	17.8	66.8	74	18.5	66.8
3	146	36.5	53.5	79	19.8	57.3	82	20.5	49.0	114	28.5	48.3
2	64	16.0	17.0	72	18.0	37.5	71	17.8	28.5	67	16.8	19.8
1	4	1.0	1.0	78	19.5	19.5	43	10.8	10.8	12	3.0	3.0
	400	100.0		400	100.0		400	100.0		400	100.0	

The outcomes of the three phases of the data set Simulation 5 (SIM5) are displayed in Table 9.6. Table 9.6 shows the outcomes of the simulation that was designed to be too hard for the cohort. The outcomes of these analyses produced some interesting comparisons. The GS analysis recalibrated the scale, having removed comparatively large numbers of defined guesses, and resulted in a very different distribution of ability outcomes. The GIP analysis identified seven items with no “correct” responses. These were identified as “extreme” cases and were removed from the redefined scale. Hence the outcomes of the GIPINIT3B are based on only 33 items. The GIPINIT3B analysis is similar to the GS analysis in indicating students in the lowest two levels and in the totals of the highest two levels. However, it is difficult to make valid comparisons given the different scale generated following the removal of the extreme response patterns. These outcomes may be significant in consideration of the impact of guessing in instances where tests are too hard for the target cohort.

Table 9.6

Comparisons of Percentages in Each Level for Simulation 5 – Mistargeted – Too Hard

Level	SIM5 LevelINIT			SIM5 LevelGS			SIM5 LevelGIP3A $p=0.6$			SIM5 LevelGIPINIT3B $p=0.6$		
	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent
6	2	0.5	100.0	106	26.5	100.0	2	0.5	100.0	52	13.0	100.0
5	75	18.8	99.5	34	8.5	73.5	75	18.8	99.5	87	21.8	87.0
4	125	31.3	80.8	26	6.5	65.0	103	25.8	80.8	41	10.3	65.3
3	126	31.5	49.5	43	10.8	58.5	76	19.0	55.0	44	11.0	55.0
2	64	16.0	18.0	44	11.0	47.8	86	21.5	36.0	59	14.8	44.0
1	8	2.0	2.0	147	36.8	36.8	58	14.5	14.5	117	29.3	29.3
	400	100.0		400	100.0		400	100.0		400	100.0	

9.2.4 Simulation Analyses – Scaled Scores

It is common practice in large-scale assessments for both described levels and a scaled score to be reported to various stakeholders. Table 9.7 shows the relative differences in the mean scaled score performances of each of four equal-number groups of candidates ($n = 100$) in the outcomes of the three analyses of the SIM1 data, together with the indicative “level” achieved by the student at the cusp cut-score of the quartile.

Table 9.7

Comparison of Mean Scaled Scores by Quartile for SIM1 Data

Quartile	SIM1 Mean SSINIT	SIM1 Mean SSGS	SIM1 Average Level INIT	SIM1 Av. Level GS	SIM1 Mean SSGIP3A $p=0.6$	SIM1 Av. Level GIP3A $p=0.6$	SIM1 Mean SS GIPINIT3B $p=0.6$	SIM1 Av. Level GIPINIT3B $p=0.6$	SIM1 Δ mean SS GIP3A $p=0.6$ with INIT	SIM1 Δ mean SS GIPINIT3B $p=0.6$ with INIT
Q4	636	864	4.7	6.0	685	5.4	708	5.6	49	72
Q3	521	587	3.8	4.4	544	3.9	551	4.1	23	30
Q2	470	486	3.0	3.3	459	2.9	472	3.0	-11	2
Q1	375	253	2.3	1.5	280	1.4	337	1.9	-95	-38
Average	500	546			491		516			
Std Dev	100	234			155		141			

Although it is appreciated that achievement levels are categorical values, an average level is presented to represent a measure of the progression “through” the level that the group achieved. This is consistent with reporting presentations such as NAPLAN, as represented in Figure 10.2. Hence a value of 4.7 represents a scaled score that indicates about three-quarters of the skills contained within Level 4 have been achieved or observed, whereas a value of 2.3 is interpreted as evidence of skills within Level 2 but in the early regions of the skills that are encompassed by that outcome.

Table 9.7 shows an anomaly in the outcomes that, at first glance, seems inconsistent with the findings in Chapters 5, 6, and 7. The mean standardised scaled score of the INIT analysis was 500 by definition. It was noted in Chapter 5 that the impact of the suppression of the defined guessing in the GS analysis reduced the mean ability estimates but increased the distribution of scores. Table 9.7 reflects this in the column SIM1 Mean SSGS, with the mean ability scaled score less than the INIT mean SS value of 500.

However, the mean of the calculated scaled scores for the column SIM1 Mean SSGIPINIT3B_{p=0.6} was 516, which is 16 scaled score points above the mean of the INIT analysis outcome. The reason for this apparent anomaly, and other instances that follow in the tables below, is that the analyses of the GIPINIT3B in Chapters 5, 6, and 7 are based on data that had been conditioned to take consideration of defined or probable guessing in the item calibration. In the calibration of the item difficulties and raw score to ability tables of the GS and GIP3A analyses, the achieved raw score of each student was reduced by the count of recoded item–person interactions determined as guesses. Although the mean score was reduced, a consequence of the guessing was that the distributions of scores increased. In the GIPINIT3B analysis, the INIT raw score data was reintroduced, which credits students for correct “guesses” and inflates the student raw score with consequent impact on the ability estimate and hence the mean statistic.

This process has been discussed and rationalised in Chapters 5, 6, and 7. In the calculation of the standardised scaled scores of both the GS and GIP3A procedures, the mean and the standard deviation of the INIT analysis were applied (Eqn 9.1) to generate the scaled score (and reflect the variations in the distributions shown in Tables 9.8 to 9.11. It follows that the use of these inflated raw scores would result in a higher mean than when raw scores were reduced to reflect the reclassification of probable guessed responses from a “1” to a missing value.

The increase in the distribution of GIPINIT3B ability estimates, translated into scaled scores, shows that the higher-ability students were underestimated, on average, by 72 points (in excess of 0.7 s.d.) in relation to the INIT analysis. The “true” value, once the defined guesses were suppressed (GS analysis) in the SIM1 data, was in excess of 200 scaled score points. This underestimation is reflected in the percentages assigned to the achievement levels shown in Table 9.2. As indicated in Chapter 5, the GIP techniques were unable to fully identify guessing (compared to the GS defined cases), especially as the observed raw score of the student increased. Table 9.2 shows that the GIPINIT3B values for the Quartile 3 students were, on average, about 30 points higher than the INIT values. The lower-ability students were overestimated by approximately 38 scaled score points – the equivalent of approximately 0.4 s.d. – a significant effect size, reflecting approximately five months of student learning.

By comparison, in comparing the “true” levels (GS analysis) to the INIT analysis derived levels, there are differences in the reported levels of each student at the cusp of a quartile (students 100, 200, 300 and 400, respectively). However, in the GIPINIT3B analysis, the students on the cusps of the upper quartiles (Q4 and Q3) shifted across the level boundary to denote improvement, while those on the cusps of the two lower quartiles (Q2 and Q1) were reported at approximately the same level as the INIT level analysis.

Tables 9.8 to 9.11 show the impact on the scaled score comparisons of the three analyses on each of the other four simulation studies. Tables 9.10 to 9.11 show similar patterns in relation to the changes in revised values to those displayed in SIM1 (see Table 9.7). The mid-range quartiles/deciles of the GIPINIT3B analysis phases are in a similar range to the INIT values but there are definite differences in the upper and lower quartiles/deciles. For the simulation cohorts of 400, four groups (quartiles) of 100 students are reported, and for the larger sample of 1000, 10 groups of 100 (deciles) were generated for comparison purposes.

Table 9.8

Comparison of Mean Scaled Scores by Quartile for SIM2 Data

Quartile	SIM2 Mean SSINIT	SIM2 Mean SSGS	SIM2 Av.Level INIT	SIM2 Av. Level GS	SIM2 Mean SSGIP3A <small>p=0.6</small>	SIM2 Av. Level GIP3A <small>p=0.6</small>	SIM2 Mean SS GIPINIT3B <small>p=0.6</small>	SIM2 Av. Level GIPINIT3B <small>p=0.6</small>	SIM2 Δ mean SS GIP3A <small>p=0.6 with INIT</small>	SIM2 Δ mean SS GIPINIT3B <small>p=0.6 with INIT</small>
Q4	634	661	4.8	5.1	704	5.6	704	5.6	70	70
Q3	532	554	4.0	4.0	585	4.4	585	4.4	53	53
Q2	461	476	3.0	3.0	473	3.3	489	3.4	12	28
Q1	376	377	2.2	2.5	280	1.6	369	2.2	-96	-7
Average	500	516			509		536			
Std Dev	100	110			167		130			

Table 9.9

Comparison of Mean Scaled Scores by Decile for SIM3 Data

Quartile	SIM3 Mean SSINIT	SIM3 Mean SSGS	SIM3 Av.Level INIT	SIM3 Av. Level GS	SIM3 Mean SSGIP3A <small>p=0.6</small>	SIM3 Av. Level GIP3A <small>p=0.6</small>	SIM3 Mean SS GIPINIT3B <small>p=0.6</small>	SIM3 Av. Level GIPINIT3B <small>p=0.6</small>	SIM3 Δ mean SS GIP3A <small>p=0.6 with INIT</small>	SIM3 Δ mean SS GIPINIT3B <small>p=0.6 with INIT</small>
Decile 10	685	881	5.2	6.0	730	5.7	730	5.7	45	45
Decile 9	611	726	4.7	5.7	647	5.0	647	5.0	36	36
Decile 8	568	605	4.0	4.5	596	4.4	596	4.4	28	28
Decile 7	529	487	4.0	3.3	545	4.0	549	4.0	16	20
Decile 6	502	434	3.4	2.8	499	3.4	516	3.9	-3	14
Decile 5	484	401	3.0	2.5	468	3.0	493	3.0	-16	9
Decile 4	462	357	3.0	2.0	432	2.9	464	3.0	-30	2
Decile 3	435	282	3.0	1.4	374	2.2	429	3.0	-61	-6
Decile 2	395	184	2.4	1.0	307	1.6	378	2.2	-88	-17
Decile 1	333	35	1.8	1.0	199	1.0	305	1.7	-134	-28
Average	500	438			479		510			
Std Dev	100	244			155		122			

Considering the outcomes of the simulated distributions that were not centrally targeted it is noticeable that in the case of the test that is too easy for the cohort (Table 9.10) the GIP process tends to indicate a greater proportion of students in the more able deciles than the INIT analysis. In the case of the test being too hard for the cohort (Table 9.11) the opposite tends to be the case with the INIT analysis overestimating the outcomes of the lower-ability students compared to the GIP analyses.

Table 9.10

Comparison of Mean Scaled Scores by Quartile for SIM4 Data

Quartile	SIM4 Mean SSINIT	SIM4 Mean SSGS	SIM4 Av.Level INIT	SIM4 Av. Level GS	SIM4 Mean SSGIP3A <small>p=0.6</small>	SIM4 Av. Level GIP3A <small>p=0.6</small>	SIM4 Mean SS GIPINIT3B <small>p=0.6</small>	SIM4 Av. Level GIPINIT3B <small>p=0.6</small>	SIM4 Δ mean SS GIP3A <small>p=0.6</small> with INIT	SIM4 Δ mean SS GIPINIT3B <small>p=0.6</small> with INIT
Q4	632	743	4.8	5.7	698	5.5	698	5.5	66	66
Q3	537	550	3.9	4.0	572	4.3	573	4.3	35	36
Q2	455	397	3.0	2.5	438	2.9	456	3.1	-17	1
Q1	377	251	2.3	1.3	301	1.6	357	2.1	-76	-20
Average	500	485			502		521			
Std Dev	100	194			156		134			

Table 9.11

Comparison of Mean Scaled Scores by Quartile for SIM5 Data

Quartile	SIM5 Mean SSINIT	SIM5 Mean SSGS	SIM5 Av.Level INIT	SIM5 Av. Level GS	SIM5 Mean SSGIP3A <small>p=0.6</small>	SIM5 Av. Level GIP3A <small>p=0.6</small>	SIM5 Mean SS GIPINIT3B <small>p=0.6</small>	SIM5 Av. Level GIPINIT3B <small>p=0.6</small>	SIM5 Δ mean SS GIP3A <small>p=0.6</small> with INIT	SIM5 Δ mean SS GIPINIT3B <small>p=0.6</small> with INIT
Q4	624	866	4.8	5.9	625	4.8	713	5.5	1	89
Q3	548	565	4.0	4.2	538	3.8	568	4.2	-10	20
Q2	460	302	3.0	1.7	408	2.6	354	2.1	-52	-106
Q1	370	81	2.2	1.0	286	1.4	172	1.0	-84	-198
Average	500	453			464		451			
Std Dev	100	319			135		216			

9.3 Study 2, Field Trial Data

Tables 9.12 and 9.13 are presented in a similar format to those presented for the simulated data and provide comparisons of the relative “reassignment” of students to performance levels in the English versions of the Field data. These results reflect the degree to which there may be misinformation in assessments that are ‘too easy’ for the target cohort. The analyses of Chapter 6 indicated that there were relatively few student/item interactions indicated as highly probable guessing which restricted the efficacy of the GIP process.

9.3.1 Year 5

As noted in Section 9.2.4, of Chapter 6, the capacity of the GIP procedure to interact with data collected from easy tests was limited. In the case of these data the test proved very easy with several items having difficulty locations at the extreme low end of the ability scale. Tables 9.12 and 9.13 highlight these limitations.

Table 9.12 shows that in comparing the students in the highest achievement level there was no discrimination observed in any of the analyses. At the lowest level of achievement, there was an increase observed in the SIG analysis compared to the INIT analysis outcomes, which reflects the lower-ability students indicating that they had guessed the relatively few harder items. The GIP analysis shows a similar trend for the lower-ability students, although the numbers in each of the observations of Level 1 were small.

Table 9.12

Comparisons of Percentages in Each Level for Field Trial Data – Y5 English Version Mathematics

Level	Y5FT LevelINIT			Y5FT LevelSIG			Y5FT LevelGIP3A $p=0.6$			Y5FT LevelGIP3B $p=0.6$ GIPINIT		
	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent
6	8	2.6	100.0	8	2.6	100.0	8	2.6	100.0	8	2.6	100.0
5	38	12.5	97.4	49	16.2	97.4	54	17.8	97.4	54	17.8	97.4
4	101	33.3	84.8	81	26.7	81.2	85	28.1	79.5	85	28.1	79.5
3	102	33.7	51.5	91	30.0	54.5	100	33.0	51.5	102	33.7	51.5
2	52	17.2	17.8	59	19.5	24.4	49	16.2	18.5	52	17.2	17.8
1	2	0.7	0.7	15	5.0	5.0	7	2.3	2.3	2	0.7	0.7
	303	100.0		303	100.0		303	100.0		303	100.0	

9.3.2 Year 7

The analysis of the Year 7 sample was limited not only by the ease of the test but also by the relatively small sample size. Table 9.13 shows small increases in the percentages of students in the lower ability levels in the SIG and GIP3A analyses compared to the INIT analysis and increases in percentages of students in the higher Levels as predicted by the outcomes of SIM4.

Table 9.13

Comparisons of Percentages in Each Level for Field Trial Data – Y7 English Version Mathematics

Level	Y7FT LevelINIT			Y7FT LevelSIG			Y7FT LevelGIP3A $p=0.6$			Y7FT LevelGIP3B $p=0.6$ GIPINIT		
	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent
6	4	2.1	100.0	2	1.1	100.0	14	7.4	100.0	14	7.4	100.0
5	25	13.2	97.9	25	13.2	98.9	27	14.3	92.6	27	14.3	92.6
4	53	28.0	84.7	39	20.6	85.7	34	18.0	78.3	41	21.7	78.3
3	84	44.4	56.6	79	41.8	65.1	77	40.7	60.3	84	44.4	56.6
2	17	9.0	12.2	22	11.6	23.3	20	10.6	19.6	16	8.5	12.2
1	6	3.2	3.2	22	11.6	11.6	17	9.0	9.0	7	3.7	3.7
	189	100.0		189	100.0		189	100.0		189	100.0	

The outcomes of the Arabic versions of the tests are not reported here. This is because the suppression of self-identified guessed responses in the SIG analysis resulted in the removal of a significant number of “extreme” items from the analysis which resulted in the scales developed for the SIG analyses being no longer directly comparable with the INIT outcomes. Similarly, the suppression of multiple items in the GIP analyses and the uncertainty of the outcomes caused by the random nature of the data made comparisons to the INIT outcomes inconclusive.

9.4 Study 3, Large-Scale Data Sets

The ultimate aim of this research is to estimate the degree to which student outcomes are biased in analyses that do not take account of the presence of guessing in student responses. The research also aimed to provide a strategy to indicate and mitigate the impact of guessing on the outcomes of large-scale assessments. The commentary and Tables that follow investigate the issue of guessing in Rasch analyses using authentic large-scale data.

9.4.1 Reported Levels for Large-Scale Data

Since there was no capacity to illicit any defined or self-identified guessing in the tests, only the INIT analysis and the GIP analysis outcomes are displayed in Tables 9.14 to 9.16. These tables show the comparisons between the outcomes of the INIT analysis and the GIP3A/GIPINIT3B analyses for the authentic large-scale data sets. They show a similar pattern to the outcomes in the simulations, particularly SIM3 and SIM5, for the INIT analysis compared to the GIPINIT3B analysis which is encouraging given the variability experienced in ‘live’ data.

9.4.1.1 Grade 4 Mathematics Levels Analysis

In comparing the LevelINIT outcomes with the LevelGIPINIT3B outcomes for Grade 4 Mathematics (Table 9.14), the percentage of students in Level 1 (lowest ability students) increased by almost 1200 cases (4.4% of the cohort). These Level 1 students represent the “at risk” students for whom the majority of the skills and concepts, as assessed by this test, were lacking and, as a consequence, may require future targeted learning. At Level 2, an additional 1500 (5.8%) students were indicated as performing below the expected level for students of the target grade.

Table 9.14

Comparisons of Percentages in Each Level for Large Scale Assessment – G4 Mathematics

Level	G4 Math LevelINIT			G4 Math LevelGIP3A _{p=0.6}			G4 Math LevelGIPINIT3B _{p=0.6}		
	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent
6	744	2.8	100.0	1333	5.1	100.0	1333	5.1	100.0
5	3780	14.4	97.2	3191	12.1	94.9	3191	12.1	94.9
4	7009	26.7	82.8	5419	20.6	82.8	7009	26.7	82.8
3	10979	41.8	56.1	3705	14.1	62.2	8281	31.5	56.1
2	3172	12.1	14.3	4662	17.7	48.1	4713	17.9	24.6
1	595	2.3	2.3	7969	30.3	30.3	1752	6.7	6.7
	26279	100.0		26279	100.0		26279	100.0	

Typically, in this type of test students performing at Levels 3 and 4 are considered to be either approaching or performing at the expected level of the cohort grade for the curriculum skills assessed. The comparison of the percentages indicated by the INIT and GIPINIT3B analyses are similar. This is the outcome predicted by the simulation studies, with little variation observed in the mid-range ability estimates noted in these regions. At the higher ability levels, there is accord between the INIT analysis outcomes and the GIPINIT3B analysis outcomes, as observed in the simulation studies. Although the increase in the number of students in the highest ability level was relatively small (2.3%), this value represents the 589 students who were underestimated in the highest achievement level by the INIT analysis.

9.4.1.2 Grade 8 Mathematics Levels analysis

Table 9.15 shows the comparisons of the percentages of students in each level for the Grade 8 Mathematics cohort. The table shows a similar set of outcomes to the G4 Mathematics cohort (see Table 9.14). At both Level 1 and Level 2, there were considerable increases in the percentage of students identified by the GIPINIT3B analysis compared to the INIT analysis. In total, an additional 2015 students (9.6% of the cohort) were identified as performing in these “at risk” levels of Levels 1 and 2. In the higher levels, the GIPINIT3B analysis identified fewer students as reclassified, but there were still 447 students (2.1%) in the highest level whose ability was underestimated by the INIT analysis.

Table 9.15

Comparisons of Percentages in Each Level for Large Scale Assessment – Grade 8 Mathematics

Level	G8 Math LevelINIT			G8 Math LevelGIP3A _{p=0.6}			G8 Math LevelGIPINIT3B _{p=0.6}		
	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent
6	788	3.8	100.0	1235	5.9	100.0	1235	5.9	100.0
5	2774	13.2	96.2	2327	11.1	94.1	2327	11.1	94.1
4	6119	29.1	83.0	3649	17.4	83.1	4598	21.9	83.1
3	8092	38.5	53.9	3691	17.6	65.7	7598	36.2	61.2
2	3036	14.5	15.4	3407	16.2	48.1	4615	22.0	25.0
1	198	0.9	0.9	6698	31.9	31.9	634	3.0	3.0
	21007	100.0		21007	100.0		21007	100.0	

The outcomes shown in Tables 9.14 and 9.15 tended to follow the pattern predicted by SIM5. In the cases of the Grade 4 and Grade 8 Mathematics the tests provided to be marginally hard for the target cohorts. SIM5 was the comparable simulation in which the GIP3A and GIPINIT3B analyses tended to report greater percentages of students in the lower-ability groups with only marginal differences in the upper ability groups. The outcomes observed in the respective analyses of Grade 4 and Grade 8 tended to follow this pattern.

9.4.1.3 Grade 4 Science Levels analysis

Table 9.16 compares the percentage of students in each level for the Grade 4 Science cohort. The overall outcomes follow a similar pattern to those shown in Tables 9.14 and 9.15. Importantly, the GIPINIT3B analysis produced an outcome with considerably more of the lower-ability students being reclassified into the “at risk” levels. Although the GIPINIT3B procedure did not indicate there were more students in the lowest level, it identified approximately 2,000 students (about 8.5%) who were probably performing below the level expected for Grade 4 students in the skills assessed by this test when guessing is not accounted for. The percentages observed in Levels 3 and 4 are relatively similar for both Mathematics and Science in the INIT and GIPINIT3B analyses. At the upper levels, there is no difference in the Level 6 outcomes, although an approximate additional 1,250 (4.8%) students were indicated by the GIPINIT3B analysis as performing at Level 5, which is marginally above the expected level of achievement for the cohort.

Table 9.16

Comparisons of Percentages in Each Level for Large-Scale Assessment – Grade 4 Science

Level	Grade 4 Science LevelINIT			Grade 4 Sci LevelGIP3A $p=0.6$			Grade 4 Sci LevelGIPINIT3B $p=0.6$		
	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent	Freq	Percent	Cumul Percent
6	1140	4.4	100.0	1140	4.4	100.0	1140	4.4	100.0
5	3204	12.3	95.6	4462	17.1	95.6	4462	17.1	95.6
4	7516	28.8	83.3	5640	21.6	78.5	6258	24.0	78.5
3	10946	42.0	54.5	4746	18.2	56.9	8973	34.4	54.5
2	2992	11.5	12.5	5139	19.7	38.7	4965	19.0	20.1
1	267	1.0	1.0	4938	18.9	18.9	267	1.0	1.0
	26065	100.0		26065	100.0		26065	100.0	

It is worth noting that the numbers and percentage of students in the lowest two levels for the GIP3A analysis are considerably different between the INIT and GIPINIT3B outcomes. This is a reflection of suppressing the identified guesses and reducing the raw score by the sum of the identified guesses for each student. As observed throughout this study, this has been a consistent and expected feature of this phase of the GIP3A process.

9.4.2 Scaled Score Comparisons of the Large-Scale Data sets

9.4.2.1 Grade 4 Mathematics Scaled Score Analysis

This sub-section shifts the focus of the analysis from Levels to reported scaled scores. When changes in scaled score between the INIT value and the GIP values are considered, a pattern emerges that confirms the impact of accounting for guessing is most notable at the extremes of the scales. Each analysis of reported scaled scores shows that there were relative low degrees of change in the percentages of mid-range ability students (deciles 3 to 8) but in all cases there were significant changes in the scores of the upper ability students (decile 10) and the lower-ability students (deciles 1, 2 and 3). This outcome was expected by the hypothesis that underpinned the methodology that took account of guessing in the student response patterns.

Table 9.17 shows the differences in the mean scaled scores of Grade 4 Mathematics students when grouped

by decile. Each decile comprised approximately 2728 students. The table shows the difference between the mean score of each decile, comparing the INIT value with the GIP3A and GIPINIT3B values, and highlights the degree to which students falling within particular decile groups were underestimated or overestimated when there was no consideration for guessing in the student responses.

Table 9.17

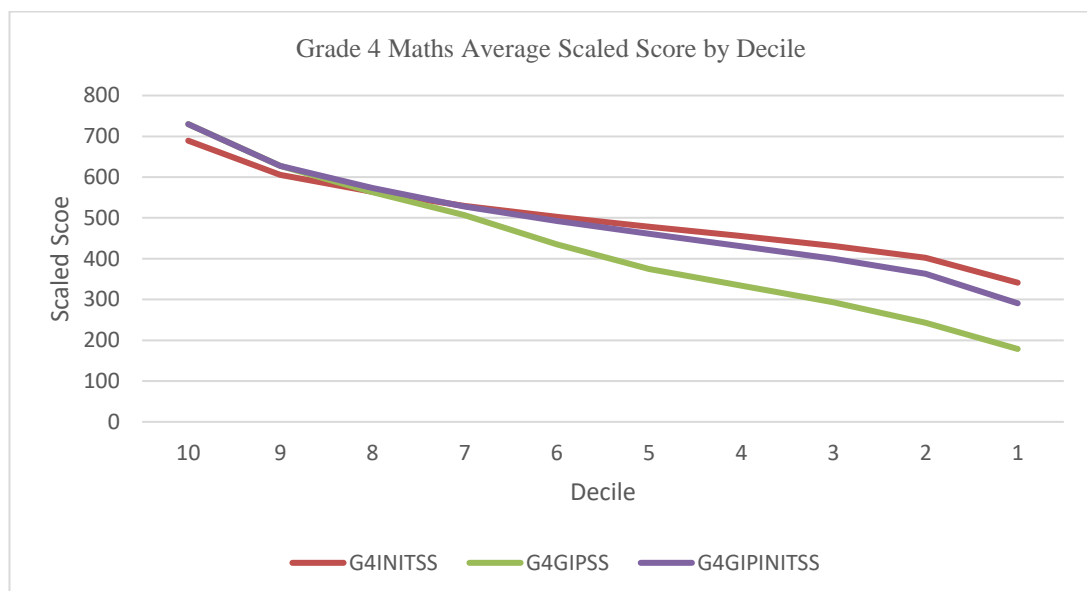
Differences Between INIT and GIP Standardised Scale Scores for Grade 4 Mathematics

Decile	Av. Scaled Score INIT	Av. Scaled Score GIP3A	Av. Scaled Score GIPINIT3B	Δ INITSS and GIP3ASS	Δ INITSS and GIPINIT3BSS
10	689	730	730	41	41
9	606	627	627	21	21
8	564	563	573	-1	9
7	529	506	528	-23	-1
6	503	435	493	-68	-10
5	479	375	461	-104	-18
4	455	334	431	-121	-24
3	432	293	400	-139	-32
2	402	243	363	-159	-39
1	341	179	291	-162	-50

Figure 9.4 shows that the “break-even” point was around the decile 8. Students positioned below the decile 8 had standardised GIPINIT3B scale scores inflated by the INIT raw scores by up to 50 scaled score points.

Figure 9.4

Comparison of Grade 4 Mathematics Scaled Scores by Decile



9.4.2.2 Grade 8 Mathematics Scaled Score Analysis

Table 9.18 and Figure 9.5 show a very similar pattern in the differences in the mean scaled scores of Grade 8 with relatively small changes in the mid-range ability deciles and larger score changes at each of the extreme ability deciles.

Table 9.18 shows that the highest ability students were, on average, underestimated by 30 scaled score points by the INIT analysis compared to the GIPINIT3B analysis and the lowest ability students were, on average, overestimated by 36 scaled score points. As shown in Figure 9.4 the “break-even” point was in decile 8, a marginally higher decile value than observed in Grade 4 Mathematics.

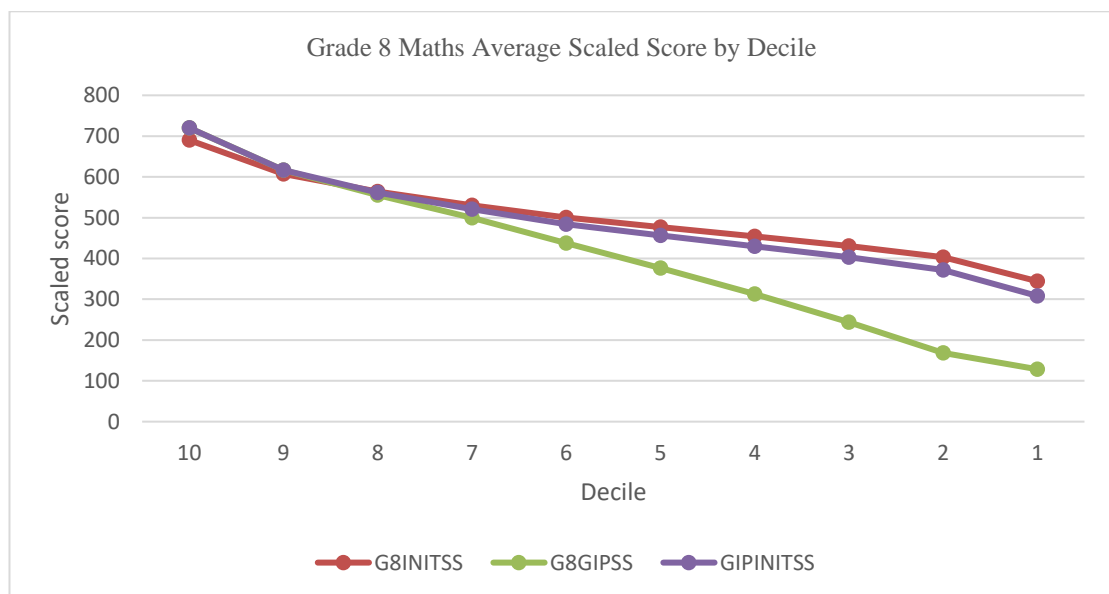
Table 9.18

Differences Between INIT and GIP Standardised Scale Scores for Grade 8 Mathematics

Decile	Av. Scaled Score INIT	Av. Scaled Score GIP3A	Av. Scaled Score GIPINIT3B	Δ INITSS and GIP3ASS	Δ INITSS and GIPINITSS
10	690	720	720	30	30
9	607	616	616	9	9
8	564	556	562	-8	-2
7	531	500	521	-31	-10
6	500	438	484	-62	-16
5	477	377	456	-100	-20
4	454	313	430	-141	-24
3	431	244	403	-187	-27
2	403	168	372	-235	-31
1	344	128	308	-216	-36

Figure 9.5

Comparison of Grade 8 Mathematics Scaled Scores by Decile



9.4.2.3 Grade 4 Science Scaled Score Analysis

The outcomes shown in Table 9.19 and Figure 9.6 are similar to those shown for the two Mathematics analyses. A difference can be seen in the position of the “break-even” point, which for Grade 4 Science was decile 6. This probably reflects the marginally better targeting of this test with respect to the item locations and the cohort ability estimates.

Whereas both the Grade 4 and Grade 8 Mathematics tests were marginally difficult for the cohorts, the G4 Science assessment was reasonably well targeted (see Figure 7.3). The impact of this better targeting appears to be a more central break-even point, with the relatively similar distributions of the differences in the numbers of over/under estimation for the decile groups about decile 6.

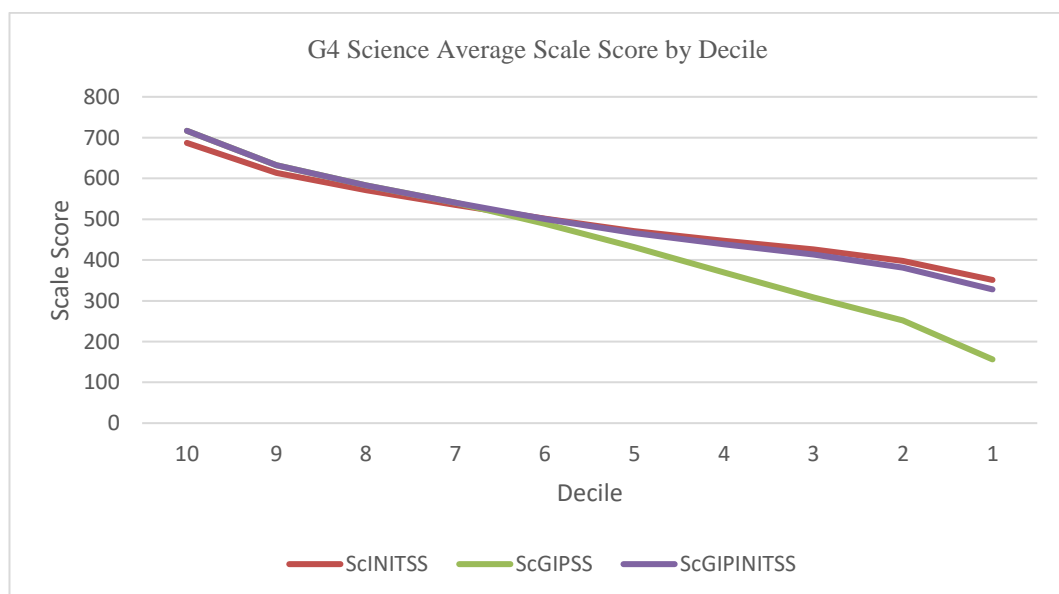
Table 9.19

Differences Between INIT and GIP Standardised Scale Scores for Grade 4 Science

Decile	Av. Scaled Score INIT	Av. Scaled Score GIP	Av. Scaled Score GIPINIT	Δ INITSS and GIPSS	Δ INITSS and GIPINIT SS
10	687	717	717	30	30
9	613	632	632	19	19
8	571	583	583	12	12
7	535	540	540	6	6
6	501	489	501	-12	0
5	471	431	466	-40	-5
4	447	370	438	-78	-9
3	426	308	413	-118	-13
2	398	252	381	-146	-17
1	351	156	328	-195	-23

Figure 9.6

Comparison of Grade 4 Science Scaled Scores by Decile



In summary, the indication and implementation of a GIP process that indicated probable guessed items and suppressed those items in a supplementary analysis resulted in considerable differences in the percentages of students' reported achievement outcomes in the both the higher and lower ability levels. The current requirement that students be credited with all correct responses, irrespective of whether they are probable guesses, reduced the degree to which the misclassification was recognised at the lower achievement levels. The GIP procedure indicated the probable guesses more frequently in the more difficult items, with the consequent impact on the difficulty location of all items. This resulted in the recalibration of student ability estimates and consequent increases in the percentages of students identified in the upper and lower ability levels compared to the INIT analysis.

The consistency between the outcomes of these large-scale authentic data with the simulation studies and the convergence of these outcomes with the hypothesis of this study suggest that a process like this should be applied to all large-scale assessments to provide a better scale of achievement and more accurate estimates of the ability of students on the revised scale.

9.5 Scale of Reclassifications

The analyses of this research reveal that although there is a consistency in the impact of the GIP procedures in improving the measurement scale and better locating student ability estimates on the scale there is variability in the degree to which student achievement is reclassified by the recalibration of the estimates. The section below discusses this element of the impact of the GIP procedure.

9.5.1 Degree of Reclassification of Reported Outcomes in Simulated Data

Tables 9.20 to 9.27 summarise the degree of reclassification by level for the various analyses. The simulation data are included to show the relative impact on "true" values when there was no adjustment for guessing in the INIT analyses. The net effect sums to zero as students are redistributed among the levels. Even so, the focus of the study is the degree to which students have been misreported by not accounting for guessing in the MC items.

Table 9.20 shows the percentages of students in each level for each analysis phase for the SIM1 data. The INIT analysis produced a relatively normal distribution of percentages in each level. When comparing the outcomes of the INIT analysis with the "true" values of the GS analysis for SIM1, the outcomes show a significant under-identification of students in the lowest achievement level, with the GS analysis revealing that 15% of students may be unrecognised at Level 1 due to successful guessing.

Although the GIP process was unable to indicate all the defined guessed responses, its overall impact on identifying students in the "at risk" Levels 1 and 2 was to better reflect the "true" achievement of students than the INIT analysis. Specifically, the GIPINIT3B analysis identified 24% of students in the "at risk" levels compared to 16.3% in the INIT analysis. At the higher end of the scale, the GIPINIT3B analysis identified 26.3% of students in the Level 5–6 range compared to 15.8% in the INIT analysis. This outcome reflects the impact of the relationship between ability estimates and item difficulty when the GIPINIT3B process has accounted for guessing in the more difficult items, despite students receiving full credit for guessed responses.

These outcomes are consistent with the expectation of the study, which is that the current practice of a simple Rasch analysis of students' assessment underestimates item locations and consequently overestimates the ability estimates of lower-ability students. In addition, the recalibration of item locations revealed an increased number of higher-ability students being reported at even higher achievement levels.

Table 9.20

Discrepancy Between Reported Percentages in Levels – the Degree of Reclassification SIM1

Level	SIM1 Level INIT	SIM1 Level GS	SIM1 Level GIP	SIM1 Level GIPINIT	Reclassifications		
	Percent	Percent	Percent	Percent	reclass% GS	reclass% GIP3A	reclass% GIPINIT3B
6	2.3	26.0	10.8	14.5	23.8	8.5	12.2
5	13.5	8.5	13.8	11.8	-5.0	0.3	-1.7
4	29.0	25.8	23.5	24.0	-3.3	-5.5	-5.0
3	39.0	13.3	25.0	25.8	-25.8	-14.0	-13.2
2	15.0	11.5	12.8	20.0	-3.5	-2.3	5.0
1	1.3	15.0	14.3	4.0	13.8	13.0	2.7
	100.0	100.0	100.0	100.0			

Note. The three “reclassifications” columns show the difference between the percentages in each level for the column compared with the percentage shown in the INIT column.

Table 9.20 shows that the increases in the number of students in the upper and lower ability levels were compensated by reductions in the percentage of students in the mid-range ability levels. Hence, in Tables 9.20 to 9.24, the increased percentages in the higher and lower ability levels are consistent with the reductions in the percentages reported in Levels 3 and 4 in the GIP3A and GIPINIT3B analyses compared to the INIT analysis. Table 9.21 shows fewer changes in the percentages in each level, which is a reflection of the fewer candidates in the sample and the fewer items from which guessing can be defined (GS) or identified (GIPINIT3B). Table 9.21 also demonstrates the bluntness of the “Level” paradigm. In the GIP3A and GIPINIT3B analyses there was an increase in the percentage of students in Levels 5 and Level 6 compared to both the INIT and GS phases. This demonstrates the limitation of a single cut score. In this case, there were students whose scale scores just exceeded the limit scale scores of 400 and 500 respectively, which caused them to be reported in the higher ability level. This demonstrates the rationale that scale-score reporting is an important adjunct to Level reporting.

Table 9.21*Discrepancy Between Reported Percentages in Levels – the Degree of Reclassification SIM2*

Level	SIM2 Level INIT	SIM2 Level GS	SIM2 Level GIP	SIM2 Level GIPINIT	Reclassifications		
	Percent	Percent	Percent	Percent	reclass% GS	reclass% GIP3A	reclass% GIPINIT3B
6	4.4	4.4	16.0	16.0	0.0	11.6	11.6
5	11.6	19.6	18.4	18.4	8.0	6.8	6.8
4	34.8	26.8	26.0	26.0	-8.0	-8.8	-8.8
3	29.2	36.8	10.8	19.6	7.6	-18.4	-9.6
2	19.2	11.2	17.2	18.8	-8.0	-2.0	-0.4
1	0.8	1.2	11.6	1.2	0.4	10.8	0.4
	100.0	100.0	100.0	100.0			

Table 9.22 shows the changes for a larger sample with a longer test length in SIM3. The outcomes of the analyses of this data set demonstrate the underestimations of the INIT analysis in reporting percentages of students in both the lower-ability group and the higher-ability group. The GS analysis shows the “true” percentages whilst the GIPINIT3B analysis shows a similar pattern to the GS analysis, but with the constraints in the efficiency of identifying guessed item/person interactions, as discussed in Chapter 7. However, even with that constraint, the GIPINIT3B analysis was a more efficient reflection of the real achievement level with respect to the reporting of performance and ability compared to the INIT analysis that did not consider guessing in the student responses.

Table 9.22*Discrepancy Between Reported Percentages in Levels – the Degree of Reclassification SIM3*

Level	SIM3 Level INIT	SIM3 Level GS	SIM3 Level GIP3A	SIM3 Level GIPINIT3B	Reclassifications		
	Percent	Percent	Percent	Percent	reclass% GS	reclass% GIP3A	reclass% GIPINIT3B
6	1.8	16.7	7.1	7.1	14.9	5.3	5.3
5	14.6	8.8	16.5	16.5	-5.8	1.9	1.9
4	27.3	7.9	20.2	25.4	-19.4	-7.1	-1.9
3	40.2	19.3	27.0	33.2	-20.9	-13.2	-7.0
2	14.5	19.8	15.0	14.4	5.3	0.5	-0.1
1	1.6	27.5	14.2	3.4	25.9	12.6	1.8
	100.0	100.0	100.0	100.0			

With regard to SIM3, Table 9.23 shows the outcomes of a test that was marginally easy for the cohort. The INIT analysis was relatively normally distributed about the Level 3 and 4 percentages. The GS analysis had a relatively equal distribution of percentages across the levels, with larger percentages reported in both Levels 1 and 6, compared to the INIT analysis. In this instance, the GIPINIT3B analysis was relatively inefficient in identifying students in the lowest ability level as the test was very easy and there were fewer items that allowed the identification of guessing strategies in the response data. Hence, the GIPINIT3B procedure was less efficient than the GS “true” estimations. However, the GIPINIT3B procedure did provide better discrimination in the upper levels compared to the INIT analysis.

Table 9.23*Discrepancy Between Reported Percentages in Levels – the Degree of Reclassification SIM4*

Level	SIM4 Level INIT	SIM4 Level GS	SIM4 Level GIP3A	SIM4 Level GIPINIT3B	Reclassifications		
	Percent	Percent	Percent	Percent	reclass% GS	reclass% GIP3A	reclass% GIPINIT3B
6	0.8	18.0	11.8	11.8	17.3	11.0	11.0
5	17.3	14.3	21.5	21.5	-3.0	4.3	4.3
4	28.5	10.5	17.8	18.5	-18.0	-10.8	-10.0
3	36.5	19.8	20.5	28.5	-16.8	-16.0	-8.0
2	16.0	18.0	17.8	16.8	2.0	1.8	0.8
1	1.0	19.5	10.8	3.0	18.5	9.8	2.0
	100.0	100.0	100.0	100.0			

The SIM5 outcomes presented in Table 9.24 provide a redistribution of Levels consistent with the changes in scaled scores. A significant percentage of students reported in the mid-range Levels when using the INIT analysis have been redistributed in both higher and lower Levels. More of higher-ability students have been recognized at Level 6 whilst a significant proportion of the lowest ability students have been identified as performing in the Lowest Level compared to those reported by the INIT phase.

Table 9.24*Discrepancy Between Reported Percentages in Levels – the Degree of Reclassification SIM5*

Level	SIM5 Level INIT	SIM5 Level GS	SIM5 Level GIP3A	SIM5 Level GIPINIT3B	Reclassifications		
	Percent	Percent	Percent	Percent	reclass% GS	reclass% GIP3A	reclass% GIPINIT3B
6	0.5	26.5	0.5	13.0	26.0	0.0	12.5
5	18.8	8.5	18.8	21.8	-10.3	0.0	3.0
4	31.3	6.5	25.8	10.3	-24.8	-5.5	-21.0
3	31.5	10.8	19.0	11.0	-20.8	-12.5	-20.5
2	16.0	11.0	21.5	14.8	-5.0	5.5	-1.3
1	2.0	36.8	14.5	29.3	34.8	12.5	27.3
	100.0	100.0	100.0	100.0			

9.5.2 Degree of Reclassification of Levels in Authentic Data

Tables 9.25 to 9.27 indicate the degree to which students would be reclassified in the reporting of these tests under the current procedures of a Rasch analysis of the data, without any consideration of probable/possible guessing in the student responses compared to the GIP outcomes. In comparing the percentages of students in each level from the INIT analysis to the GIPINIT3B analysis, Table 9.25 shows that 1,157 students (4.4%) had their abilities overestimated by the INIT analysis in the lowest achievement level. It is common for students indicated in Level 1 (more than 2 standard deviations below the mean) of a reporting scale to be considered “at risk” and for targeted intervention to be developed to support these students.

Similarly, at Level 2, the INIT analysis overestimated the ability of about 1,500 students (5.8%) in Grade 4 Mathematics. Based on this INIT analysis, these students would be considered as approaching the mid-range of the curriculum expectations when, in fact, as indicated by the GIPINIT3B analyses, they were more likely functioning well below the expected curriculum level.

At the higher end of the scale, approximately 600 students were not indicated at the level of their skills and knowledge because of the impact of the successful guessing of lower-ability students and the “knock-on” effect of the bias on item difficulty calibration and subsequent ability estimate calculations.

Table 9.25

Comparison of Analysis Outcomes – Percentage in Levels – Grade 4 Mathematics

Level	G4 Maths Level INIT		G4 Maths Level GIP		G4 Maths Level GIPINIT		reclass INIT and GIP3A	reclass INIT and GIPINIT3B
	Freq	Percent	Freq	Percent	Freq	Percent		
6	736	2.8	1340	5.1	1340	5.1	2.3	2.3
5	3784	14.4	3180	12.1	3180	12.1	-2.3	-2.3
4	7017	26.7	5414	20.6	7017	26.7	-6.1	0.0
3	10985	41.8	3705	14.1	8278	31.5	-27.7	-10.3
2	3180	12.1	4652	17.7	4704	17.9	5.6	5.8
1	604	2.3	7963	30.3	1761	6.7	28.0	4.4
	100.0		100.0		100.0			

Table 9.26 shows a similar overall pattern to the Grade 4 Mathematics outcomes. The general impact of the GIPINIT3B analysis was to increase the number of students in the lower two levels and in the higher level compared to the INIT analysis, with these differences being redistributed in the mid-range levels. However, a consistent pattern is observed regarding the differences in percentages in levels in comparing the INIT outcomes with the GIP outcomes.

Table 9.26

Comparison of Analysis Outcomes – Percentage in Levels – Grade 8 Mathematics

Level	G8 Maths Level INIT		G8 Maths Level GIP		G8 Maths Level GIPINIT		reclass INIT and GIP3A	reclass INIT and GIPINIT3B
	Freq	Percent	Freq	Percent	Freq	Percent		
6	798	3.8	1239	5.9	1239	5.9	2.1	2.1
5	2772	13.2	2331	11.1	2331	11.1	-2.1	-2.1
4	6111	29.1	3654	17.4	4599	21.9	-11.7	-7.2
3	8085	38.5	3696	17.6	7602	36.2	-20.9	-2.3
2	3045	14.5	3402	16.2	4620	22.0	1.7	7.5
1	189	0.9	6699	31.9	630	3.0	31.0	2.1
	100		100		100			

Referring to Table 9.27, in Grade 4 Science changes were observed in the proportion of students overestimated in Level 2 and underestimated in Level 5; however, at the extremes, the same students fell within the level boundaries.

Table 9.27*Comparison of Analysis Outcomes – Percentage in Levels – Grade 4 Science*

Level	G4 Science Level INIT		G4 Science Level GIP3A		G4 ScienceLevel GIPINIT3B		reclass INIT and GIP3A	reclass INIT and GIPINIT3B
	Freq	Percent	Freq	Percent	Freq	Percent		
6	1144	4.4	1144	4.4	1144	4.4	0.0	0.0
5	3198	12.3	4446	17.1	4446	17.1	4.8	4.8
4	7488	28.8	5616	21.6	6240	24.0	-7.2	-4.8
3	10920	42.0	4732	18.2	8944	34.4	-23.8	-7.6
2	2990	11.5	5122	19.7	4940	19.0	8.2	7.5
1	260	1.0	4914	18.9	260	1.0	17.9	0.0
		100.0		100.0		100.0		

9.6 Review of Findings

9.6.1 Study 1, Simulated Data

The simulations of Study 1 provided a framework for estimating the impact of the GIP procedure in Studies 2 and 3. Reflecting upon the outcomes of SIM1 and SIM3 (large sample and large item pool, which are typical of large-scale assessments), the GIPINIT3B analysis shows a definite increase in the students in the highest and lowest bands compared to the INIT analysis outcomes. These outcomes indicate that, for larger cohorts with a test length of 40 items, the GIPINIT3B procedure was more efficient in indicating the ability of students in the higher and lower levels than the simple Rasch analysis with no accounting for guessed items.

Although the scale of the differences in percentages within Levels varies in the comparisons shown in Table 9.28, the data were randomly generated and hence both the percentages of random guessing and the recovery rate of the GIP application were variable. However, the critical feature of interest demonstrated in the table is that the same general conditions were noted: the overestimation of the lower-ability students and the underestimation of the higher-ability students.

These observed differences are relevant in measuring the learning growth of students. An artefact of the INIT analysis observable in all analyses is the compression of the scale when there was no consideration of guessing. The impact of this compression was the potential misclassification of lower-ability students into relatively higher levels, and the underestimation of achievement of the higher-ability students.

Table 9.28

Summary of Differences in Percentages in Levels, Study 1, Simulations 1 to 5

Data Set	N	Treatment	Percentages Within Specific Levels			
			Percent Level 1	Percent Level 1 and Level 2	Percent Level 5 and Level 6	Percent Level 6
SIM 1	400	INIT	1.3	16.3	15.8	2.3
		GS	15.0	26.5	34.5	26.0
		GIP3A	14.3	27.0	24.6	10.8
		GIPINIT3B	4.0	24.0	26.3	14.5
SIM 2	250	INIT	0.8	20.0	16.0	4.4
		GS	1.2	12.4	24.0	4.4
		GIP3A	11.6	28.8	34.4	16.0
		GIPINIT3B	1.2	20.0	34.4	16.0
SIM 3	1000	INIT	1.6	16.1	16.4	1.8
		GS	27.5	47.3	25.5	16.7
		GIP3A	14.2	29.2	23.6	7.1
		GIPINIT3B	3.4	17.8	23.6	7.1
SIM 4 easy	400	INIT	1.0	17.0	18.1	0.8
		GS	19.5	37.5	32.3	18.0
		GIP3A	10.8	28.5	33.3	11.8
		GIPINIT3B	3.0	19.8	33.3	11.8
SIM 5 hard	400	INIT	0.3	11.3	8.3	0.0
		GS	27.8	46.0	33.3	19.8
		GIP3A	14.5	36.0	23.3	0.5
		GIPINIT3B	29.3	44.0	34.8	13.0

It is worth noting that when comparing percentages within levels in this manner, an issue arising is the bluntness of these indicators. With a bandwidth of 100 scaled score points for a Level, it is possible to “improve” by up to 98 scaled score points (.98 s.d.) and still be re-coded as within the same level after one year of study. In relation to a measure of student “growth”, this is potentially misleading. Consideration of the scaled scores in conjunction with the reported level provides a more refined measure of changes in ability over a period. It is in this level of analysis that the action of not accounting for guessing in MC items was more problematic.

9.6.2 Study 2, Small-Scale Sample

The English versions of the test implemented with the small-scale field study were shown to be generally too easy for the students. The similarities in the outcomes between the INIT and GIPINIT3B values shown in Table 9.29 may reflect the interaction of the relatively small samples in the field trial and the tests proving to be relatively easy for the target sample. In reviewing the outcomes of the SIG analysis, the observed outcomes may be associated with the uncertainty regarding the students’ capacity to indicate which responses (correct and/or incorrect) were guessed.

Table 9.29*Summary of Differences in Percentages in Levels, Study 2, Field Trial Data Year 5 and Year 7*

Field Trial Summary		Percentages Within Specific Levels				
Data Set	N	Treatment	% level 1	%Level 1 and 2	% level 5 and 6	% level 6
Y5 Mathematics	303	INIT	0.7	17.8	15.1	2.6
		SIG	5.0	24.4	18.8	2.6
		GIP3A	2.3	18.5	20.4	2.6
		GIPINIT3B	0.7	17.8	20.4	2.6
Y7 Mathematics	189	INIT	3.2	12.2	15.3	2.1
		SIG	11.6	23.3	14.3	1.1
		GIP3A	9.0	19.6	21.7	7.4
		GIPINIT3B	3.7	12.2	21.7	7.4

9.6.3 Study 3, Large-Scale Outcomes

The summary outcomes shown in Table 9.30 are similar to the outcomes observed in SIM1 and SIM3, with the percentages observed in the highest achievement levels and those in the lowest achievement level increased in the GIPINIT3B analysis compared to the INIT analysis.

The improvements in the scale and the capacity of the GIP procedure to better estimate ability is not as pronounced as the simulations, but this reflects the variation expected in live data that do not accord exactly with the theoretical situation and the lower number of MC items in these tests. In general, the similarities in direction observed in these analyses support the contention that the GIP procedure improves the scale and provides an improved estimate of student achievement across the scale.

Table 9.30*Summary of Differences in Percentages in Levels, Study 3, Large-Scale Data*

			Percentages Within Specific Levels			
Data Set	N	Treatment	% level 1	%Level 1 and 2	% level 5 and 6	% level 6
Grade 4 Mathematics	26279	INIT	2.3	14.3	17.2	2.8
		GIP3A	30.3	48.1	17.2	5.1
		GIPINIT3B	6.7	24.6	17.2	5.1
Grade 8 Mathematics	21003	INIT	0.9	15.4	17.0	3.8
		GIP3A	31.9	48.1	17.0	5.9
		GIPINIT3B	3.0	25.0	17.0	5.9
Grade 4 Science	26067	INIT	1.0	12.5	16.7	4.4
		GIP3A	18.9	38.7	21.5	4.4
		GIPINIT3B	1.0	20.1	21.5	4.4

9.7 Summary

This chapter has reported the outcomes of the Initial (INIT) and Guessing Identification Protocol (GIP3A, GIPINIT3B) analyses on a set of large-scale authentic data. The tables and figures presented show the impact on reported achievement levels as a result of ignoring the effect of probable guessing in student responses. The application of the GIP procedures, developed through the observation of the simulation parameters and the confirmation of the determined parameters for the GIP, and subsequent GIPINIT3B values, were effective in providing a structured approach to the identification of probable guessing in large-scale data.

The impact of the application of the GIP procedure on the scaled scores follow the expected pattern, with the higher-ability students underestimated and adjusted upward by the GIP process and the lower-ability students overestimated and adjusted lower by the GIP procedure. The GIPINIT3B analysis tended to replicate the GIP3A values in the higher levels, as the higher raw scores were less impacted by the GIP processes. However, the reinstatement of the credit for the guessed responses (GIPINIT3B) had a larger impact on the capacity of the process to “correctly” locate the lower-ability students. The mid-range ability students had only marginal changes in their pre- and post-GIP ability estimates, and the overall distribution was structurally normal. The analyses in this chapter also demonstrate the importance of well-targeted tests of sufficient item difficulty range and length to allow for accurate estimation of student ability (Wright, 2008) and a more effective application of the GIP process.

The INIT outcomes reported above represent the current practice in the Australian National Assessment Program contexts (NAPLAN). The outcomes displayed and the commentary provided highlight the deficiency in this practice in which no account is taken of probable guessing in student responses.

Chapter 10

Discussion: Limitations and Merits of the GIP

10.1 Introduction – Limiting Factors

The investigations of the simulated data indicate that the extent of change in student ability estimates, when employing the Guessing Indication Protocol (GIP) process compared to typical current practice, was up to 100 scaled score points as a result of accounting for guessing, which is at the extremes of the scale. This means that up to 15% (see Table 9.4) of students who were classified as meeting a minimum level of proficiency based on current approaches would be identified as actually being below the proficiency level when the GIP process was applied to account for guessing. However, the GIP procedure experienced some challenges in identifying all of the defined guesses in the simulation data and indicating guesses in the subsequent applications to authentic data, as shown in Chapter 7. The next section discusses some of the factors that limited the efficiency of the GIP procedure to indicate probable guessing across the full range of student abilities.

10.2 Reflections on the Circular Constraints in Identifying Guessing in Student Response Data

10.2.1 Calculation of Item Difficulty

Item difficulty is a function of the number of correct responses observed in an item compared to the number of students who have responded to the item. Hence, the number of responses omitted or suppressed becomes a factor in the overall item difficulty calculation, as expressed in Eqn 10.1.

A simple estimate of item difficulty (δ_i) is:

$$\delta_i = \log \frac{[1 - \lambda]}{[\lambda]} \quad \text{Eqn 10.1}$$

where λ is the ratio $\frac{n_i}{N_i}$

n_i is the number of correct responses for item(i), and

N_i is the number of students who have attempted item (i).

When guessing is present, the number of correct responses (n_i) is inflated, on average at approximately the rate of $1/d$ (where d is the number of distractors) for each response in which the difficulty of the item exceeds the ability of the student, which is the item/student interaction region of the item in which random guessing occurs. The literature describes in detail the estimation procedures for item difficulty location (Anderson, 1970, Ghosh, 1995, Molenaar, 1995a, Wilson, 2004). A mathematical feature that characterises the RM is that the estimation of item difficulty is independent of the distribution of ability of the students who have participated in the test (Andrich, 1998; Rasch, 1960).

However, the interaction that engenders guessing is a function of item difficulty and the student ability at the individual student/ item interaction level. The inflation of n_i as a consequence of guessing inflates the difficulty location of the items and consequently impacts on the number of responses included in the calculation of the item difficulty location, the probability of a successful response, and the degree to which the interaction misfits the RM.

In the case of dichotomous items, there is a 2-by-2 matrix of potential response outcomes for any two adjacent items. This contributes to the conditional probability calculation that underpins item difficulty estimation in the RM, as shown in Table 10.1 (Andrich, 1988, p. 27).

Table 10.1
Item Estimation – Conditional Probability of Dichotomously Scored Responses

Item (n)	Item (n+1)	Joint Outcome	Joint Outcome matrix			
Correct (1)	Correct (1)	Both Correct (1,1)	Item		l_i	l_{i+1}
Incorrect (0)	Correct (1)	Incorrect/correct (0,1)		outcome	0	1
Correct (1)	Incorrect (0)	Correct/incorrect (1,0)	l_i	0	0,0	0,1
Incorrect (0)	Incorrect (0)	Incorrect/Incorrect (0,0)	l_{i+1}	1	1,0	1,1

In the Rasch item estimation procedure, the only response patterns that impart information regarding the relative difficulty of the items is the correct/incorrect and incorrect/correct responses of adjacent items. Using Eqn 10.2, the total probability of that subset of responses can then be mathematically derived:

$$\Pr\{(x_{j1} = 1, x_{j2} = 0) | (x_{j1} = 1, x_{j2} = 0) \text{ OR } (x_{j1} = 0, x_{j2} = 1)\} = e^{-\delta_1} / (e^{-\delta_1} + e^{-\delta_2}) \quad \text{Eqn 10.2}$$

Where x_{j1} is a correct/incorrect response by student j to item 1
 x_{j0} is an incorrect/correct response by student j to item 1
 $e^{-\delta_n}$ is the inverse of the difficulty of item 1 and 2 respectively.

Eqn 10.2 shows that estimate of item difficulty independent of student ability (β_j). When guessing is not suppressed, there is contamination of the difficulty estimate of each item. It is logical to assume a direct relationship between the difficulty of the item and the number of occasions in which guessing will be employed, with the more difficult items more greatly impacted by guessing in a test. In the limit, the true value of n_i will be overestimated to the degree $n_i(1 + 0.25 \times (N_i - n_i))$. Hence, there is ‘noise’ around the estimates of item difficulty caused by the presence of guessed items.

10.2.2 Distribution of Item Difficulties

Eqn 10.1 shows the difficulty of an item in terms of the correct/incorrect response ratio whilst Eqn 10.2 shows the relative difficulty of an item as calculated in the RM. In a RM analysis the item difficulty locations are, by default, centred on zero. Hence, when the proportions of correct responses and overall responses change, the relative difficulty locations may change due to the degree to which n_i and N_i change. In the case of suppressing item responses indicated as guessed student/item interactions, both n_i and N_i are reduced. However, it has been shown that the suppression of items observed as likely guesses for harder items is greater than for easier items. Hence, the harder items become harder than initially calculated. In a RM analysis, the item locations are centred on zero and hence the easier items have lower locations on the scale and, overall, the distribution of the item locations on the scale increases as a result. When guessing is not accounted for, the distribution of item locations is contracted in a RM analysis.

10.2.3 Estimation of Student Ability

Chapter 3 described how student abilities are estimated. In the RM there is a one-to-one relationship between a student's raw score and the ability estimate assigned to the response pattern of correct answers. Although the calibration of item difficulty is independent of the ability(s) of the students participating in the assessment, the converse is not the case. The raw score/ability interaction is a direct result of the item difficulty calibrations and estimates.

Eqn. 10.3 provides the definition of the interaction between raw score (r_j), item difficulty, and student ability estimate. This equation is solved iteratively for possible values of ability (β_j) that, given defined values of item difficulties (δ_1 through δ_n), provide natural number equivalents for the potential scores of a test of N maximum score points (items if all questions are dichotomous MC items).

$$r_j = \sum_{i=1}^N \frac{e^{\beta_j - \delta_i}}{1 + e^{\beta_j - \delta_i}} \quad \text{Eqn 10.3}$$

This relationship derived by the iterative process determines the natural number equivalents of possible ability estimates, given the set of difficulties estimated for the set of items that comprise the test. The relationship between the student raw score/ability estimation and the item difficulties within the test gives rise to two further sources of uncertainty in the estimates derived:

1. The inflation of item difficulties impacts the iterative interaction with the possible ability estimates and depresses the relative ability estimate associated with each raw score across the full spectrum of possible values. Hence, at the higher ability level, the higher raw scores are not attributed the full measure of ability on the trait, and at the lower ability level, the lower scores are attributed a degree of inflation in their estimate of ability on the trait.
2. Since the number of items in which guessing may occur is likely to be inversely proportional to the ability of a student, the effect of successful guessing thus has a greater impact on lower-ability students. The fact that these students' raw scores are inflated, relative to their "true" score, causes an overestimation of the ability estimates in the lower-ability students.

11.2.4 Calculation of Item/Student Probability of a Correct Response.

The probability of a correct response by a student of any estimated ability to an item of derived difficulty is a simple application of the RM, as follows:

$$\Pr\{x_{ji}=1|\beta_j, \delta_i\} = e^{\beta_j - \delta_i} / 1 + e^{\beta_j - \delta_i} \quad \text{Eqn 10.4}$$

When the item difficulty estimate has been contaminated by guessing, the value of δ_i is underestimated. Concurrently, the raw score of student j is overestimated due to the positive impact of unconditioned guessing. The interaction of these two factors causes the probability of a correct response, as calculated using the RM, to be overestimated, meaning that the efficacy of using the interaction of individual item difficulty with specific student ability is diminished by this dual impact of a correct guess. The impact is exacerbated by the randomness of the number of successful guesses, as demonstrated in the simulated data in which students with pre-defined low ability in a 40-item test have “randomly guessed” up to 14 of the 39 items that are defined as beyond their ability in the trait.

When a response is a successful random guess, then the ability of the student is effectively inflated in the calculation in Eqn 10.5 by the difference an additional response impacts on the ability estimate of the student. This contaminates all of the resulting measures that derive from the RM.

10.2.5 Calculation of Item/Student Residual

The item/student residual is a measure of the misfit between an individual response to an item by a student of a calculated ability. Given that ability is a continuous variable, and the natural number raw score is a representation of the trait of interest about that difficulty level, there will always be some “noise” about the convergence of the estimated score derived by the RM and the natural number value (1 or 0) that represents success or failure on the specific item. Conceptually, the value of the residual is the difference between the observed response(s) and expected response(s) across all items (Andrich, 1986; Masters, 1982; Wright & Stone, 1979).

The value of the item/student residual is calculated by using the conditional formula below (Andrich, 1988), which results in two possible standardised values.

$$z_{ji} = (x_{ji} - E[X_{ji}]) / \sqrt{V[X_{ji}]}$$

which resolves to

$$\text{for: } X_{ji} = 0: -\sqrt{\pi_{ji}(1 - \pi_{ji})} \quad \text{and}$$

$$\text{for: } X_{ji} = 1: \frac{\sqrt{(1 - \pi_{ji})}}{\sqrt{\pi_{ji}}}$$

where $E[X_{ji}]$ is the probability of a correct response derived from the RM for the calculated values of β_j and δ_i .

These values are contaminated by the influence of correct guessing patterns, and these derived parameters have a circular nature of. The interrelationships discussed above explain to some degree the challenge in using the outputs of the RM and the initial analysis parameters to identify guessing patterns at the individual item/student level.

10.3 Merits of the GIP: Student Scaled Scores and Reported Levels

The limitations articulated in Section 10.2 above help explain the challenges in accurately indicating guessing when using the RM. Despite these limitations militating against the efficiency of indicating guessing in student responses, this study shows that these challenges can be addressed in a psychometrically sound manner that ensures that at least a proportion of the guesses can be removed with confidence. This in turn means that the final measures of student ability are more valid than those obtained without using the GIP procedure, both at the scale and individual student level.

10.3.1 Summary Observation in Simulated Data Sets

In the simulated data, in which the random guesses in each student response pattern were defined, the GIP procedure was applied and the responses identified as guessed were re-coded as missing. The new set of data were then re-analysed; the resultant recovery rate of identified guesses, compared to defined guesses, ranged between 29% and 48%, as shown in Chapter 5, Table 5.12, and reproduced in Table 10.2.

Closer examination of those recovery rates, by partitioning the data into sub-groups, revealed a significant improvement in the identification of guessing (identified and defined) among the lower-ability groups, as shown in quartiles 1 and 2, and deciles 1, 2, and 3, in Table 10.2. It is also noted that in no instance was a correct response identified as a guess by the application of the GIP that was not predefined as a guess.

In the higher-ability regions – quartiles 3 and 4, and deciles 8, 9, and 10 in SIM3 – the recovery rate of “defined” guesses was, on average, less than 6%, except for the SIM5 test. In the lower quartiles (deciles), which report the outcomes of the lower-ability students, the recovery rate improved significantly and, in the lower ability category, at least 50% of the “defined” guesses were “identified”.

Given the relationship between guessing and ability, the GIP functioned most effectively in the region where guessing is most prevalent; that is, among the lower-ability students. Hence, the GIP can be considered to contribute favourably to the refinement of the scale, particularly in areas of increased need.

Table 10.2*Proportion of GIP Identified Guesses by Ability Group Compared to Simulated Data INIT Analysis*

Simulation	Group	Count of defined random Guesses (Table 5.12)	Count of Guesses recovered by GIP _{p=0.6}	Recovery rate Identified vs actual (%)
SIM1, Normal 400,40	Q4, most able	72	0	0.0%
	Q3, able	446	59	13.2%
	Q2, less able	478	195	40.8%
	Q1, least able	775	537	69.3%
	Overall mean	1771	791	44.7%
SIM2, Normal 250,20	Q4, most able	14	0	0.0%
	Q3, able	74	0	0.0%
	Q2, less able	114	20	17.5%
	Q1, least able	158	84	53.2%
	Overall mean	360	119	28.9%
SIM3 Normal 1000,40	Top decile	14	0	0.0%
	Decile 9	98	0	0.0%
	Decile 8	259	0	0.0%
	Decile 7	400	28	7.0%
	Decile 6	481	128	26.6%
	Decile 5	438	177	40.4%
	Decile 4	473	237	50.1%
	Decile 3	591	407	68.9%
	Decile 2	791	502	63.5%
	Decile 1	759	572	75.4%
Overall mean	4304	2051	47.6%	
SIM4, Easy 400,40	Q4, most able	72	0	0.0%
	Q3, able	217	7	3.2%
	Q2, less able	389	150	38.6%
	Q1, least able	633	432	68.2%
	Overall mean	1311	589	44.9%
SIM5, Hard 400,40	Q4, most able	303	0	0.0%
	Q3, able	417	90	21.6%
	Q2, less able	722	409	56.6%
	Q1, least able	722	532	73.9%
	Overall mean	2164	1031	47.6%

10.3.2 Large-Scale Data Sets

The evidence demonstrated in the simulation analyses and in the GIP analyses of the large-scale authentic data set, makes it apparent that the proposed GIP model is an improvement on a process that takes no action to account for guessing.

Tables 10.3 to 10.8 make explicit the reclassification of results for groups of students (see also Chapter 7, Figures 7.4, 7.5, and 7.6, for each of the GIP3A and GIPINIT3B analyses). In each case, the tables present the amount by which student results may be misrepresented by an analysis method that does not consider the influence of guessing in student data. In Table 10.3, the scaled scores have been approximately aligned to indicate at what point of progress through the level the scaled score had located the students on a particular score.

Table 10.3*G4 Mathematics Comparison of Proportions in Levels by Scaled Score and Analysis Phase*

Level	SSMathG4INIT			SSMathG4GIP3A			SSMathG4GIPINIT3B		
	Scaled Sc	Freq	Percent	Scaled Sc	Freq	Percent	Scaled Sc	Freq	Percent
6+	858	79	0.3	919	79	0.3	919	79	0.3
	782	237	0.9	838	237	0.9	838	237	0.9
6	728	428	1.6	777	428	1.6	777	428	1.6
	689	589	2.2	732	589	2.2	732	589	2.2
5	658	874	3.3	693	874	3.3	693	874	3.3
	631	1066	4.1	659	1066	4.1	659	1066	4.1
	606	1251	4.8	628	1251	4.8	628	1251	4.8
4	583	1448	5.5	598	1448	5.5	598	1448	5.5
	561	1597	6.1	569	934	3.6	569	1597	6.1
	539	1816	6.9	540	1437	5.5	540	1816	6.9
	516	2148	8.2	511	1600	6.1	511	2148	8.2
3	494	2562	9.7						
	470	2780	10.6	481	1236	4.7	481	2562	9.7
	445	2939	11.2	450	987	3.8	450	2780	10.6
	416	2698	10.3	417	1482	5.6	417	2939	11.2
2	384	2015	7.7	381	2209	8.4	381	2698	10.3
	344	1157	4.4	341	2453	9.3	341	2015	7.7
1	288	477	1.8	293	2857	10.9	293	1157	4.4
	211	118	0.4	230	2667	10.1	230	477	1.8
Below Level 1				146	2445	9.3	146	118	0.4
	Total	26279	100.0	Total	26279	100.0	Total	26279	100.0

The GIP3A and GIPINIT3B values of the scaled scores of the students were the same as a result of the anchoring process, but the percentages in the levels show the difference as a result of suppressing, compared to not suppressing, the score associated with the GIP indicated guesses. Noticeably, the percentages shown for the GIP3A and GIPINIT3B in the upper levels do not change due to the GIP process not indicating probable guesses in this region. However, at the lower regions there are considerable differences in the frequencies of students at each level. The equivalence of the frequencies between the INIT outcomes and the GIPINIT3B outcomes is due to the INIT raw scores being applied in the GIPINIT3B analyses. However, the scaled scores and level variations reflect the impact of accounting for guessing.

In Table 10.3, the GIP3A and GIPINIT3B outcomes show a group of students who were described “below Level 1” with a scaled score of 146, which is approximately 0.5 s.d. below the cut score for Level 1 (i.e. as scaled score below 200). This level of achievement was not reported in the INIT analysis. This outcome indicates that the content in the test was beyond the ability of the students when guessing was accounted for. Yet, in a simple Rasch analysis, these students would be identified as among the 595 students at Level 1. In contrast, the GIP3A/GIPINIT3B analysis indicates that there were many more students at this level who were misclassified in a simple Rasch analysis. The GIPINIT3B analysis indicates that 6.6% of students were at/or below Level 1, compared to 2.2% identified by the Rasch INIT analysis.

At the higher ability end of the scale, the GIPINIT3B analysis indicates that 1,333 students performed at Level 6 or above, compared to 744 in the INIT analysis. The similarity in the scaled scores and proportions of students in the mid-range ability Levels (3 and 4) reflects the expected achievement for these levels. These outcomes are consistent with those of the simulated data and are predicted by the GIP processes applied.

Table 10.4 aggregates the data of each analysis into three performance groups. The shaded area at the bottom of the table indicates the aggregation of Level 1 and below with Level 2 and represents students who were performing below the expectation of the cohort. These students are considered “at risk” of not achieving the learning outcomes for the cohort. The unshaded area (the aggregation of Levels 3 and 4) represents the students who were “approaching or at” the expected achievement level of the cohort. The shaded area at the top of the table represents students who were achieving above the expected level (the aggregation of Levels 5, 6, and 6+).

Table 10.4

G4 Mathematics Comparison of Percentages in Levels

Level	INIT			GIP3A			GIPINIT3B		
	Frequency	% at level	Total %	Frequency	% at level	Total %	Frequency	% at level	Total %
6+	79	0.3		316	1.2		316	1.2	
6	665	2.5	17.2	1017	3.9	17.2	1017	3.9	17.2
5	3780	14.4		3191	12.1		3191	12.1	
4	7009	26.7	68.5	5419	20.6	34.8	7009	26.7	58.2
3	10979	41.8		3705	14.1		8281	31.5	
2	3172	12.1		4662	17.7		4713	17.9	
1	595	2.3	14.4	5524	21.0	48.0	1634	6.2	24.6
Below 1	0	0.0		2445	9.3		118	0.4	

Note. There are minor rounding errors in aggregating proportions within levels

Table 10.4 highlights the challenge of condensing information in reports. At first glance, the percentages of higher-ability students in Levels 5, 6, and 6+ is the same for each analysis (17.2%). However, the disaggregation into the component levels shows a variation in the percentages in each level. The alignment of the frequencies in the GIP3A and GIPINIT3B outcomes is a function of the reduced capacity of the GIP process to indicate guessing in the higher-ability students when the student ability estimate is high compared to a minor differentiation between the item locations for harder items. However, both of these analyses show an increase in the percentage of students identified as performing at Levels 5, 6, and 6+ compared to the INIT analysis.

By comparison, the variation in the percentage of students in each level for the lower-ability students is significant. The GIP3A analysis (suppressed indicated guesses) displays 48.0% of students in the “at risk” region, with high percentages of that group in or below Level 1. The INIT analysis resulted in the majority of these students achieving the expected level of achievement for the Grade 4 students. The GIPINIT3B analysis displays 24.6% of students in the “at risk” levels, which is approximately 10% more than the INIT

analysis. This outcome shows the negative impact of crediting student with the results of highly probable guesses. The lower-ability students were most advantaged by guessing.

Table 10.5

G8 Mathematics Comparison of Percentages in Levels by Scaled Score and Analysis Phase

Level	SSMathG8INIT			SSMathG8GIP			SSMathG8GIPINIT		
	Scaled Sc	Freq	Percent	Scaled Sc	Freq	Percent	Scaled Sc	Freq	Percent
6+				996	23	0.1	996	23	0.1
	928	23	0.1	900	69	0.3	900	69	0.3
	842	69	0.3	831	133	0.6	831	133	0.6
6	781	133	0.6	781	215	1.0	781	215	1.0
	739	215	1.0	740	348	1.7	740	348	1.7
	706	348	1.7	705	447	2.1	705	447	2.1
5	677	447	2.1	673	643	3.1	673	643	3.1
	652	643	3.1	644	763	3.6	644	763	3.6
	629	763	3.6						
4				617	921	4.4	617	921	4.4
	608	921	4.4						
	587	974	4.6	591	974	4.6	591	974	4.6
	567	1058	5.0	566	871	4.1	566	1058	5.0
	548	1227	5.8	542	868	4.1	542	1227	5.8
3	528	1339	6.4	518	936	4.5	518	1339	6.4
	508	1521	7.2						
	488	1852	8.8	494	899	4.3	494	1521	7.2
	466	2040	9.7	469	874	4.2	469	1852	8.8
2	443	2185	10.4	444	973	4.6	444	2040	9.7
	418	2015	9.6	418	945	4.5	418	2185	10.4
	390	1608	7.7	389	931	4.4	389	2015	9.6
	357	992	4.7	357	1136	5.4	357	1608	7.7
1	314	436	2.1	321	1340	6.4	321	992	4.7
				276	1363	6.5	276	436	2.1
	254	161	0.8	213	1802	8.6	213	161	0.8
Below 1	168	37	0.2						
				124	3533	16.8	124	37	0.2

Tables 10.5 and 10.6 present the results for the Grade 8 students, in the same format as Tables 10.3 and 10.4, respectively. Table 10.5 shows a similar pattern to the Grade 4 Mathematics distributions. The GIPINIT3B analysis reveals more students in both the higher and lower ability achievement levels than the INIT analysis. The variation evident in the scores achieved is an expected outcome of the recalibration of the scale and the “purer” GIP variable against which the students were measured.

The variations in the scaled scores, resulting from these recalibrations, shifted groups of students across the cut score for particular levels and provided a better estimate of their achievement relative to the expected standard. For instance, the 133 students with scaled scores of 781 in the INIT analysis had been classified as Level 5, whereas the same students achieved a recalibrated scaled score of 831 and were classified at Level 6 in the GIPINIT3B analysis.

Table 10.6 displays the movement of students between levels as a consequence of the recalibration of the Grade 8 scale when comparing the INIT outcome to the GIPINIT3B outcome. The notable variation is the proportion of students who were classified as “at risk” in the GIPINIT3B analysis compared to the INIT analysis

Table 10.6

G8 Mathematics Comparison of Percentages in Levels

Level	INIT			GIP3A			GIPINIT3B		
	Frequency	% at level	Total %	Frequency	% at level	Total %	Frequency	% at level	Total %
6+	92	0.4		225	1.1		225	1.1	
6	696	3.3	16.9	1010	4.8	17.0	1010	4.8	17.0
5	2774	13.2		2327	11.1		2327	11.1	
4	6119	36.2		3649	17.4		4598	21.9	
3	8092	38.5	74.7	3691	17.6	34.9	7598	36.1	58.0
2	3036	14.5		3407	16.2		4615	22.0	
1	161	0.8	15.4	3165	15.1	48.1	597	2.8	25.0
Below 1	37	0.2		3533	16.8		37	0.2	

Note. There are minor rounding errors in aggregating proportions within levels

Tables 10.7 and 10.8 present the Grade 4 Science summary statistics. These outcomes reflect the different distribution of item locations and ability estimates shown in Figure 7.6 in Chapter 7. Table 10.7 shows a close relationship between the scaled scores of the INIT and GIPINIT3B values, which maintained the variation in the proportions in levels observed in the previous analyses.

The lowest scale score in the GIPINIT3B analysis was lower than that reported in the INIT analysis, and the highest score in the GIPINIT3B was higher than the INIT analysis. The impact on the final scaled scores and the variation in the percentages in the levels follows a similar pattern to those observed in Grade 4 Mathematics. Despite the Grade 4 Science assessment displaying a more condensed scale in the INIT analysis, the GIP outcomes demonstrate that the students located in the lowest ability level were likely to be performing below the level reported the INIT analysis outcomes.

At the higher end of the scale, the percentage of students in the “better than expected” achievement group is also greater than the comparative INIT group. The GIP3A process and the GIPINIT3B application provided a better discrimination of students across the scale when guessing was considered in the calibration process.

Table 10.7*Grade 4 Science Comparison of Percentages in Levels by Scaled Score and Analysis Phase*

Level	SS SciG4INIT			SS SciG4GIP3A			SS SciG4GIPINIT3B		
	Scaled Sc	Freq	Percent	Scaled Sc	Freq	Percent	Scaled Sc	Freq	Percent
6+				869	69	0.3	869	69	0.3
	826	69	0.3						
6				791	375	1.4	791	375	1.4
	753	375	1.4	735	696	2.7	735	696	2.7
5				694	908	3.5	694	908	3.5
	667	908	3.5	662	1120	4.3	662	1120	4.3
	639	1120	4.3	635	1176	4.5	635	1176	4.5
	615	1176	4.5	610	1258	4.8	610	1258	4.8
				594	1258	4.8	588	1228	4.7
4				575	1228	4.7			
	557	1245	4.8	567	1245	4.8	567	1245	4.8
	540	1238	4.7	547	1238	4.7	547	1238	4.7
	524	1271	4.9	527	1271	4.9	527	1271	4.9
	507	1276	4.9	508	658	2.5	508	1276	4.9
3				491	1393	5.3	489	1393	5.3
				475	1560	6.0	470	1560	6.0
				458	1815	7.0	451	1815	7.0
				441	2049	7.9			
				423	2156	8.3	431	2049	7.9
				403	1973	7.6	410	2156	8.3
2				381	1544	5.9	387	1973	7.6
				357	977	3.7	362	1544	5.9
				328	471	1.8	334	977	3.7
							302	471	1.8
1				292	186	0.7			
							261	186	0.7
							239	52	0.2
Below Level 1				203	1713	6.6	203	52	0.2
	165	29	0.1						
Total	26065	100.0		Total	26065	100.0	Total	26065	100.0

Table 10.8*Grade 4 Science Comparison of Percentages in Levels*

Level	INIT			GIP			GIPINIT		
	Frequency	% at level	Total %	Frequency	% at level	Total %	Frequency	% at level	Total %
6+	69	0.3		69	0.3		69	0.3	
6	1071	4.1	16.7	1071	4.1	21.5	1017	4.1	21.5
5	3204	12.3		4462	17.1		4462	17.1	
4	7516	28.8	70.8	5640	21.6	39.8	6258	24.0	58.4
3	10946	42.0		4746	18.2		8973	34.4	
2	2992	11.5		5139	19.7		4965	19.0	
1	238	0.9	12.5	3028	11.6	38.7	238	0.9	20.1
Below 1	29	0.1		1910	7.3		29	0.1	

10.4 Overall Comments

The Guessing Indication Protocol (GIP) developed during this study provides a method for taking account of the guessing in individual student response patterns in a probabilistic manner. In particular, even though the GIP does not capture 100% of guesses across the full range of ability estimates, the strength of the procedure is that it meets a need in areas where the issue is greatest, namely, the overestimation of students at risk. This chapter has considered the factors that limit the capacity of the GIP's efficiency in identifying guessing; however, these limitations do not diminish the findings of the study, or its potential contribution to providing more accurate information to educational stakeholders. Rather, these limitations provide the agenda for future research directions.

In considering the outcomes in Tables 10.3 to 10.8, in conjunction with those presented of the simulation studies, these outcomes clearly indicate that analyses that do not take account of guessing underestimate the ability of higher-ability students and overestimate the ability of lower-ability students.

The consolidated impact of the suppression of highly probable guessed items in individual student response patterns was to generate a more refined variable that was less contaminated by the noise associated with guessing. A further consequence of this refinement of the variable was to stretch the distribution of the cohort outcomes. Specifically, there was an increase in the ability estimates of the higher-ability students as a consequence of removing the depression caused by false positive-guessed answers in the more difficult items. The identification of the probable guesses in the more difficult items is a significant benefit compared to current practice in that it generates better estimates of the difficulty of those items and contributes to an improvement in the overall scale and estimates of student ability within the scale.

At the lower end of the scale, the overestimated ability of lower-ability students was also recognised when outcomes contaminated by probable guessing were identified. In particular, the presence of guessed items masked the true ability level of the lower-ability students.

10.4.1 Degree of Reclassification

The outcomes shown in Tables 10.3 to 10.8 suggest that stakeholders should first consider the possible degree of reclassification of students indicated by the GIP procedure, which represents the best approximation of the “true” ability estimate that has taken account of the probable guessing. Table 10.4 shows approximately 14% of students in Levels 1 and 2 when the INIT Rasch results are reported. By comparison, the GIP results display approximately 30% in the “highly at risk” Level 1 or below, and approximately 18% at the “at risk” Level 2. These outcomes indicate that approximately 8,960 students were misclassified using the INIT results for the G4 Mathematics cohort.

The GIPINIT analysis shows that almost 2,900 of the lower-ability students were overestimated and misclassified by the INIT analysis. At the higher ability end of the scale of the G4 Mathematics students, the INIT, GIP, and GIPINIT analyses report the same number of students to be in the “beyond expectation” group. However, the disaggregation of the group by level shows that the GIP3A and GIPINIT3B analyses reveal a degree of underestimation of the growth achieved. The INIT analysis displays 297 students in Level 5 with a scaled score of 782 points, while the GIP and GIPINIT analyses show these students as achieving a scaled score of 838 and more likely to be in Level 6.

The potential consequence of these misclassifications is that lower-ability students who need support in achieving the expected curriculum outcomes, as assessed by the test, are inaccurately measured and the unlearned content and skills remain unidentified. Simultaneously, educators may overlook the higher-ability students and fail to provide extension and experiences that challenge their true ability.

Tables 10.5 through 10.8 show similar patterns, with the GIP3A and GIPINIT3B analyses reporting higher percentages of lower-ability students in levels *lower* than those reported by the INIT analysis. The misclassifications impacted up to 3,900 students in the Grade 8 Mathematics analysis and nearly 2,900 students in the Grade 4 Science analysis. In respect of the higher-ability students, approximately 1,000 students are reported *above* the level of the INIT analysis for both Grade 8 Mathematics and Grade 4 Science.

In comparing the GIPINIT3B scale to the INIT scale, these outcomes indicate misclassification in reporting of the lower ability of the order of 10%. These students were systematically overestimated. At the higher end of the scale there is a greater accord between the overall performances reported by the INIT and GIPINIT analyses; however, there is a variation in the percentages reported in each level, which generally underestimates the performance of the higher-ability students.

In general, comparison between the INIT and the GIP outcomes indicates that the students in the mid-range ability deciles were reported at approximately the appropriate level by the outcomes of the INIT analysis; however, the abilities of the lower-ability groups were overestimated and those of the higher-ability groups were underestimated.

10.5 Summary

The introduction to this chapter highlighted the constraints of the RM on the capacity of a process designed to identify guessing in the response patterns of students. The chapter highlights that even though the proposed $GIP_{p=0.6}$ process is not able to identify all instances of probable guessing over the full range of student abilities, it produces a more accurate set of estimates of student ability than does ignoring the influence of guessing in student responses. In the case of lower-ability students, stakeholders who use the currently reported outcomes of the Rasch (INIT) analysis are typically using unreliable data in their considerations. Messick (1989) would suggest that these misclassifications may produce invalid consequential actions.

Chapter 11

Conclusion

11.1 Introduction to the Chapter

The underlying premise of this thesis is that if educational professionals and other stakeholders are to make reliable decisions about large-scale assessments, they should have high-quality, data-driven information that minimises measurement errors and maximises the reliability and validity of those assessments. This issue is significant because educational systems, schools, and individual students use the reported outcomes from large-scale assessments in multiple ways.

This research examined the potential impact of guessing on estimations of item difficulty and student ability. Its particular focus was degree to which student ability may be misclassified in large-scale assessments in which multiple-choice (MC) items are the dominant item type, and then to propose and evaluate a protocol that addresses this issue.

11.2 Summary of the Research

The initial chapters of this thesis explained the research context and outlined the development of Modern Test Theory and the analysis techniques that have evolved to meet the requirements of educational stakeholders. A key feature of reporting to stakeholders is the presumption that assessment outcomes are accurately measured according to stable and reliable scales that compare student achievements. The Rasch Model (RM) conforms to such principles of measurement and is the analysis technique mostly used in the Australian national assessment context, specifically the National Assessment Program (NAP) suite of assessments. However, the RM does not directly account for guessing.

This research involved simulations that were grounded in the assumptions of the RM. They included student response patterns that were contaminated with defined guesses. The simulations clearly demonstrated a degree of bias in the results of the estimations of item difficulty and consequent estimates of student ability and Level of achievement in which the defined guessing was accounted for.

The protocol that was proposed and evaluated in this study is termed the Guessing Indication Protocol (GIP). The GIP was evaluated using a set of small-scale data and three sets of large-scale data. It reduced the bias seen in the simulations and improved the reliability of student ability estimates. The application of the proposed protocol with large-scale data returned outcomes that were consistent with the outcomes of the simulations which validated the process. Hence the GIP process goes some way to addressing the problems articulated in the research question. The penultimate chapters explicated the degree to which reports to stakeholders may be misrepresented in cases where guessing, as indicated by the GIP, is not accounted for.

11.3 Summary of the Key Results

Chapter 2 described the multiple strategies that have been employed to remove the bias caused by guessing in MC items. These range from soft options such as encouraging students not to guess, to more punitive actions such as reducing the achieved score to penalise students for incorrect answers that are assumed to be guesses. In the NAPLAN and many other large-scale assessments, there is no penalty for guessing.

The simulations of Study 1 indicated that a relatively poor student on a “good guessing day” could achieve a score in the mid-range of ability estimates, thus demonstrating the potential for obfuscation of the true learning and ability of the student due to successful guessing. Aligning with previous research (Andrich et al., 2012, 2015), this piece of research confirmed that correct responses to items beyond the calculated ability of the students contributed to the misfit of an item to the RM. These student/item responses contaminated the “purity” of the variable as a unidimensional measure of student achievement. Advancements developed in this study are that the RM can be used to indicate probable guesses in the response pattern of an individual student, and the suppression of the individual probable guesses can improve the scale that accrues from the modification of the data.

Overall, the outcomes observed in Studies 1, 2 and Study 3 have shown that when the RM is used, the quantity of misfit at the individual item/student interaction level can indicate aberrant interactions that are strong indicators of probable guessing. The use of a set of threshold values with respect to the quantum of the misfit, the GIP can be used to indicate item/student interactions that are probable guesses that may be suppressed. The subsequent analysis of the conditioned data in which the indicated aberrant interactions have been suppressed improve both the fit of the data to the RM and the reliability statistics.

A significant effect of the GIP is that it consistently increased the distribution of item difficulties, which increased the consequent distribution of student ability estimates. This meant there was greater differentiation between the ability estimates of the higher-ability students and, in most cases, an improved reflection of the achievements of those students. In addition, the GIP resulted in a reduction in the ability estimates of the lower-ability students that better reflected their achievements when the guessing was accounted for. The increased reliability of the ability estimates of the lower-ability students is a critical and important outcome of this research that has not been evident in previous studies.

Another significant feature of the GIP is that the first phase – GIP3A – redefined the scale of the assessment. This phase took account of the guessing and improved the item difficulty estimates so that they could better reflect the true location of each item. Hence, the revised described scale is potentially more valid and the annotations of the skills that describe the levels ascribed to the scale were more accurate.

Despite the bias in the student ability estimates caused by the reintroduction of the initial response data in the second phase – GIPINIT3B – the students scaled score results and the assignment of levels to students on the revised scale were based on a more accurate and reliable calibration of achievement.

11.4 Implications of the study

A significant implication of this study is that when the RM is used to analyse large-scale data in which guessing is present in the student response patterns, the reported results may be constrained with respect to the range of item difficulties. This is a direct consequence of not accounting for any guessing that is present in the data. This can produce an overestimation of the ability estimates and percentages of students in the lower ability levels and an underestimation of the percentages of students in the higher ability levels.

The GIP appears to function more effectively when there is discrimination between the item difficulties of the MC items in the test. When the items function relatively homogeneously, the measurement scales can be negatively affected in two ways. The first is a contraction of the student ability estimates that are functions of the item locations. Second, the homogeneity of the item locations means there are fewer instances of differences in item location and student/item ability interactions exceeding 1.1 logits, which in turn restricts the number of cases of probable guessing indicated by the GIP procedure. It is recommended that test constructors address this issue to ensure MC items not only transcend the full range of student abilities in a test but also cover a wide range of item difficulties.

Another implication is the impact of targeting of the test to the cohort which can deliver variable distributions of outcomes for the cohort. In a Rasch analysis, the well-targeted test items and student ability estimates cover a significant range of the scale and tend to produce a relatively normal distribution of item locations and student ability estimates. Tests that are too easy tend to underestimate the abilities of the higher-ability students and, to a lesser extent, overestimate the outcomes of lower-ability students. Tests that are too hard also underestimate the abilities of the higher-ability students and significantly overestimate the abilities of the lower-ability students when guessing is ignored. Associated with these issues is the misrepresentation of the achievement levels of students whose scaled scores are impacted by not accounting for guessing.

In recent decades there has been a trend in large-scale assessments to involve multiple test forms with linking items to enable concurrent estimations of item difficulty and student ability (e.g., TIMSS, PISA, NAP-SL, NAP-ICTL, NAP-CC). However, alternative forms of adaptive test design in which student responses determine the items students interact with are now being introduced (e.g., NAPLAN since 2019). For more reliable student ability estimates to be determined in these test formats, it is imperative that when the RM is applied as the analysis technique it is accompanied by a process such as the GIP to negate the contamination of the item difficulties caused by guessing.

Given these findings, it is recommended that a protocol such as the GIP be implemented to assess the impact of probable guessing and to provide more reliable ability estimates upon which to make data-driven decisions in instances where the RM is implemented.

11.5 Limitations of the Study

Although the GIP indicated probable guessing, it functioned differentially across ability levels within a cohort. It did not have a significant impact of the ability estimations of the mid-range ability students. This was due to the discrimination between items difficulty locations, and consequent student ability estimations about the default centralised values in a RM analysis, especially when tests were well targeted to the participating students.

The targeting of the tests – the alignment of the cohort ability with the range of item difficulty – affected the GIP's ability to indicate probable guessing. The basic assumption of the GIP is that misfit is an indicator of probable guessing. Misfit is a function of the accord of the data with the RM, and it is a typical feature of a Rasch analysis when there is poor test targeting it. This limitation of the GIP in instances of poor test targeting is a function of the general characteristics of the test/student interaction, rather than a malfunction of the GIP itself.

In this research, the data for the large-scale component of Study3 were extracted from mixed model test constructs that included both MC and constructed response (CR) items. Only the sub-set of MC items in these tests were analysed which resulted in smaller effective test lengths than observed in the majority of the simulation studies. Consequently, the reliability statistic of the student ability estimates of Study 3 was impacted by the smaller number of items, a phenomenon which was demonstrated in SIM2. However, the overall results were consistent with the outcomes of the simulated data. Since the mixed model test construct reflects current practices in educational assessment the impact of the GIP procedure in improving the reliability and fit of these components of the test result in a more precise estimation of student outcomes in the combined test construct.

As mentioned earlier, the GIP has two phases. The GIP3A phase conditions the data to recalibrate the item difficulties and generates a revised set of student ability estimates, thus producing a more reliable scale and a better fit of the data to the model. In other words, it generates a purer item calibration and student achievement scale. The GIPINIT3B phase re-introduces the misfit by acknowledging that a correct response cannot be discounted at the individual student level, even if it is highly probable that it is a guess. This limitation introduced by the second component of the GIP process is an issue that may be best handled in the reporting of student achievement, which requires more consideration and investigation.

11.6 Recommendations for Further Research

An issue that could be further investigated is the optimum p value to apply to the calculations of the GIP parameters. In this conceptual piece of research a p value of 0.6 was used to demonstrate the principles of the procedure. The implementation of alternative p values could be evaluated in the simulation data to find the optimum value to maximise the indication of a guess in the simulated data.

If the $\Pr(1) = 0.25$ is maintained, the calculated difference between item location and person ability was approximately 1.1 logits. Hence the $p = 0.6$ threshold means approximately 0.7 logits ($1.1 - 0.4$) represents the difference in the learning achieved by the student. Cohen (1985) contends that this relates to

approximately 18 months of learning. If the p value increases the impact is the reduction in the contribution attributed to learning. Hence, the optimal level is not simply a mathematical application of an alternative p value but rather a more complex psychometric issue that requires both quantitative and qualitative research.

Given the Australian context an area that requires further investigation is the interaction of the GIP procedure with the adaptive model now current for the NAPLAN assessments in Reading, Numeracy and Language Conventions. In each of these subjects the adaptive model introduces each test with one-third of the items with relatively easy difficulty to engage the students and assess their potential for items that are more or less challenging as appropriate to their observed performance of the initial item testlet.

However, the NAPLAN Technical Report (2019) notes that in each subject at least 66% of the items of the total test (100% in the case of Grammar and Punctuation) and probably the majority of items in the initial testlet (not specified in public reports) will be multiple choice items. As such the general principles that underpin the GIP procedure are highly relevant.

A revolutionary approach to address the issue of guessing in a RM analysis would be to include the GIP3A outcomes, which take full account of highly likely guessing, in reports to stakeholders. In this study, the GIP3A phase was effective in indicating probable guesses with the lower-ability students, who are likely to be the most advantaged by engaging in guessing. It also consistently generated the highest reliability statistics and produced a better fit of the data to the RM and better estimations of item difficulty and student ability.

Reporting the GIP3A results should, to some extent, neutralise the advantage of guessing and support the soft approach to informing students that guessing has few benefits. Such reporting would also indicate to all stakeholders which items had been suppressed by the GIP3A process and whether a correct response to the item was credited by chance or was a function of student knowledge.

11.7 Next Steps

Reports of student achievements to stakeholders take several forms. In NAPLAN, for example, they are typically public and technical reports for the educational systems, and school- and student-based reports for teachers, students, and parents (<https://www.nap.edu.au>). The school-based reports are usually in an electronic format that allows mining of the data at the item/student interaction level. Future researchers might investigate not only how to introduce the GIP into these reporting regimes, but also how to educate stakeholders to extract, interpret, and implement the associated information.

11.8 Conclusion to the Chapter

The use of MC items in large-scale assessments has many benefits. However, failing to account for guessing is problematic when RM procedures are used to analyse student data, particularly when reports are provided to the students. This study has shown that ignoring guessing when using the RM can lead to overestimating the abilities of lower-ability students and underestimating those of higher-ability students.

Hence, when data-driven decisions are made using outcomes of measurements that ignore guessing, stakeholders themselves will be imprecise in their recommendations and interventions.

The outcomes presented in this piece of research evidence the potential contribution of the GIP process to improve the quality of the data from MC items and the consequent information reported. Hence the process has the potential to assist decision makers in their actions based on more precise information. The outcomes presented support the contention that a process such as GIP is a necessary adjunct to any analysis of MC data when the RM is the chosen analysis methodology.

List of References

- Abu-Sayf, F. K. (1979). The scoring of multiple choice tests: A closer look. *Educational Technology*, 19, 5–15.
- ACARA (2020). Version 3 of National Literacy and Numeracy Learning Progressions. Australian Curriculum, Assessment and Reporting Authority, Sydney.
<https://www.australiancurriculum.edu.au/resources/national-literacy-and-numeracy-learning-progressions/>
- Adams, R. J., & Khoo, S. (1996). Quest: The interactive test analysis system. Camberwell: Australian Council for Educational Research.
- Alnabhan, M. (2002). An empirical investigation of the effects of three methods of handling guessing and risk taking on the psychometric indices of a test. *Social Behavior and Personality*, 30(7), 645–652.
- Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Royal Statistical Society*. Volume 32, Issue 2, July 1970, pp 283-301.
- Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123-140.
- Andersen, E.B. (1995). Residual analysis in the polytomous Rasch model. *Psychometrika*, 60 (3), 375-393.
- Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, 2, 449-460.
- Andrich, D. (1982). An Index of Person Separation in Latent Trait Theory, The Traditional KR-20 Index and the Guttman Scale Response Pattern. *Education Research and Perspectives*. 9:1, 95-104.
- Andrich, D. (1985). An Elaboration of Guttman Scaling with Rasch Models for Measurement. *Sociological Methodology* Vol 15 (1985) pp 33-80. <https://www.jstor.org/stable/270846>
- Andrich, D. (1986). Intellectual development of pre-adolescent and adolescent children from a psychometric perspective. International conference on longitudinal methodology. Budapest, Hungary. September 1986.
- Andrich, D (1988). Rasch Models for Measurement. Sage University Series on Quantative Applications in Social Sciences 68. Beverly Hills, Sage Publications.
- Andrich, D. (1989). Advanced Social and Educational Measurement E444, Unit Materials Semester 2, School of Education. Murdoch University.
- Andrich, D., Marias, I., Humphry, S.M. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple-choice items. *Journal of Educational and Behavioural Statistics*, 37, 417-442.

- Andrich, D., & Marais, I. (2014). Person proficiency estimates in the dichotomous Rasch model when random guessing is removed from difficulty estimates of multiple choice items. *Applied Psychological Measurement*, 36, 432-449.
- Andrich, David., Marais, Ida., Humphry, S.M. (2015). Controlling Guessing Bias in Dichotomous Rasch Model Applied to a Large-Scale, Vertically Scaled Testing Program. *Educational and Psychological Measurement*. 2016, Vol 76(3), 412-435.
- Andrich, D., Sheridan, B., & Lou, G. (2001). RUMM 2020. Rasch Unified Measurement Models. RUMM Laboratory. University of Western Australia. Perth.
- Andrich, D., Sheridan, B. S., & Luo, G. (2013). RUMM2030: An MS Windows computer program for the analysis of data according to Rasch Unidimensional Models for Measurement. Perth, Australia: RUMM Laboratory.
- Australian Curriculum and Reporting Authority. (2019). 2018 NAPLAN Technical Report https://www.nap.edu.au/docs/default-source/default-document-library/2018_naplan_technical_report_full_v1.pdf?sfvrsn=0
- Bandaranayake, R. C. (2008). Setting and maintaining standards in multiple choice examinations: *AMEE Guide* No. 37. *Medical Teacher*, 30, 836–845.
- Bansilal, S., Long, C. & Juan. A. (2019). Lucky Guess? Applying Rasch Measurement Theory to Grade 5 South African Mathematics Achievement Data. *Journal of Applied Measurement*. 2019. 20(2) 206-220.
- Bar-Hillel, M., Budescu, D. & Attali, Y. (2005). Scoring and keying multiple choice test: a case study in irrationality. *Mind & Society*, 4, 3–12.
- Barnard, J.J (2013). Option Probability Theory: A quest for better measures. University of Sydney / EPEC Pty Ltd. Available from http://www.epecat.com/EPEC_Option_Probability_Theory.
- Barnes, L.L. (1988). Correcting for guessing in the one-parameter logistic item response theory model: An investigation with small samples. (January 1, 1988). ETD collection for University of Nebraska - Lincoln. Paper AAI8824912. <http://digitalcommons.unl.edu/dissertations/AAI8824912>
- Ben-Simon, A., Budescu, D. V. & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, 21(1), 65–88.
- Bereby-Meyer, Y., Meyer, Y. & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making*, 15, 313–327.
- Betts, L. R., Elder, T. J., Hartley, J. & Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? *Assessment & Evaluation in Higher Education*, 34(1), 1–15.
- 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* (pp. 397-479). Reading, MA: Addison-Wesley.

- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51,
- Bond, T.G. & Fox, C.M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Second Edition. Lawrence Erlbaum associates, Publishers, Mahwah, New Jersey.
- Bradbard, D. A., Parker, D. F. & Stone, G. L. (2004). An alternate multiple-choice scoring procedure in a macroeconomics course. *Decision Sciences Journal of Innovative Education*, 2(1), 11–26.
- Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.
- Brown, C., Templin, J. & Cohen, A. (2015). Comparing the Two- and Three Parameter Logistic Models via Likelihood Ratio Tests: A Commonly Misunderstood Problem. *Applied Psychological Measurement*. 2015, Vol 39(5) 335-348.
- Budescu, D. & Bar-Hillel, M. (1993). To guess or not to guess: a decision-theoretic view of formula scoring. *Journal of Educational Measurement*, 30(4), 277–291.
- Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, 26(1), 41–50.
- Burton, R. F. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Education*, 36, 805–811.
- Burton, R. F. (2004). Multiple choice and true/false tests: reliability measures and some implications of negative marking. *Assessment & Evaluation in Higher Education*, 29(5), 585–595.
- Bush, M. (1999). Alternative marking schemes for online multiple-choice tests. 7th Annual Conference on the Teaching of Computing, Belfast.
- Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education*, 25(2), 157–163.
- Chiu, T-W & Camilli, G. (2012). Comment on 3PL IRT Adjustment for Guessing. *Applied Psychological Measurement* (2013) 37:76-86.
- Choppin, B. (1983). The Rasch Model for Item Analysis. *CSE Report No 19, 1983*, 1-30.
- Choppin, B. (1985). A fully conditional estimation procedure for Rasch model parameters. *Evaluation in Education: An International Review Series* 9: 29-42.
- Choppin, B. H. L. (1985). A two-parameter latent trait model. *Evaluation in Education*, 9, 43-62.
- Choppin, B. H. (1988). Correction for guessing. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: an international handbook* (pp. 384–386). Oxford: Pergamon Press.
- Cliff, R. (1958). The predictive value of chance-level scores. *Educational and Psychological*

Measurement, 1958, 18, 607 – 616.

- Cohen-Schotanus, J. & van der Vleuten, C. P. M. (2010). A standard setting method with the best performing students as point of reference: practical and affordable. *Medical Teacher*, 36, 154–160.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.
- Davis, F.B. (1964). *Educational measurements and their interpretation*. Belmont, California: Wadsworth 1964.
- Davis, F.B. (1967). A note on the correction for guessing for chance success. *Journal of Experimental Education*, 1967, 35, 42-47.
- Davis, F.B. & Fifer, G. (1959). The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, 1959, 19, 159 – 170.
- Department of Education. (2018). *Through Growth to Achievement (Gonski 2.0). Report of the Review to Achieve Educational Excellence in Australian Schools*. Commonwealth of Australia. March 2018. <https://www.education.gov.au>
- DeMars, C.E. (2007). ‘Guessing Parameter Estimates for Multidimensional Item Response Theory Models. *Educational and Psychological Measurement*. Volume 67 Number 3 433-446.
- Diamond, J. & Evans, W. (1973). The correction for guessing. *Review of Educational Research*, Vol43, No2 (Spring, 1973) 181-191.
- Dochy, F., Kyndt, E., Baeten, M., Pottier, S. & Veestraeten, M. (2009). The effects of different standard setting methods and the composition of borderline groups: A study within a law curriculum. *Studies in Educational Evaluation*, 35, 174–182.
- Downing, S. M., Lieska, N. G. & Raible, M. D. (2003). Establishing passing standards for classroom achievement tests in medical education: a comparative study of four methods. *Academic Medicine*, 78(10), 85–87.
- Downing, S. M. (2004). On guessing corrections. *Medical Education*, 38, 113.
- Dunn, L., Parry, S. & Morgan, C. (2007). Seeking quality in criterion referenced assessment. Retrieved 25 April, 2011, from <http://www.leeds.ac.uk/educol/documents/00002257.htm>.
- Education Council. (2014). *The Adelaide Declaration on National Goals for Schooling in the Twenty-First Century*. <http://www.educationcouncil.edu.au/EC-Publications/EC-Publications-archive/EC-The-Adelaide-Declaration.aspx>
- Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179-197.
- Embretson, S.E., & Reise, S.P. (2000). *Item Response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

- Espinosa, M. P. & Gardezabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, 54(5), 415–425.
- Fisher, R.A. (1935). *Statistical Methods, Experimental Design and Scientific Inference*. Republished Oxford University Press, USA, 1990.
- Foley, B.P. (2016). Getting Lucky: How Guessing Threatens the Validity of Performance Classifications. *Practical Assessment Research & Evaluation*. Volume 21, Number 3, February 2016.
- Fowell, S. & Jolly, B. (2000). Combining marks, scores and grades. Reviewing common practices reveal some bad habits. *Medical Education*, 34, 785–786.
- Frary, R.B. (1969). Elimination of the guessing component of multiple-choice test scores: effect on reliability and validity. *Educational and Psychological Measurement*, 1969, 29, 655 – 680.
- Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement*, 4, 79–90.
- Frary, R. B. (1988). Formula scoring of multiple choice tests (Correction for guessing). *Educational Measurement: Issues and Practice*, 7(2), 33-38.
- Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, 2(1), 79–96.
- Gardner-Medwin, A. R. (1995). Confidence assessment in the teaching of basic science. *Research in Learning Technology*, 3(1), 80–85.
- Gay, L.R., Mills, G.E., & Airasian, P. (2009) *Education Research. Competencies for Analysis and Applications*. 9th Ed. Pearson Education Inc., Upper Saddle River, New Jersey.
- Goldstein, H. (1979). Consequences of using the Rasch model for educational measurement. *British Journal of Educational Research*, 5, 211 – 220.
- Goldstein. H. (2011). *Multilevel Statistical Models*. Wiley & Sons, West Sussex. U.K.
- Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statistics and Probability Letters* 23: 165–170.
- Guo, H., Rios, J., Haberman, S., Liu, O. L., Wang, J. & Paek, I. (2016). A New Procedure for Detection of Students' Rapid Guessing Responses Using Response Time. *Applied Measurement in Education* 29(3). April 2016
- Guttman, L.A. (1944). A basis for scaling qualitative data. *American Sociological Review*, 91, 139–150.
- Guttman, L.A. (1950a). The basis for scalogram analysis. In Stouffer, S.A., Guttman, L.A.,
- Guttman, L. (1950b). The problem of attitude and opinion measurements. In S.A. Souffer and others (Eds), *Measurement and Prediction*. New York.
- Guttman, L. (1954). The principle components of scalable attitudes. In P.F. Lazarus (Ed.) *Mathematical*

Thinking in the Social Sciences. New York. Free Press.

- Hambleton, R.K. (1982). Item Response Theory: The Three-Parameter Logistic Model. *CSE Report No 220. Centre for the Study of Evaluation*, University of California, Los Angeles.
- Hambleton, R.K. & Cook, L.L. (1977). Latent Trait Models and Their Use in the Analysis of Educational Test Data. *Journal of Educational Measurement*. Volume 14, No 2, 75-96. Summer 1977.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hammond, E. J., McIndoe, A. K., Sansome, A. J. & Spargo, P. M. (1998). Multiple-choice examinations: adopting an evidence-based approach to exam technique. *Anaesthesia*, 53, 1105–1108.
- Han, K.T. (2012). Fixing the c Parameter in the Three-Parameter Logistic Model. *Practical Assessment, Research and Evaluation*; Vol 17, Number 1, January 2012.
- Harper, R. (2003). Correcting computer-based assessment for guessing. *Journal of Computer Assisted Learning*. (2003) 19, pp.2-8.
- Harris, D. (1989). Comparison of 1-, 2-, and 3-Parameter IRT Models. *Educational Measurement: Issues and Practice*. Spring 1989. 35-41.
- Hattie, J. (2003). Teachers Make a Difference What is the research evidence? *University of Auckland*. Paper for Australian Council for Education Research. October 2003.
- Hattie, J. (2017). Hattie's 2017 Updated List of Factors Influencing Student Achievement. <http://www.evidencebasedteaching.org.au/hatties-2017-updated-list/>
- Hendrickson, G.F. (1971). The effect of differential option weighting on multiple-choice objective tests. Report No.3, *The Center for the Study of Social Organisation of Schools*, The John Hopkins University, 1971.
- Humphry, S.M. (2010). Modelling the effects of person group factors on discrimination. *Educational and Psychological Measurement*, 70, 215 – 231.
- Humphry, S.M. (2015). Using a Rasch Model to Account for Guessing as a Source of Low discrimination. *Journal of Applied Measurement*, 16(2), 193 – 203.
- Jennings, S. & Bush, M. (2006). A comparison of conventional and liberal (free-choice) multiple-choice tests. *Practical Assessment, Research & Evaluation*, 11(8).
- Jia, B., Zhang, X. & Zhu, Z. (2019). A short note on aberrant responses bias in item response theory. *Frontiers in Psychology*, Vol 10, Jan 31, 2019. ArtID:43.
- Kan, Adnan & Bulut Okan (2015.) Examining the Language Factor in Mathematics Assessments. *Journal of Education and Human development*. March 2015, Vol4, No.1, 133 -146.
- Karandikar, R. L. (2010). On multiple choice tests and negative marking. *Current Science*, 99(8), 1042– 1045.

- Kline, T. J. (2005). *Modern test theory: assumptions, equations, limitations, and item analyses*. In *Psychological testing: A practical approach to design and evaluation (pp. 107-166)*. SAGE Publications, Inc., <https://www.doi.org/10.4135/978148338569>
- Kubinger, K.D., Holocher-Ertl, S., Reif, M., Hohensinn, C. & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment*, 18(1), 111–115.
- Kurz, T. B. (1999). A review of scoring algorithms for multiple-choice tests. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.
- Lee, Y-H. & Yue, J. (2014). Using response time to investigate students' test-taking behaviours in a NAEP Computer-Based Study, *Large-scale Assessments in Education*, v2 Article 8, 2014.
- Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement*, 17 (2), 28-30.
- Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6, 83-102.
- Linacre, J. M., and B. D. Wright. (1994). (Dichotomous mean-square) chi-square fit statistics. *Rasch Measurement Transactions* 8: 360.
- Linacre, J. M., & Wright, B. D. (2000). *WINSTEPS: A Rasch computer program*. Chicago: MESA Press.
- Lissitz, R.W. (ed). (2009) *The Concept of Validity. Revisions, New Directions and Applications*. Information Age Publishing Inc, Charlotte, NC.
- Lord, F.M. (1963). Formula scoring and validity. *Educational and Psychological Measurement*, 1963, 23, 663 – 672.
- Lord, F.M. (1964). The Effect of Random Guessing on Test Validity. *Educational and Psychological Measurement*, Vol. XXIV, No4, 1964, 745-747.
- Luce, R.D. (1959). *Individual Choice Behaviours: A Theoretical Analysis*. New York: J. Wiley.
- Macquarie University. (1981). *The Macquarie Dictionary, Revised Edition 1985*. The Macquarie Library. Griffin Press Ltd, Netley. South Australia.
- Marais, I. (2015). Implications of removing random guessing from Rasch item estimates in vertical scaling. *Journal of Applied Measurement*, 16, 113-128.
- Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement*, 51, 315-327. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0263224114000645>
- Masters, G.N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, 25(1), 15-29.
- Mattson, D. (1965). The effects of guessing on the standard error of measurement, and the reliability of

test scores. *Educational and Psychological Measurement*, 1965, 25, 727 – 730.

McDonald, R.P. (2013). Psychology, Psychological Methods and Measurement. The Oxford Handbook of Quantitative Methods in Psychology, Vol 1. Ed. T.D Little. Oxford University Press USA. 2013.

McLeod, S. A. (2019). What a p -value tells you about statistical significance. *Simply psychology*: <https://www.simplypsychology.org/p-value.html>

McLeod, S. A. (2019). What does effect size tell you? *Simply psychology*: <https://www.simplypsychology.org/effect-size.html>

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Meaning and Values in Test Validation: *The Science and Ethics of Assessment*, *Education Researcher*, Vol 18. No.2, pp5-11.

Messick, S. (1989b). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Michaelidies, M.P., Ivanova, M., Nicolaou, c. (2020). The Relationship between Response-Time Effort and Accuracy in PISA Science Multiple Choice Items. *International Journal of Testing*. Volume 20, 2020 – Issue 3, p187-205.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88, 355-383.

Ministerial Council on Education, Employment, Training and Youth Affairs. (2008). *Achieving the Educational Goals for Young Australians*. Melbourne. 2008.
http://www.curriculum.edu.au/verve/resources/National_Declaration_on_the_Educational_Goals_for_Young_Australians.pdf

Mislevy, R. J. (1992). Linking educational assessments: Concepts, issues, and prospects. Princeton, NJ: Educational Testing Service.

Molenaar, I. W. (1995a). *Estimation of item parameters*. In *Rasch Models, Foundations, Recent Developments and Applications*, ed. G. H. Fisher and I. W. Molenaar, 39–51. New York: Springer.

Moss, E. (2001). Multiple-choice questions: their value as an assessment tool. *Current Opinion in Anaesthesiology*, 14, 661–666.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement*, 17(2), 6-12.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational measurement. *Review of Research in Education*, 30, 109–162.

Muijtjens, A. M. M., van Mameren, H., Hoogenboom, R. J. I., Evers, J. L. H. & van der Vleuten, C. P. M. (1999). The effect of a ‘don’t know’ option on test scores: number-right and formula scoring compared. *Medical Education*, 33, 267 – 275.

- NAEP computer-based study (2014). *Large-scale Assessments in Education* volume 2, Article number: 8, 2014.
- Ng, A. W. Y. & Chan, A. H. S. (2009). Different methods of multiple-choice test: implications and design for further research. *Proceedings of the International MultiConference of Engineers and Computer Scientists 20*, Vol. 2, Hong Kong.
- Norcini, J. J. (2003). Setting standards on educational tests. *Medical Education*, 37, 464 – 469.
- Obinne, A.D. (2012). Using IRT in Determining Test Item Prone to Guessing. *World Journal of Education*, Vol 2
- Park, R., Pituch, K.A., Kim, J., Dodd, B.A., Chung, H. (2015). Marginalized Maximum Likelihood Estimation for the 1PL-AG IRT Model. *Applied Psychological Measurement*, 2015, Vol 39(6) 448 – 464.
- Paek Insu, Xu Jie, Lin Zhongtian. (2019). Detection Rates of the M2 Test for Nonzero Lower Asymptotes Under Normal and Nonnormal Ability Distributions in the Applications of IRT. *Applied Psychological Measurement 2019*, Vol. 43(1) 84–88.
- Paek, I. (2015). An Investigation of the Impact of Guessing on Coefficient α and Reliability. *Applied Psychological Measurement*, 39(4) 264 – 277.
- Prieto, G. & Delgado, A. R. (1999). The effect of instructions on multiple-choice test scores. *European Journal of Psychological Assessment*, 15(2), 143 – 150.
- Pendrill, L. (2014). Man as a measurement instrument [Special Feature]. NCSLi Measure: *The Journal of Measurement Science*, 9(4), 22-33. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/19315775.2014.11721702>
- Prihoda, T. J., Pinckard, R. N., McMahan, C. A. & Jones, A. C. (2006). Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *Journal of Dental Education*, 70(4), 378 – 386.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press. Copenhagen, Denmark: Danmarks Paedogiske Institut. Retrieved from www.rasch.org/books.htm.
- Rasch Measurement Transactions: www.rasch.org/rmt/rmt223d.htm
- Rizopoulos, D. (2006). An R package for latent variable modeling and item response theory analyses, *Journal of Statistical Software* 17: 1–25.
- Rogers, H. J. (1999). Guessing in multiple choice tests. In: Masters, G. N. & Keeves, J. P. (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 235 – 243). Amsterdam: Pergamon.
- Sabers, D.L & Feldt, L.S. (1968). An empirical study of the effect of the correction for chance success on the reliability and validity of an aptitude test. *Journal of Educational Measurement*, 1968, 5, 251-258.
- Sabers, D.L & White, G.W. (1969). The effect of differential weighting of individual item responses on

- the predictive validity and reliability of an aptitude test. *Journal of Educational Measurement*, 1969, 6, 93-96.
- Samejima, F. (1973). A comment on Birnbaum's three parameter logistic model in latent trait theory. *Psychometrika*, 38, 221-233.
- San Martin, E., del Pino, G., and DeBoeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, 30, 183 – 203.
- Setzer J.C., Wise, S.L., van den Heuvel, J.L. & Ling G. (2013). An Examination of Examinee Test-Taking Effort on a Large-Scale Assessment. *Applied Measurement in Education*, 26:1, p34-49.
- Shepard, L.A. (1993). "Evaluating Test Validity." In L. Darling-Hammond (Ed.), *Review of Research in Education*, Vol. 19. Washington, DC: AERA.
- Socha, A. & DeMars, C.E. (2013). A Note on Specifying the Guessing Parameter in ATFIND and DIMTEST. *Applied Psychological Measurement* 37(1) 87-92.
- Smith, R.M. (1991). IPARM: Item and person analysis with the Rasch Model. Chicago: MESA Press.
- Thayn, S. (2011). An evaluation of multiple choice test questions deliberately designed to include multiple correct answers. Retrieved 28 June, 2011, from <http://contentdm.lib.byu.edu/ETD/image/etd4168.pdf>.
- Thurstone, L.L. (1927). A law of comparative judgement. *Psychological Review*, 34, 273-286.
- Thurstone, L.L. (1929). The Measurement of Psychological Value. In T.V. Smith and W.K. Wright (Eds.), *Essays in Philosophy by Seventeen Doctors of Philosophy of the University of Chicago*. Chicago: Open Court.
- Thurstone, L.L. (1959). *The Measurement of Values*. Chicago: The University of Chicago Press.
- Tognolini, J. (1989). Psychometric Profiling and Aggregating of Public Examinations at the Level of Test Scores. *Doctoral Thesis, June 1989*, Murdoch University.
- Tindal, G. & Haladyna, T.M (ed). (2002) *Large-Scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation*. Lawrence Erlbaum Associates, Mahwah, New Jersey.
- Tognolini, J. & Davidson, M. (2012). Assessment, standards-referencing and standard setting. In M.M.C. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific*. New York: Springer. [http://link.springer.com.ezproxy1.library.usyd.edu.au/book/10.1007%2F978-94-007-4507-0Links to an external site.](http://link.springer.com.ezproxy1.library.usyd.edu.au/book/10.1007%2F978-94-007-4507-0Links%20to%20an%20external%20site)
- Traub, R. E., Hambleton, R. K. & Singh, B. (1969). Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. *Educational and Psychological Measurement*, 29, 847 – 861.
- Traub, R.E. (1983). A-priori considerations in choosing an item response model. In R.K. Hambleton (ed.) *Applications of item response theory*. Vancouver: Educational Research Institute of British Columbia.

- Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer
- van der Vleuten, C. P. M. (2010). Setting and maintaining standards in multiple choice examinations: Guide supplement 37.1 – Viewpoint. *Medical Teacher*, 32, 174 – 176.
- Victorian Curriculum and Assessment Authority. (n.d.). Using assessment data. http://usingassessmentdata.vcaa.vic.edu.au/naplan/tut1_1/mod1.aspx#
- Waller, M.I. (1974). Removing the effects of Random Guessing from Latent Trait Ability Estimates. *Educational Testing Service*, New Jersey. A paper presented at the Annual Convention of the American Educational Research Association, Chicago, Illinois, April 19, 1974.
- Waller, M. I. (1989). Modelling guessing behaviour: A comparison of two IRT models. *Applied Psychological Measurement*, 13, 233-242.
- Waterbury, G.T. & Mars, C.E. (2019). The Effects of Probability Threshold Choice on an Adjustment for Guessing using the Rasch Model. *Journal of Applied Measurement*. 2019, Vol 20 issue 1, p1-12. 12p.
- Wilson, M. & Engelhard, G., Jr (2000). Objective Measurement: Theory Into Practice. *Rasch Measurement Transactions*, Volume 5, 2000, 14:2 p.744
- Wilson, M. (2004). *Constructing Measures: An Item Response Theory Approach*. Erlbaum, Mahwah, NJ.
- Wright, B.D. (1977). Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement*, 14, 97-116.
- Wright, B.D. (2008). What is the “Right” Test Length? *Rasch Measurement Transactions*: www.rasch.org/rmt/rmt61.htm
- Wright, B.D., & Masters, G.N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.
- Wright, B.D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29, 23-48.
- Wright, B. & Stone, M. (1979). *Best Test Design*. MESA Press: Chicago, IL.
- Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest: Generalised item response modelling software*. Camberwell: Australian Council for Educational Research
- Yen, W.M., Burket, G.R., & Sykes, R.C. (1991). Nonunique solutions to the likelihood equations for the three parameter logistic model. *Psychometrika*, 56, 39-54.
- Zimmerman, D. W. & Williams, R. H. (2003). A new look at the influence of guessing on reliability of multiple-choice tests. *Applied Psychological Measurement*. 27(5), 357 – 371.

APPENDIX A

Detail of the Data Construct of Study 1: the Simulated Data

A.1 Simulation Algorithm Conceptualisation

A.1.1 A Guttman-like Arithmetic Progression

The algorithm assumes that a set of items can be created with a defined percentage correct (facility) that defines the number of students that will correctly answer the item based on their ability. Hence it is possible to simulate a 40-item unidimensional test in which the range of facilities can be defined uniformly as an arithmetic progression with Term1 defined as 90% and the difference being -2%.

Hence $a = 90$, $n = 40$ and $d = -2$. Each of the 40 terms (T_n) are defined by $T_n = a + (n-1)d$ as percentage of theoretical correct answers for the sample of the students in a Guttman-like response pattern. In the algorithm the number of students is a random number, but for the sake of consistency of the comparisons of outcomes, the simulated data for Simulation1 was defined to have 400 students.

Given the arithmetic progression defined for Simulation 1(SIM1) data T_{20} (Item 20) will have a defined percentage correct of 52% and therefore have an expected number of correct responses defined as $52\% * 400 = 208$, with those responses accruing to students whose ability is equal to or greater than the difficulty of Item 20.

Table A.1

Extract of Simulated Data (SIM1)

	Easier				Harder			
Person	Item 1	Item 2	Item 3	Item 38	Item 39	Item 40	X_j
Case 1	0	0	0		0	0	0	0
Case 2	0	0	0		0	0	0	0
Case 6	1	0	0		0	0	0	1
Case 200	1	1	1		0	0	0	56
Case 201	1	1	1		0	0	0	56
Case 399	1	1	1		1	1	0	39
Case 400	1	1	1		1	1	1	40
Facility	90.0%	88.0%	86.0%		16.0%	14.0%	12.0%	
Increasing Difficulty								

Given the starting point for the research study is the impact of guessing in an assessment, the initial simulated data set is a pure Guttman scale pattern that is contaminated in two ways;

1. in order to replicate the presence of random errors (e.g. through carelessness and/or misconceptions) amongst those students who have scored very well, some random errors have been impregnated into the data set as incorrect answers;

- in order to replicate the presence of random guessing in the region ‘beyond’ the “ability” of the student, a random number generator has embedded responses in the “incorrect” region of each student’s responses pattern.

Assuming that in a typical multiple choice item of four distractors it is reasonable that one in four random guesses will be correct. The data are ‘contaminated’ by a specifically defined value that can be coded as a correct answer and defined as a correct random guess.

A.1.2 The Data Construct

Table A.2 is an overview of a simulated data set construct and explained below.

Table A.2

Assumptions Regarding Construct of Simulated Data Matrix

Assumptions -															
Item%corr	90%	88%	58%	56%	54%	52%	50%	20%	18%	16%	14%	12%	
EdGuess(1)	1%	1%	1%	1%	1%	1%	2%	2%	2%	2%	2%	3%	
Item%(1)	91%	89%	59%	57%	55%	53%	52%	22%	20%	18%	16%	15%	
Guess(1)	2%	3%	10%	11%	11%	12%	12%	19%	20%	20%	21%	21%	
Observed(1)	93%	91%	69%	68%	66%	65%	64%	42%	40%	39%	37%	36%	
key	1	1	1	1	1	1	1	1	1	1	1	1	40
Data															
ID	Q01	Q02	Q17	Q18	Q19	Q20	Q21	Q36	Q37	Q38	Q39	Q40	RSc

Note 1. The row *Item%corr* is the defined facility or relative ‘difficulty’ of the item. As mentioned, it is an arithmetic progression of 40 items with the difference between each item -2%. With 400 students the expected number of correct responses (score 1) for item Q01 will be 360.

Note 2. The row *EdGuess(1)* is defined as an estimate of the ‘zone of uncertainty’ within which individual student’s ability on the trait is approximately equal to the difficulty of the item and hence the response will return a random ‘1’ or ‘0’ on 50% of occasions.

Note 3. The row *Item%(1)* is the sum of *Item%corr* and *EdGuess(1)* and represents the inflation of the true facility due to ‘Educated Guessing’.

Note 4. The row *Guess(1)* is the expected percentage of correct guesses calculated as one-quarter of the number of potential incorrect answers (Total items – Correct and Educated Guesses).

Note 5. The row *Observed(1)* is the expected percentage of correct responses being the sum of *Item%corr* with *EdGuess(1)* and *Guess(1)*.

The construct assumes that every student attempts every item, and the overall test has three major components for each student;

- A ‘zone’ in which the student’s ability is greater than the item difficulty with a score of ‘1’.
- A zone in which the student’s ability is approximately equal to the item difficulty in which the student’s will be randomly assigned a code of ‘5’ a score of ‘1’, or assigned a code of ‘6’ which is scored as ‘0’ (incorrect). This was termed ‘the zone of uncertainty’.
- A zone in which the student’s ability is less than the item difficulty with responses coded from ‘7’ to ‘10’ for which a ‘9’ is defined as a correct guess (scored ‘1’) and other responses (7, 8, and 10) are coded as incorrect (‘0’).

This construct provides three potential analyses depending upon the coding of the responses that are in each zone for each student;

1. items for which the difficulty of the item exceeds the ability of the student the random guessing responses of ‘9’ can be coded as correct – score ‘1’ ($Pr(9) = \frac{1}{4}$), and all other responses are coded as incorrect (‘0’) which reflects the current typical analysis methodology in which there is no accounting for guessing;
2. items for which the difficulty exceeds the ability the guess responses (‘9’) can be coded as incorrect (‘0’) to confirm compliance with the Guttman scale conditions; and
3. items that are defined as guesses (code 9) are suppressed and coded in the data set as ‘missing data’ so that they are not included in the calibration of item difficulty.

Table A.3 below is an extract of the simulated data set comprising multiple choice items (4-options) each dichotomously scored.

Table A.3
Abbreviated Components of the SIM1 Data Structure

Assumptions -																			
Item%corr	90%	88%	58%	56%	54%	52%	50%	20%	18%	16%	14%	12%					
EdGuess(1)	1%	1%	1%	1%	1%	1%	2%	2%	2%	2%	2%	3%					
Item%(1)	91%	89%	59%	57%	55%	53%	52%	22%	20%	18%	16%	15%					
Guess(1)	2%	3%	10%	11%	11%	12%	12%	19%	20%	20%	21%	21%					
Observed(1)	93%	91%	69%	68%	66%	65%	64%	42%	40%	39%	37%	36%					
key	1	1	1	1	1	1	1	1	1	1	1	1					40
Data																			
ID	Q01	Q02	Q17	Q18	Q19	Q20	Q21	Q36	Q37	Q38	Q39	Q40	RSc	TRUE (1)	EG (1)	RG(1)	
S001	1	1		1	1	1	1	1		1	1	1	1	1	39	38	1	0	
S002							1	1		1	1	1	1	6	38	38	0	0	
S003							1	1		1	1	1	1	6	38	38	0	0	
S004							1	1		1	1	1	1	1	40	39	1	0	
S030	1	1		1	1	1	1	1		1	1	6	6	5	37	36	1	0	
S031	1	1		1	1	1	1	1		1	5	5	6	6	37	35	2	0	
S032	1	1		1	1	1	1	1		4	5	6	6						
S033	1	1		1	1	1	1	1		1	6	6	6						
S079	1	1		1	1	1	1	1		9	7	8	9						
S080	1	1		1	1	1	1	1		8	9	8	9	7	33	30	1	2	
S081	1	1		1	1	1	1	1		9	10	7	9	8	35	30	3	2	
S082	1	1		1	1	1	1	1		7	9	7	7	10	34	30	2	2	
S217	1	1		1	1	1	6	6							23	18	0	5	
S218	1	1		1	1	1	6	5							24	18	1	5	
S219	1	1		1	1	1	6	6							22	18	0	4	
S398	4	9		8	9	7	9	8		10	8	9	9	7	11	0	0	11	
S399	2	9		7	9	9	10	8		8	9	9	7	7	13	0	0	13	
S400	3	7		9	8	9	7	7		8	9	9	9	8	14	0	0	14	
															\bar{x}	24.6	18.9		
															sd	8.5	11.1		

Taking item 20 as an example (shaded blue) a defined Item % correct of 52% with an adjustment for 1% for Educated Guesses means that 47% of the target sample ‘should’ score zero for this item. However, given that the probability of a correct guess in a four-distractor multiple choice item is $\frac{1}{4}$ then the expected percentage of correct guess is $\frac{1}{4}$ of 47 = 11.75% (rounded 12%).

This is the logic that has been attributed to the calculation of the values assigned to the row *Guess(I)*. These are the response that contaminate the Guttman scale and add nothing to the understanding of the characteristics of the scale or the ability of the participants in the test to measure the scale.

The row *Observed(I)* represents the sum of *Item(I)* and *Guess(I)* and shows the expected number of correct responses for this item. This value is compared to the number recovered by the application of the algorithm given the random nature of the assignment of codes '7' to '10' in the zones that represent difficulty beyond ability for each test-taker.

Note for Item 20 with a defined percent correct of 52% the expected Observed percent correct is 65% and for the most difficulty item (Item 40) the defined percent correct of 12% is inflated to 36% due to the guessing factor. Table A.4 provides an extract of a data set SIM1.

Table A.4
Data Extract of SIM1 Data Structure

ID	Q01	Q02	Q17	Q18	Q19	Q20	Q21	Q36	Q37	Q38	Q39	Q40	RSc
S001	1	1		1	1	1	1	1		1	1	1	1	5	39
S002	1	1		1	1	1	1	1		1	1	1	1	6	38
S003	1	1		1	1	1	1	1		1	1	1	1	6	38
S004	1	1		1	1	1	1	1		1	1	1	1	5	40
S031	1	1		1	1	1	1	1		1	5	5	6	6	37
S032	1	1		1	1	1	1	0		1	5	6	6	6	36
S033	1	1		1	1	1	1	1		1	6	6	6	5	36
S079	1	1		1	1	1	1	1		9	7	8	9	10	34
S080	1	1		1	1	1	1	1		8	9	8	9	7	33
S081	1	1		1	1	1	1	1		9	10	7	9	8	35
S082	1	1		1	1	1	1	1		7	9	7	7	10	34
S217	1	1		1	1	1	6	6		9	8	7	7	9	23
S218	1	1		1	1	1	6	5		7	10	7	8	9	24
S219	1	1		1	1	1	6	6		9	9	7	10	10	22
S398	4	9		8	9	7	9	8		10	8	9	9	7	11
S399	2	9		7	9	9	10	8		8	9	9	7	7	13
S400	3	7		9	8	9	7	7		8	9	9	9	8	14

There are three component sections to these data, together with the aggregated score Initial Raw Score (RSc).

Component 1 – Mastery

For each student there is a string of '1's that represent the correct responses an indication that, under the Guttman conditions, students have mastery of the skills and/or knowledge implicit within the items as it contributes to the scale.

Student S030 has mastery of all the items up to and including Item 37, whilst student S217 has a lower ability on this scale with mastery in the Guttman pattern up until and including Item 19.

Note that student S032 has a result of '0' for Item 21 despite evidence that he/she has mastered the skill up to and including Item 36. These random incorrect responses in the Guttman patterns for most students represent the random errors that may accrue due to misconceptions, misinterpretations and carelessness often evidenced in test papers in which capable student answer easy items incorrectly.

Component 2 – The ‘Zone of Uncertainty’

For each student there is an area in which the student ability is about the same as the item difficulty which is termed, for the purpose of this work, the zone of uncertainty. In this area of variable width students may achieve a correct or incorrect response. Random codes of ‘5’ and ‘6’ have been assigned in the ‘zone of uncertainty’ for which a code of ‘5’ is a correct ‘educated guess’ and a code of ‘6’ is an incorrect ‘educated guess’. Student S218 displays one incorrect response and one correct response. This region has been included in an attempt to replicate the observations of student responses in the field.

Component 3 – Random Guessing Zone.

The Random Guessing Zone represents the region beyond the ability of each student relative to the difficulty of the items that have been ordered by increasing difficulty. Random codes of ‘7’, ‘8’, ‘9’, and ‘10’ have been assigned with the code ‘9’ representing a correct response. In the scored data this will result in credit of one mark for that item for which the student has incomplete or little knowledge.

The Raw Score, which in the Guttman scale defines the response pattern, and in a Rasch model is a sufficient statistic for the estimation of student ability, is the sum of the correct answers (code 1), the correct Educated Guesses (code 5) and the correct Random Guesses (code 9).

Table A.5

Extract of Data Structure SIM1 Showing Summary Scoring Structure

ID	Q01	Q02	Q17	Q18	Q19	Q20	Q21	Q36	Q37	Q38	Q39	Q40	RSc	TRUE (1)	EG (1)	RG(1)
S001	1	1		1	1	1	1	1		1	1	1	1	5	39	38	1	0
S002	1	1		1	1	1	1	1		1	1	1	1	6	38	38	0	0
S003	1	1		1	1	1	1	1		1	1	1	1	6	38	38	0	0
S004	1	1		1	1	1	1	1		1	1	1	1	5	40	39	1	0
S031	1	1		1	1	1	1	1		1	5	5	6	6	37	35	2	0
S032	1	1		1	1	1	1	0		1	5	6	6	6	36	35	1	0
S033	1	1		1	1	1	1	1		1	6	6	6	5	36	35	1	0
S079	1	1		1	1	1	1	1		9	7	8	9	10	34	30	2	2
S080	1	1		1	1	1	1	1		8	9	8	9	7	33	30	1	2
S081	1	1		1	1	1	1	1		9	10	7	9	8	35	30	3	2
S082	1	1		1	1	1	1	1		7	9	7	7	10	34	30	2	2
S217	1	1		1	1	1	6	6		9	8	7	7	9	23	18	0	5
S218	1	1		1	1	1	6	5		7	10	7	8	9	24	18	1	5
S219	1	1		1	1	1	6	6		9	9	7	10	10	22	18	0	4
S398	4	9		8	9	7	9	8		10	8	9	9	7	11	0	0	11
S399	2	9		7	9	9	10	8		8	9	9	7	7	13	0	0	13
S400	3	7		9	8	9	7	7		8	9	9	9	8	14	0	0	14
															\bar{X}	24.6	18.9	
															sd	8.5	11.1	

Table A.5 is the deconstruction of the way in which a Raw Score is computed for a range of students.

This table includes column *RSc* – the observed raw score achieved by the student, *True(1)* the defined True score defined under a Guttman structure with some random error to represent ϵ_j , a column *EG(1)* that represents the correct Educated Guesses attributed as a random code of ‘5’ in the zone of uncertainty, and a column *RG(1)* to represent the number of ‘Correct’ Random Guesses assigned to each student through a code of ‘9’ in the zone beyond his/her ability.

Hence student S031 has a raw score of 37 obtained through 35 marks from correct answers as a result of knowledge and two marks due to the random code ‘5’ in the ‘zone of uncertainty’.

Student S218 has a raw score of 24, constructed of 18 correct answers, one educated guess and five ‘correct’ random guesses, whilst student S400 has a raw score of 14 made up entirely of ‘correct’ random guesses.

These the data allow for observations to be made about the parameters that are generated from a Rasch analysis to provide a base line for models that may be discernible to identify guessing in student response patterns and hence by accounting for these violations of the requirements of the RM improve the quality of the item calibration and the accuracy of variable of interest.

Table A.6 below is an extract of 80 sets of statistics from Simulation 1 (SIM1). The items have been ordered from easiest to hardest as annotated by the difficulty label (δ) in Row 1.

Student item interaction statistics are sorted by Student ID (ID) – column 2. They include:

1. Marked score (per item) e.g. Rows 4, 7 etc
2. $Pr(1)$ - the probability of a correct response for the student item interaction given the difficulty of the item and the estimated ability of the student
3. Residual – the calculated difference between the observed and expected result for that student/item interaction.

The Raw Score achieved (column 3) is the sum of all correct responses including correct defined guesses. The INIT ability estimate (β) (column 4) is used to calculate the probability of a correct response ($Pr(1)$). Column 4 is the score including correct guesses (9) and column 5 the RS when the correct defined guesses are accounted for (RS - CG).

The Pink shaded statistics are those interactions that have a $Pr(1)$ less than 0.25 with the whole interaction shaded when the response is a “1” – derived from the defined guesses. The green shaded statistics are the “5”s that have been randomly assigned in the “Zone of Uncertainty”.

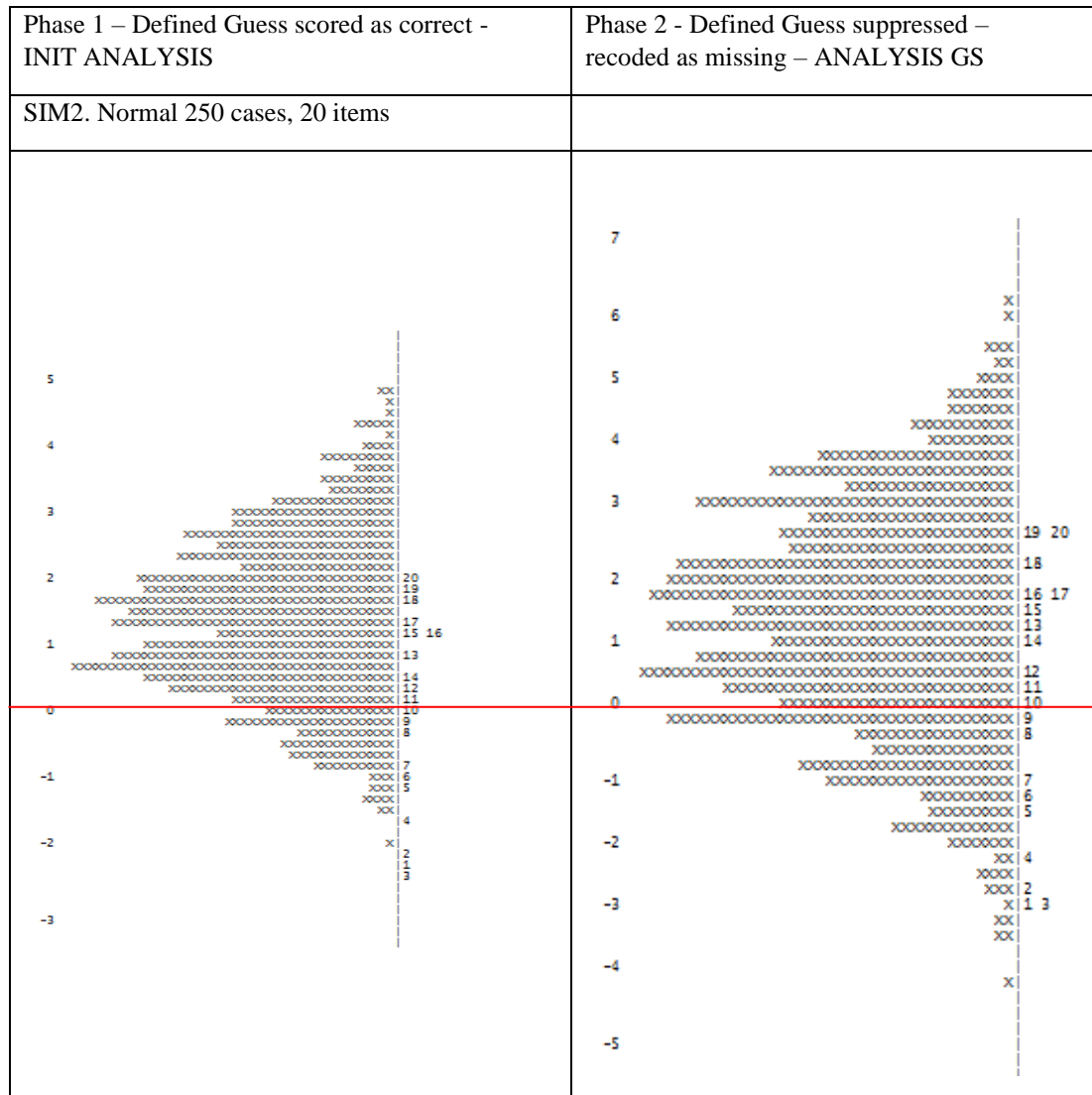
A.2 Selection of item/student maps – INIT and GS analyses $p = 0.5$

A.2.1 Item Maps for Simulations 2, 4, and 5

The item/student maps for SIM1 and SIM3 are provided in the body of the study. The summary item/student maps shown below in Figures A.1 to A.4 provide information of alternative simulated data structures.

Figure A.1

SIM2 Item/Student Map for INIT Analysis and GS Analysis

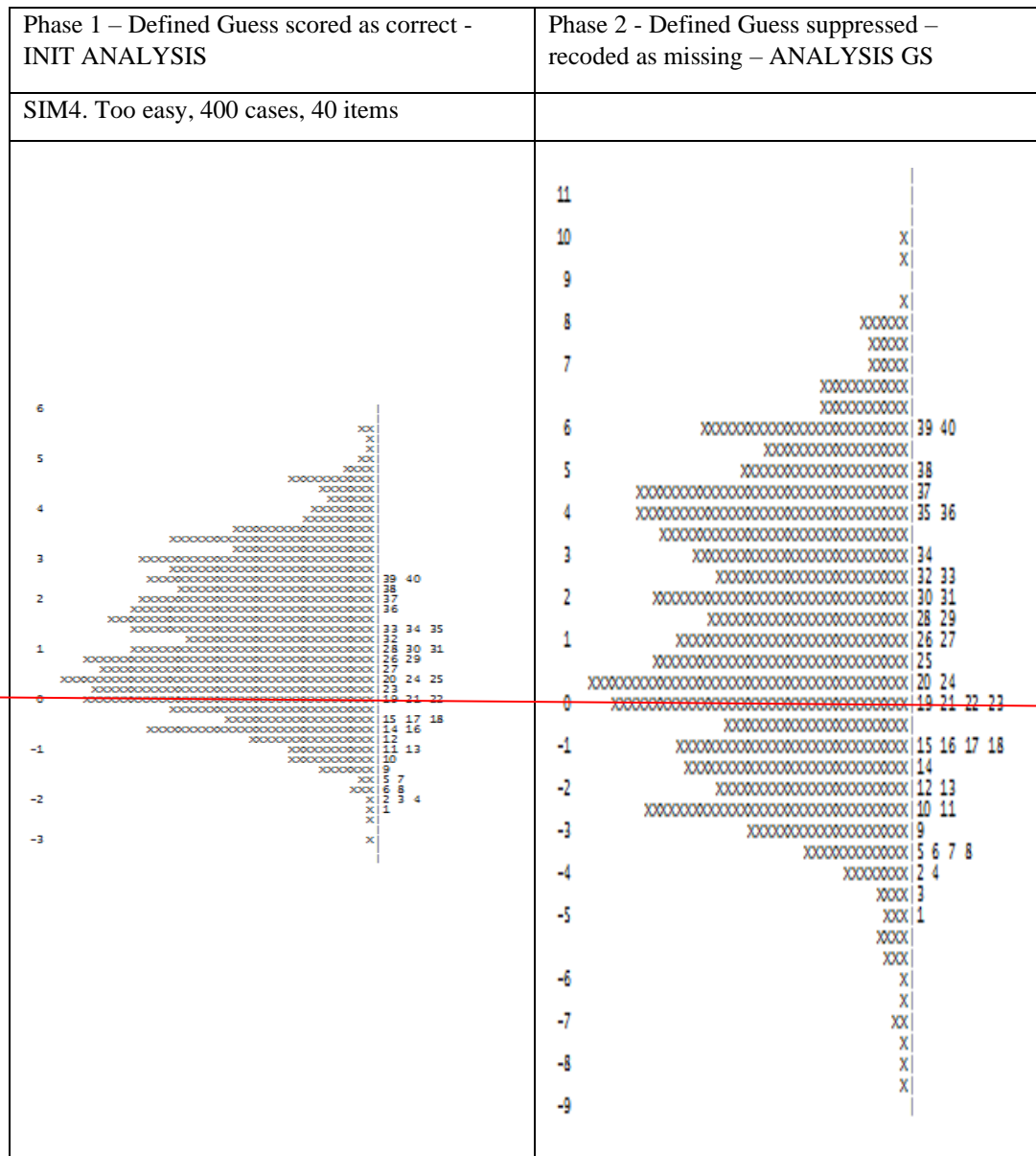


Simulation 2 comprised only 20 items and 250 cases and as such was a smaller sample. This variation was designed to investigate the effects of sample size and test length on the identification of guessing in student response patterns.

Figure A.1 shows that the test was too easy for the sample with the hardest item (20) being accessible by about 40% of the population. The removal of the defined guesses in the sample increased the distribution of the item locations and the range of ability estimates of the students with the higher-ability students achieving higher estimates and the lower-ability students being identified as having a lower ability estimate than indicated in the INIT analysis. This pattern is consistent with the outcomes of SIM1 and SIM3.

Figure A.2

SIM4 Item/Student Map for INIT Analysis and GS Analysis



The design of Simulation 4 attempted to represent a positively skewed distribution which relates to a test which is relatively easy for the target population. This is made explicit in the item/student maps with the analyses item locations centred on a mean difficulty of zero. The student ability distributions are observed to be concentrated well above the zero point on the common scale.

The similarity of the distributions with respect to the relative compression of the INIT scale and the wider distribution of the item difficulties and student ability estimates in the GS analysis was more apparent in this data set than that observed in SIM2.

The suppression of the defined guesses highlighted a proportion of the lower-ability students whose ability is overestimated in the INIGT analysis. There are also significant proportions of higher-ability students with revised GS estimates beyond the highest value observed in the INIT scale

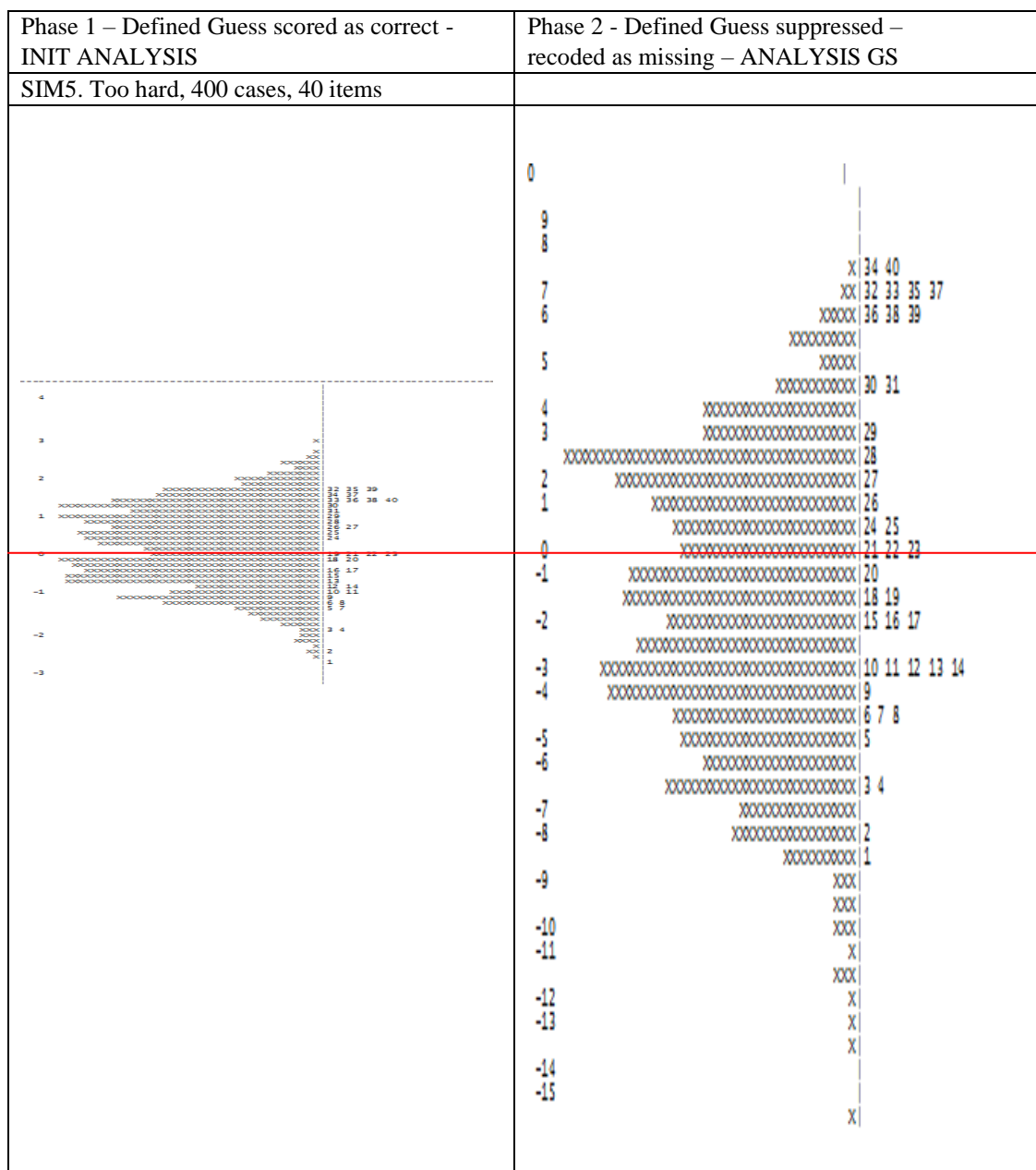
Simulation 5 data, as shown in Figure A.3, provide the obverse of Simulation 4 data with the test generally too hard for the target population resulting in a negatively skewed distribution of student ability estimates.

A feature that was highlighted when the defined guessing has been removed (GS) was the number of items that are calibrated as the most difficult in the INIT analysis were re-calibrated as too hard in the GS analysis. (viz. items 32 to 40 in the upper right-hand region of the figure).

The degree to which the scales were extended in the GS analysis compared to the INIT analysis when guessing was removed from the data was significant. Also significantly changed was the skew of the ability estimates of the population with lower-ability student having estimates recalibrated significantly lower than the INIT value.

Figure A.3

SIM5 Item-Student Map for INIT Analysis and GS Analysis



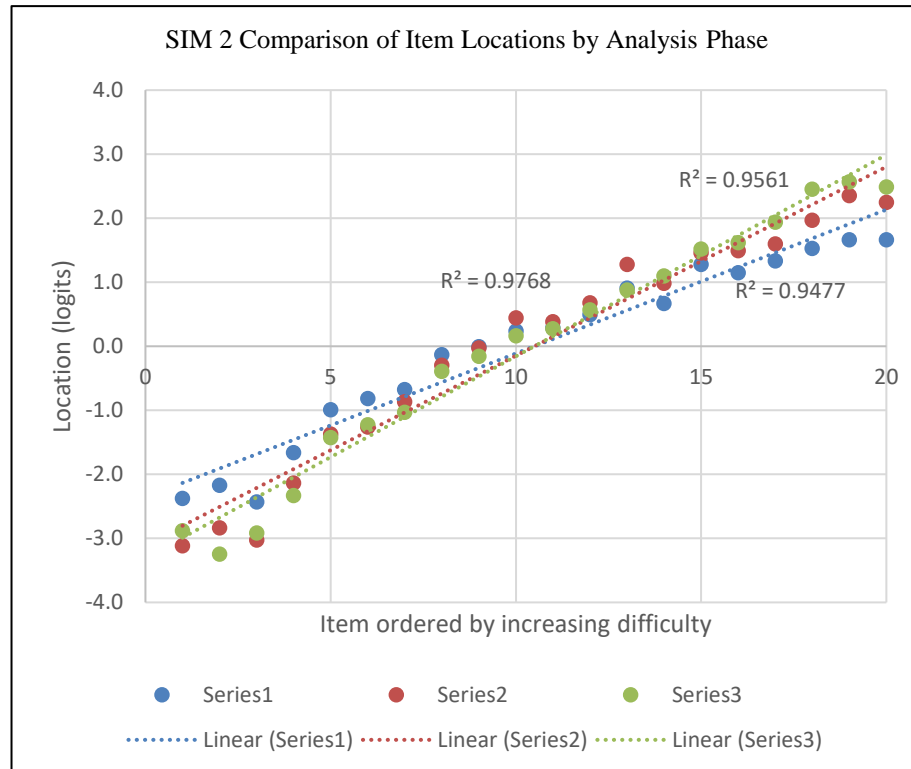
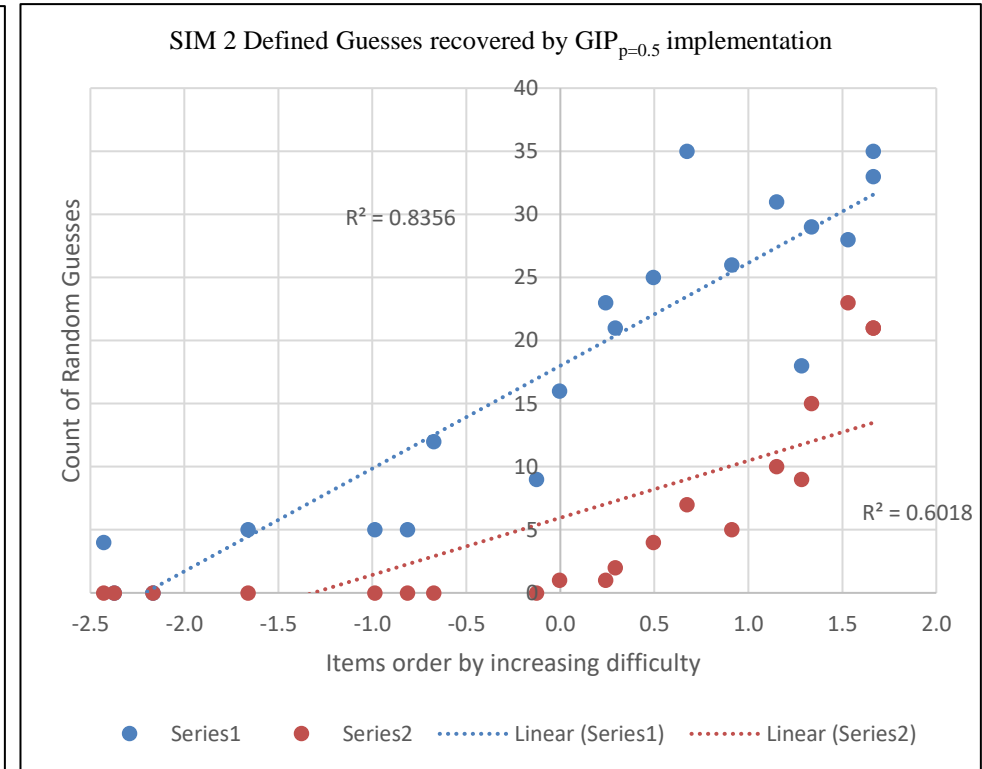
A.2.1 Item Maps for Simulations 2, 4, and 5 including $GIP_{p=0.5}$ Analyses

The Figures below detail the implementation of the $GIP_{p=0.5}$ process with the GIP3A item/student .map for each of the data sets SIM2, SIM4 and SIM5.

Figure A.4

SIM2 Comparison of Item Student Maps for INIT Analysis, GS Analysis and $GIP3A_{p=0.5}$ Analysis

Defined Guess scored as 1 - INIT Analysis	Defined Guess suppressed GS Analysis	Identified Guess suppressed $GIP3A_{p=0.5}$ Analysis
<p>SIM2. Normal 250 cases, 20 items</p>		

Figure A.5*SIM2 Comparison of Locations – INIT, GS and GIP3A_{p=0.5} Analysis***Figure A.6***SIM2 Comparison of Defined Guesses Recovered From GS by GIP_{p=0.5} Analysis*

Note. Figures A.4 to A.6 relate to the SIM2 data, a dataset of 20 items attempted by 250 students. Overall, the targeting of this test was moderately easy for the simulated students which is reflected in the observation of the relative location of the means of the student distributions shown for each analysis in Figure A.4.

The lower co-efficient of determination in respect of efficiency of the GIP application ($R^2 = 0.60$) in Figure A.6 compared to the GS analysis value may be a function of the targeting of the easy test and the fewer items in the analysis which constrains the possible distribution of item locations. The homogeneity of the items locations of Item 5 through Item 15 shown in Figure A.5 evidence the lack of discrimination in the difficulty of the items. Only in the most difficult items (16 through 20) and the more easy items (1 through 4) was there a notable difference in the item locations for the various analyses.

The SIM4 data displayed in Figure A.7 were generated to show the impact of a mis-targeted test which was too easy for the target cohort. The sample size was defined as 400 cases to control for this variable to allow for comparison with SIM1. The GS analysis shows considerable differences in the item locations and student ability estimates with guessing suppressed compared to the INIT analysis, with the GIP analysis showing a wider distribution of item locations and student estimates.

Figure A.7

SIM4 Comparison of Item/Student Maps for INIT Analysis, GS Analysis and GIP $p=0.5$ Analysis

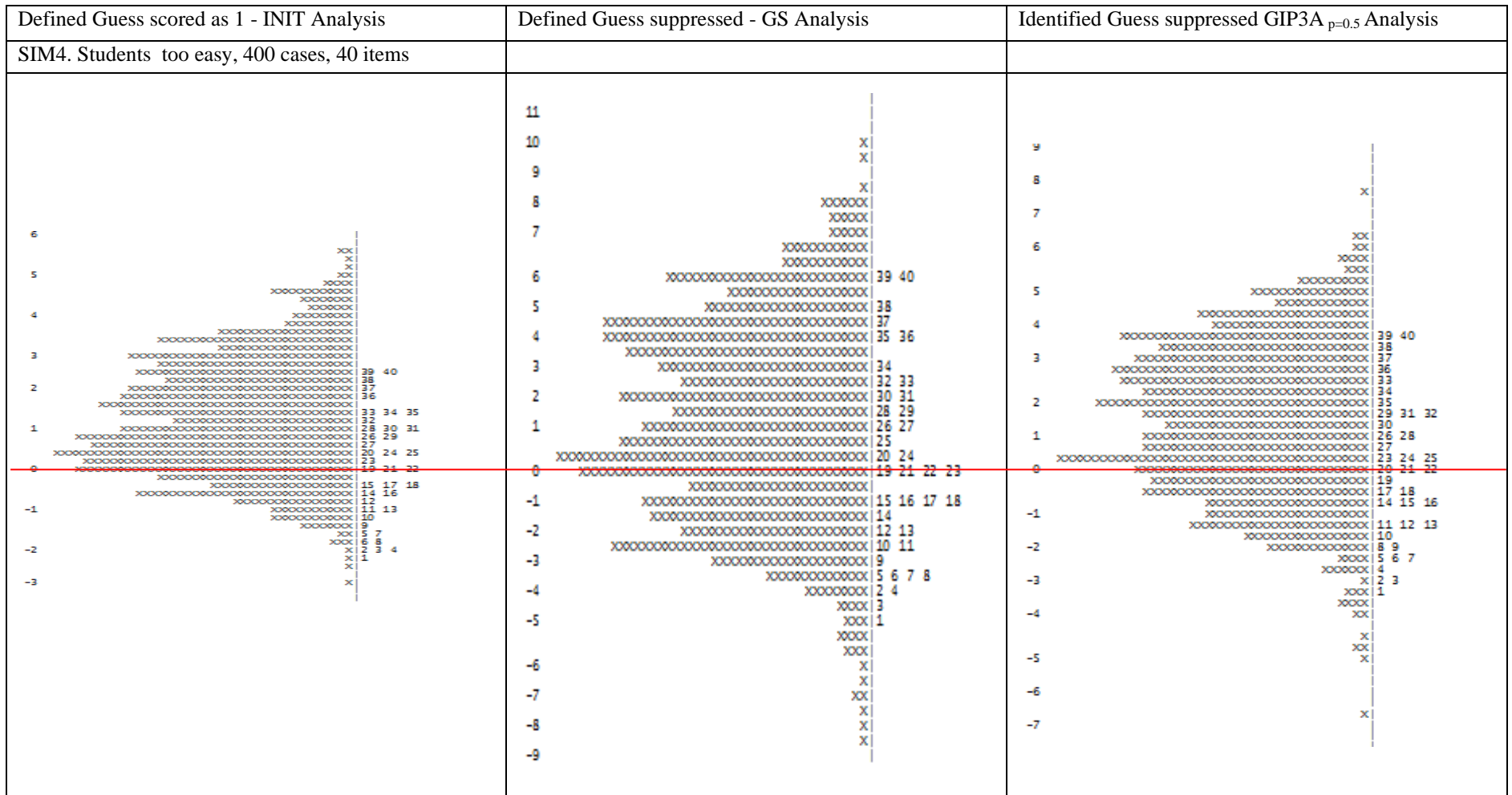


Figure A.8

SIM4 Comparison of Item Locations: INIT, GS Analysis and GIP3A_{p=0.5} Analysis

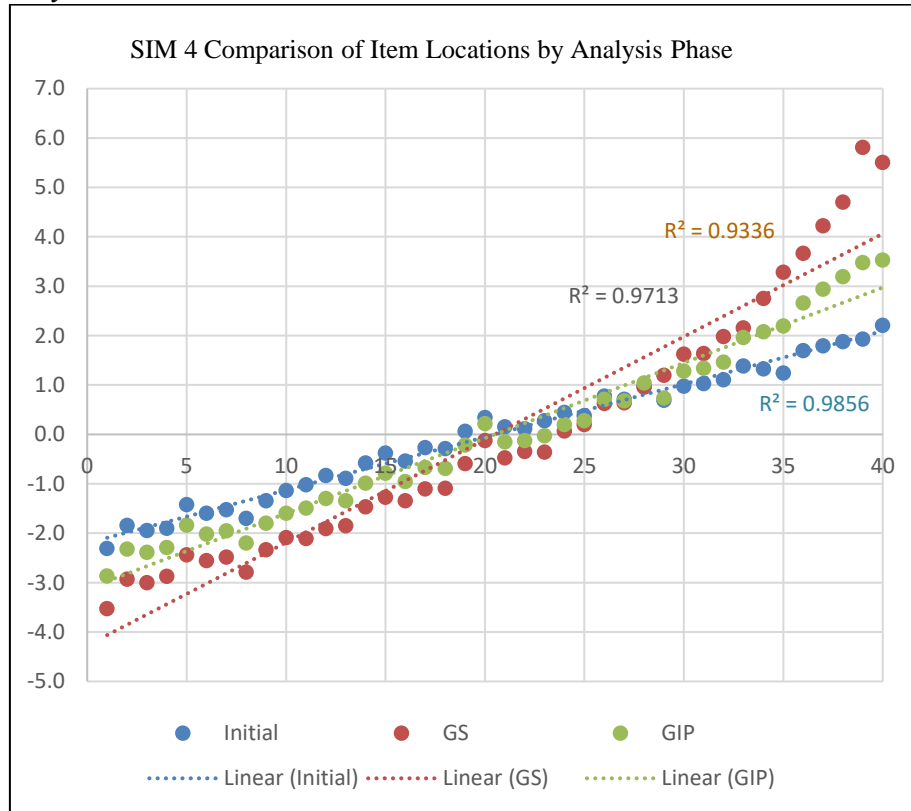


Figure A.9

SIM4 Comparison of Defined Guesses Recovered From GS by GIP_{p=0.5} Analysis

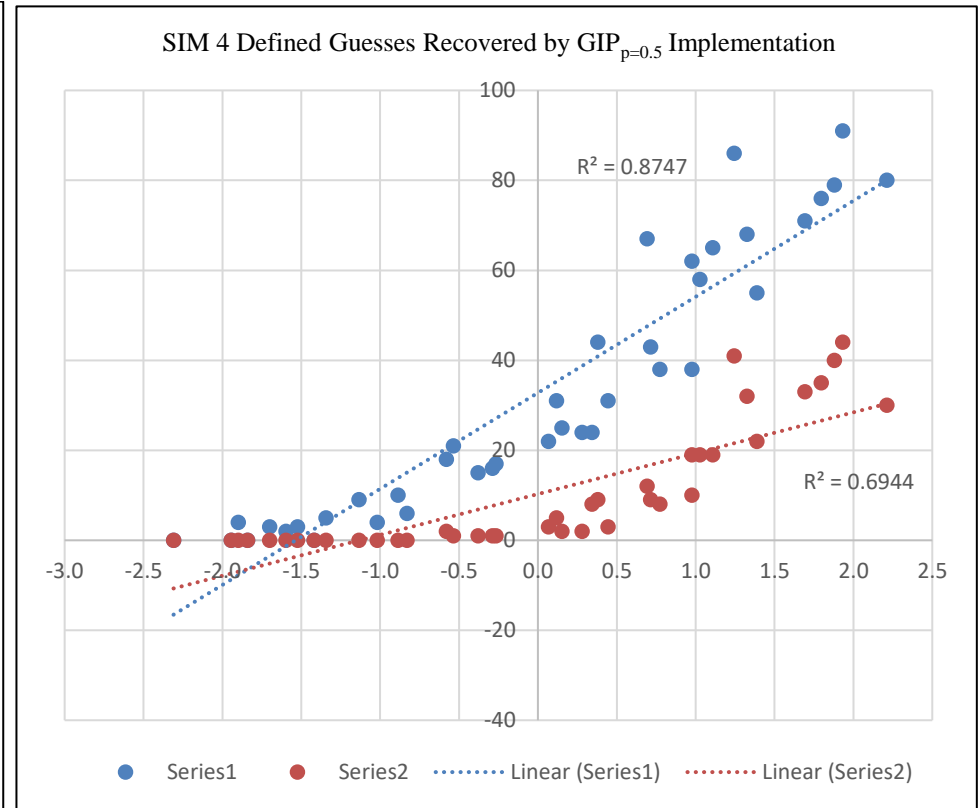
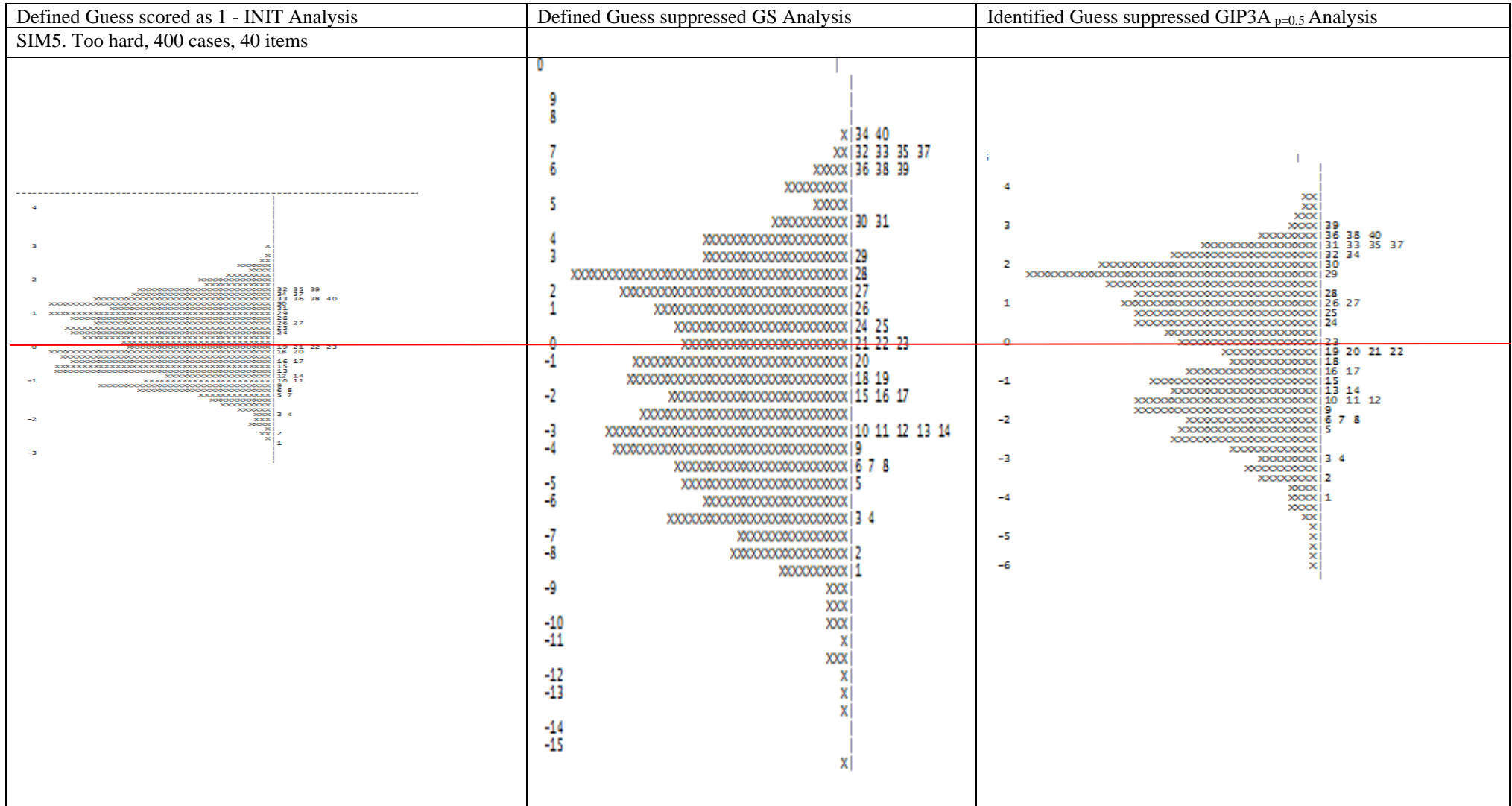


Figure A8 displays the difficulty location of each item for each analysis. As per SIM2 there is relatively little difference in the mid-range items. However at the upper end of the scale there is significant discrimination in item locations generated by the different analyses. In the easier items (items 1 through 10) there was an apparent consistency in the degree to which each analysis impacted on the item difficulty location.

Figure A.9 shows a similar pattern as the previous analyses in relation to the recovery rates of the GIP procedure compared to the full accounting for defined guesses in the GS analysis. However, it is noted that this distribution that reflects an ‘easy’ test has a lower coefficient of determination (R^2) than that observed in better targeted tests SIM1 and SIM3. This statistic represents the proportion of the variance observed in the dependent variable that is explained by the independent variable. In the figures above the dependent variables are the item order and the dependent variables are the item difficulty and count of random guesses respectively. The lower R^2 coefficient was also observed in SIM2 the previous easy test.

Figure A.10

SIM5 Comparison of Item Student Maps for INIT Analysis, GS Analysis and GIP Analysis



The SIM5 data shown in Figure A.10 were generated to show the impact of a mis-targeted test which was too hard for the target cohort. The sample size was also defined as 400 cases to control for this variable to allow for comparison with SIM1.

The INIT analysis appears to have a slightly bimodal distributed about the mean of zero for both items and students. When the defined guessing is suppressed (GS) the 'true' distribution is revealed in the graphic in the centre of Figure A.10. The bi-modal distributions were maintained however there are significant differences in the item locations. In the GS analysis the nine most difficult item are recalibrated as at the extreme of the ability range of the students, whilst the nine easiest items have item locations recalibrated below the lowest value of the INIT analysis. In the GS analysis the distribution of the student ability estimates was also far wider than the INIT analysis.

The GIP3A analysis followed a similar pattern to the GS analysis but to a lesser extent. The distribution of item locations and student ability estimates were twice that observed in the INIT analysis although the most difficult items tended to be within the range of the higher-ability students. The distributions remained bimodal but with higher discriminations between the locations.

In this data structure not only are the lower-ability students located even lower on the scale, but the more able students are also located relatively higher on the scale compared to SIM1. The R^2 statistic shown in Figure A.12 below of 0.76 suggests that the GIP procedure may be more effective for hard tests than easy tests.

An observation of interest is a common feature in each of Figures A6, A9 and A12. In each the paucity of items indicated by the GIP3A process in the region where the item location is less than zero, the default value of each analysis is noticeable. Past this point the recovery of GIP identified guesses climbs at about the rate of the actual defined guesses indicated by the GS analysis.

Figure A.11
SIM5 Comparison of Item Locations: INIT, GS Analysis and GIP3A_{p=0.5} Analysis

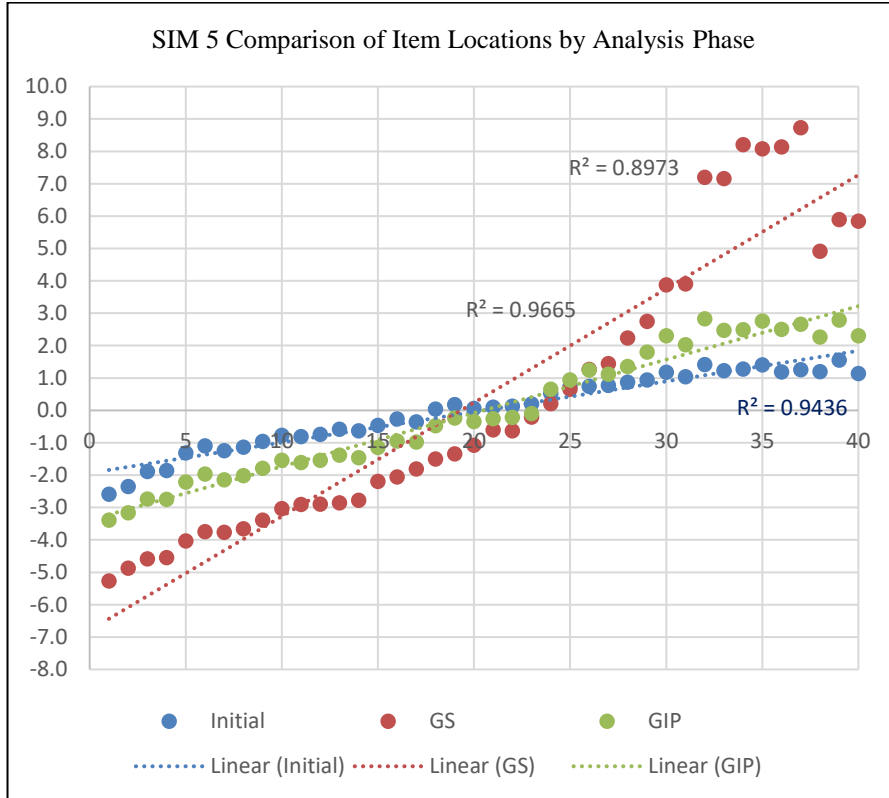
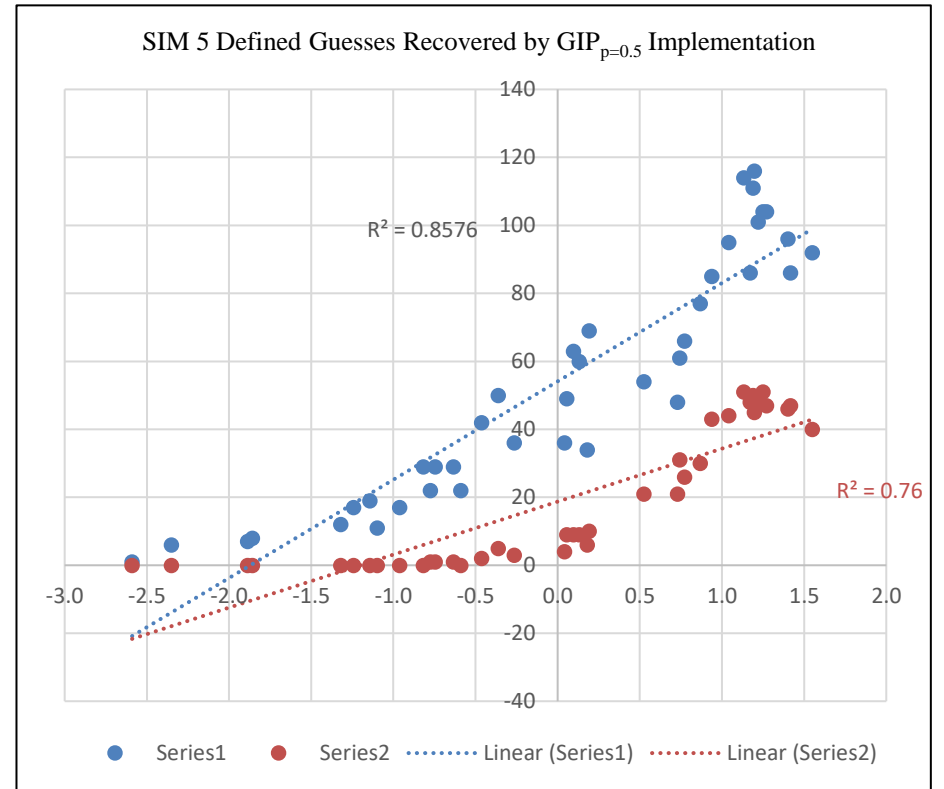


Figure A.12
SIM5 Comparison of Defined Guesses Recovered From GS by GIP_{p=0.5} Analysis



B.1 Example of GIP Calculations for SIM1 data

Table B.1 above shows examples of item/student interactions that fail the GIP parameters proposed in Section 5.4 for a selection of lower-ability students.

Row 1 shows the item locations in logits from the $INIT_{p=0.5}$ analysis

Row 2 is the item label in the original item order

Each set of student data are presented over three rows: 1. Scores (for Student (n), 2. Pr(1) (for student (n) and 3. Residual (for student (n)

Row 3 is the observed score for each item for Student S392

Row 4 is the calculated probability of a correct response for student S392 for the specific item/student interaction given the difficulty location of the item and the $p=0.6$ ability estimate of the student determined from the INIT analysis.

Row 5 is the calculated item/student residual for student S392 the observed score given the probability of a correct response.

The sequence is then repeated for student S393, S394, etc.

The highlighted cells are those for which the Residual is greater than 1.75 which also indicate the interactions in which the observed score is “1”- correct and the probability of a correct response (Pr(1)) is less than 0.25.

The highlighted cells represent those interactions identified by the $GIP_{p=0.6}$ process.

The example provided displays the manner in which the aberrant responses are indicated. When these responses are suppressed, the conditioned data become more Guttman-like.

APPENDIX C

Analysis of Small-Scale Field Study Data

C.1 Frequency Analysis – English version Self-Identified Guessing (SIG)

C.1.1 Year 5

Table C.1 below shows the relationships between the facility of each item and the self-identified logic that generated the student response with items ordered from most difficult to least difficult. The estimate of the ‘True Facility’ is calculated as the difference between the observed facility and the proportion of correct self-identified guesses.

Table C.11.1

Year 5 Item Facility Rates INIT and Self-Identified Guessing (SIG) Sorted by Item Facility

Item	Facility	Non Attempts	SIG*	Correct SIG *	True Facility	Incorrect responses	Incorrect Guess
MkM5EQ37	38.3%	1.0%	26.4%	2.3%	36.0%	60.7%	24.1%
MkM5EQ40	46.5%	1.0%	26.1%	6.9%	39.6%	52.5%	18.5%
MkM5EQ30	54.1%	1.0%	35.0%	9.6%	44.6%	44.9%	25.4%
MkM5EQ18	54.8%	4.3%	34.0%	5.3%	49.5%	40.9%	28.1%
MkM5EQ29	55.4%	0.7%	24.8%	2.6%	52.8%	43.9%	22.1%
MkM5EQ33	56.1%	0.7%	12.2%	5.6%	50.5%	43.2%	6.6%
MkM5EQ35	56.8%	1.7%	18.5%	5.6%	51.2%	41.6%	12.9%
MkM5EQ31	58.7%	1.0%	23.8%	4.0%	54.8%	40.3%	19.5%
MkM5EQ39	60.1%	2.0%	23.1%	6.9%	53.1%	38.0%	16.2%
MkM5EQ32	62.7%	2.3%	19.1%	5.3%	57.4%	35.0%	13.5%
MkM5EQ19	64.0%	1.3%	13.5%	6.3%	57.8%	34.7%	7.3%
MkM5EQ17	69.0%	3.6%	28.4%	10.6%	58.4%	27.4%	17.8%
MkM5EQ21	69.3%	0.7%	9.6%	3.3%	66.0%	30.0%	6.3%
MkM5EQ36	69.3%	1.3%	16.8%	6.9%	62.4%	29.4%	9.9%
MkM5EQ38	71.3%	0.3%	14.9%	3.0%	68.3%	28.4%	11.9%
MkM5EQ34	71.6%	1.0%	18.2%	5.3%	66.3%	27.4%	12.9%
MkM5EQ24	72.9%	0.0%	4.0%	0.3%	72.6%	27.1%	3.6%
MkM5EQ28	72.9%	1.0%	15.5%	4.3%	68.6%	26.1%	11.2%
MkM5EQ22	73.3%	1.3%	9.2%	3.0%	70.3%	25.4%	5.6%
MkM5EQ27	73.6%	0.0%	11.9%	3.6%	70.0%	26.4%	8.3%
MkM5EQ15	73.6%	0.7%	12.9%	5.6%	68.0%	25.7%	7.3%
MkM5EQ13	74.9%	1.7%	15.5%	5.9%	69.0%	23.4%	9.6%
MkM5EQ12	75.2%	1.7%	17.2%	7.9%	67.3%	23.1%	8.6%
MkM5EQ23	78.5%	0.7%	6.6%	4.0%	74.6%	20.8%	2.6%
MkM5EQ09	80.5%	0.7%	4.3%	1.3%	79.2%	18.8%	3.0%
MkM5EQ26	83.2%	2.6%	6.9%	4.0%	79.2%	14.2%	3.0%
MkM5EQ16	84.5%	0.7%	12.2%	4.3%	80.2%	14.9%	7.6%
MkM5EQ25	86.1%	2.0%	4.6%	0.7%	85.5%	11.9%	4.0%
MkM5EQ20	86.5%	1.3%	4.6%	2.0%	84.5%	12.2%	2.3%
MkM5EQ11	87.1%	0.0%	5.9%	3.6%	83.5%	12.9%	2.3%
MkM5EQ14	87.5%	2.6%	4.6%	2.0%	85.5%	9.9%	2.6%
MkM5EQ05	89.8%	2.0%	10.6%	5.9%	83.8%	8.3%	4.6%
MkM5EQ02	90.1%	1.3%	5.9%	1.7%	88.4%	8.6%	4.3%
MkM5EQ06	91.4%	0.7%	3.0%	1.7%	89.8%	7.9%	1.3%
MkM5EQ10	92.4%	0.3%	4.6%	2.0%	90.4%	7.3%	2.6%
MkM5EQ04	92.7%	0.0%	2.3%	0.3%	92.4%	7.3%	2.0%
MkM5EQ08	94.7%	0.0%	3.6%	3.0%	91.7%	5.3%	0.7%
MkM5EQ03	95.4%	0.7%	2.3%	2.0%	93.4%	4.0%	0.3%
MkM5EQ07	98.3%	0.3%	1.3%	0.7%	97.7%	1.3%	0.7%
MkM5EQ01	98.7%	0.7%	0.7%	0.7%	98.0%	0.7%	0.0%

* the column ‘SIG’ reports the proportion of self-identified guesses for each item and the column Correct

SIG reports the percentage of responses that are self-identified as a guess AND are a correct response.

Table C.1 of Year 5 shows that in Item 37 (the first item in the table) approximately $\frac{1}{10}$ (2.3 of 26.4) of the reported-attempted-guessed-items were correct, whilst in Item 40 about $\frac{1}{4}$ (6.9 of 26.1) of the self-reported guesses were correct.

Item 37 was the hardest item being correctly answered by 38.3% of the 303 students. One percent (3 students) did not attempt the item. Of the 300 students who attempted the item, 26.4% (79 students) indicated that they had guessed the response and seven (2.3%) of these students guessed correctly. Of the 60.7% of students who answered incorrectly about one-third indicated that the response was a guess. Hence approximately two-thirds applied an incorrect logic in responding to the item.

Item (37) is presented below in Figure 6.2, together with its facility annotated below each distractor.

Figure C.1

Year 5 Item 37

<p>37</p> <p>At football training, 3 boys wear red shirts and 5 boys wear blue shirts. If the coach kicks the ball to one of the boys, what is the probability that he kicks the ball to a boy wearing a red shirt?</p> <p>(A) $\frac{1}{3}$ (B) $\frac{3}{5}$ (C) $\frac{1}{8}$ (D) $\frac{3}{8}$</p>	17.2%	33.3%	10.2%	38.3%
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------	-------	-------	-------

The response pattern shows a definite logic in the incorrect distractors which have attracted a proportion of the students. Distractor B is strongly endorsed and the logic that generates this response easily unpacked. These data support the contention that not all incorrect responses are random guesses. Interrogation of the other items in Table 6.1 supports that contention. This observation is elaborated in tables and commentary that follow although there is no attempt within this study to further investigate the difference between random guesses and guesses from other strategies that may be employed by the students.

C.1.2 Year 7

Table C.2 shows the relevant statistic from Year 7 in the format presented for Year 5 students.

An interesting feature of the table is the relatively high non-attempt rates for the more challenging items. This result was surprising as there was no penalty in guessing and the fact that students were encouraged to guess and indicate the 'type' of guess using the coloured marker. Students were requested to indicate when they had completed the test during the administration. There were no reports of students having insufficient time to complete the test.

Table C.11.2*Year 7 Item Facility Rates INIT and Self-Identified Guessing (SIG) Sorted by Difficulty*

Item	Facility	Non		Correct		True	Incorrect	Incorrect
		Attempts	Guesses*	Guess*	Facility	Responses	Guess	
M7EQ42	23.3%	43.9%	5.3%	2.1%	21.2%	32.8%	3.2%	
M7EQ28	27.0%	13.2%	8.5%	5.3%	21.7%	59.8%	3.2%	
M7EQ43	27.6%	41.8%	7.9%	2.1%	25.5%	30.6%	5.8%	
M7EQ29	28.6%	17.5%	39.7%	10.1%	18.5%	53.9%	29.6%	
M7EQ41	29.7%	40.2%	19.0%	9.5%	20.2%	30.1%	9.5%	
M7EQ44	33.9%	40.7%	21.2%	11.1%	22.8%	25.3%	10.1%	
M7EQ34	39.2%	23.8%	6.9%	3.7%	35.5%	37.0%	3.2%	
M7EQ37	42.9%	30.7%	23.3%	13.8%	29.2%	26.4%	9.5%	
M7EQ31	44.0%	15.3%	24.9%	4.8%	39.2%	40.7%	20.1%	
M7EQ32	44.5%	20.6%	36.5%	15.9%	28.6%	34.9%	20.6%	
M7EQ36	45.0%	29.1%	11.1%	4.8%	40.3%	25.9%	6.3%	
M7EQ40	45.1%	36.5%	10.6%	4.8%	40.3%	18.4%	5.8%	
M7EQ38	45.6%	34.9%	24.9%	6.9%	38.7%	19.5%	18.0%	
M7EQ35	48.2%	29.1%	17.5%	7.4%	40.8%	22.7%	10.1%	
M7EQ39	50.9%	34.4%	13.2%	6.9%	44.0%	14.7%	6.3%	
M7EQ30	58.2%	14.3%	33.9%	14.3%	44.0%	27.5%	19.6%	
M7EQ26	61.4%	10.6%	38.1%	18.5%	42.9%	28.0%	19.6%	
M7EQ13	61.9%	2.1%	18.0%	9.0%	52.9%	36.0%	9.0%	
M7EQ33	64.1%	23.3%	6.9%	3.7%	60.4%	12.6%	3.2%	
M7EQ09	67.7%	6.9%	27.5%	8.5%	59.3%	25.4%	19.0%	
M7EQ25	70.4%	7.4%	19.0%	10.1%	60.3%	22.2%	9.0%	
M7EQ16	71.4%	4.2%	13.8%	5.8%	65.6%	24.3%	7.9%	
M7EQ10	72.0%	1.1%	8.5%	5.8%	66.1%	27.0%	2.6%	
M7EQ21	72.5%	3.7%	9.5%	4.8%	67.7%	23.8%	4.8%	
M7EQ24	73.6%	3.2%	16.4%	8.5%	65.1%	23.3%	7.9%	
M7EQ27	74.6%	10.1%	10.1%	4.2%	70.4%	15.3%	5.8%	
M7EQ14	75.1%	0.0%	14.8%	2.6%	72.5%	24.9%	12.2%	
M7EQ19	76.7%	2.6%	16.9%	2.6%	74.1%	20.6%	14.3%	
M7EQ17	78.8%	0.0%	10.6%	3.7%	75.1%	21.2%	6.9%	
M7EQ07	78.8%	1.1%	11.6%	2.6%	76.2%	20.1%	9.0%	
M7EQ22	78.9%	5.8%	6.9%	3.2%	75.7%	15.3%	3.7%	
M7EQ18	79.9%	2.1%	11.6%	7.4%	72.5%	18.0%	4.2%	
M7EQ05	83.1%	0.0%	6.3%	4.8%	78.3%	16.9%	1.6%	
M7EQ15	83.6%	1.1%	5.8%	4.8%	78.8%	15.3%	1.1%	
M7EQ23	87.8%	2.1%	10.6%	7.4%	80.4%	10.0%	3.2%	
M7EQ06	88.9%	1.1%	7.4%	5.8%	83.1%	10.0%	1.6%	
M7EQ20	91.0%	2.1%	7.9%	6.9%	84.1%	6.9%	1.1%	
M7EQ03	91.5%	2.1%	2.6%	2.6%	88.9%	6.3%	0.0%	
M7EQ11	92.6%	1.1%	9.0%	5.3%	87.3%	6.3%	3.7%	
M7EQ12	93.1%	1.1%	2.1%	1.6%	91.5%	5.8%	0.5%	
M7EQ01	94.7%	3.2%	1.1%	1.1%	93.7%	2.1%	0.0%	
M7EQ02	95.2%	3.7%	2.1%	2.1%	93.1%	1.1%	0.0%	
M7EQ04	97.9%	0.0%	0.0%	0.0%	97.9%	2.1%	0.0%	
M7EQ08	97.9%	0.0%	4.2%	3.2%	94.7%	2.1%	1.1%	

C.2Extracts of GIP indicated Guessed Items for Selected Students

C.2.1 Indicating Probable Guessing in the Year 5 data

A selection of student responses is provided in Tables C.3, C.4, C.5 and C.6 that follow to demonstrate the relationships between student responses, the proposed GIP parameters and the self-indicated guessing for differing ability levels.

Table C.3

Response Pattern for High-range Ability Year 5 Students Ordered by Item Difficulty Showing the Self-identified Response Mode (G, Y or P)

ID	Person	β	-4.23	-3.12	-1.97	-1.5	-1.46	-1.32	-1.13	-1.11	-1.07	-0.95	-0.66	-0.66	-0.65	-0.47	-0.45	-0.05	0.078	0.2	0.257	0.36	0.4	0.402	0.403	0.447	0.5	0.53	0.533	0.62	0.655	0.934	1.001	1.077	1.257	1.281	1.315	1.359	1.418	1.47	1.942	2.367
			EQ01	EQ07	EQ03	EQ10	EQ08	EQ04	EQ05	EQ02	EQ06	EQ14	EQ11	EQ25	EQ20	EQ26	EQ16	EQ09	EQ23	EQ12	EQ13	EQ22	EQ15	EQ27	EQ28	EQ24	EQ34	EQ38	EQ17	EQ36	EQ21	EQ19	EQ32	EQ39	EQ31	EQ35	EQ33	EQ18	EQ29	EQ30	EQ40	EQ37
G5E205	203	1.431	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	1	1	0	1	1	0	0
G5E205	203	1.431	0.997	0.990	0.968	0.949	0.948	0.940	0.928	0.927	0.924	0.916	0.890	0.890	0.889	0.870	0.868	0.815	0.795	0.774	0.764	0.745	0.737	0.737	0.737	0.728	0.717	0.711	0.711	0.692	0.685	0.622	0.606	0.588	0.543	0.537	0.529	0.518	0.503	0.490	0.375	0.282
G5E205	203	1.431	G	G	G	G	G	G	G	Y	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	Y	G	G	G	Y	G	G	G	G	G
5E209	203	1.431	0.059	0.103	0.183	0.231	0.235	-3.97	0.278	-3.56	0.287	0.304	0.351	0.352	0.354	0.386	0.39	0.477	-1.97	-1.85	0.556	0.585	0.597	0.598	0.598	0.611	0.628	0.637	0.638	-1.5	-1.47	0.78	0.806	-1.19	0.917	0.928	0.944	1.02	-0.77	-0.63		
G5E21C	204	1.431	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1	1	0	0	
G5E21C	204	1.431	0.997	0.990	0.968	0.949	0.948	0.940	0.928	0.927	0.924	0.916	0.890	0.890	0.889	0.870	0.868	0.815	0.795	0.774	0.764	0.745	0.737	0.737	0.737	0.728	0.717	0.711	0.711	0.692	0.685	0.622	0.606	0.588	0.543	0.537	0.529	0.518	0.503	0.490	0.375	0.282
G5E21C	204	1.431	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	Y	G	G	G	G	G	G	G	G	G	Y	G	G	G	G	G	G	G	Y	G	Y	G	G	
5E210	204	1.431	0.059	0.103	0.183	0.231	0.235	0.252	0.278	0.281	0.287	0.304	0.351	0.352	0.354	0.386	0.39	-2.1	-1.97	0.54	0.556	0.585	0.597	-1.67	0.598	-1.64	0.628	0.637	0.638	-1.5	-1.47	0.78	0.806	0.838	0.917	0.928	-1.06	-1.04	0.994	1.02	-0.77	-0.63
G5E19E	192	1.557	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	9	1	1	0	1	1	0	1	1	1	1	0	1	0	1	0	1	0	1	1	0	0	
G5E19E	192	1.557	0.997	0.991	0.971	0.955	0.953	0.947	0.936	0.935	0.932	0.925	0.902	0.902	0.901	0.884	0.882	0.833	0.814	0.795	0.786	0.768	0.761	0.760	0.760	0.752	0.742	0.736	0.736	0.718	0.711	0.651	0.636	0.618	0.574	0.569	0.560	0.549	0.535	0.522	0.405	0.308
G5E19E	192	1.557	G	G	G	G	G	G	G	G	G	G	G	G	Y	G	G	Y	9	G	G	G	Y	Y	G	G	G	G	Y	Y	G	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	P
5E198	192	1.557	0.055	0.096	0.171	0.217	0.221	0.237	0.261	0.264	0.269	0.285	0.33	0.33	0.333	0.363	0.366	0.448	0.477	*	0.522	0.55	-1.78	0.561	0.562	-1.74	0.59	0.599	0.6	0.626	-1.57	0.733	-1.32	0.787	-1.16	0.871	-1.13	-1.1	0.933	0.957	-0.83	-0.67
G5E027	27	1.574	1	1	1	1	1	1	1	1	0	1	1	9	0	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	0	1	1	0	0	1	0
G5E027	27	1.574	0.997	0.991	0.972	0.956	0.954	0.948	0.937	0.936	0.933	0.926	0.904	0.903	0.902	0.886	0.884	0.835	0.817	0.798	0.789	0.771	0.764	0.764	0.763	0.755	0.745	0.740	0.739	0.722	0.715	0.655	0.639	0.622	0.579	0.573	0.564	0.554	0.539	0.526	0.409	0.312
G5E027	27	1.574	G	G	G	G	G	G	G	G	G	G	G	9	G	G	G	G	G	9	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
5E027	27	1.574	0.055	0.095	0.17	0.215	0.219	0.235	0.259	0.262	-3.75	0.283	0.327	*	-3.03	0.359	0.363	-2.25	0.473	0.503	0.518	0.545	0.556	0.556	-1.8	0.569	0.584	0.593	0.594	0.621	0.632	-1.38	0.751	0.78	0.853	-1.16	0.878	0.898	-1.08	-1.05	1.202	-0.67
G5E007	7	1.589	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0	0
G5E007	7	1.589	0.997	0.991	0.972	0.956	0.955	0.948	0.938	0.937	0.934	0.927	0.905	0.905	0.903	0.887	0.885	0.837	0.819	0.800	0.791	0.774	0.767	0.766	0.766	0.758	0.748	0.742	0.742	0.725	0.718	0.658	0.643	0.625	0.582	0.576	0.568	0.557	0.543	0.530	0.413	0.315
G5E007	7	1.589	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
5E007	7	1.589	0.055	0.095	0.169	0.213	-4.6	0.233	0.257	0.26	0.265	0.28	-3.09	0.325	0.327	0.357	0.36	0.441	0.47	0.499	-1.95	-1.85	0.552	0.552	0.552	0.565	-1.72	0.589	0.59	0.616	0.627	0.721	0.745	0.774	0.847	-1.17	-1.15	0.891	0.918	0.942	-0.84	-0.68

Table C3 demonstrates the challenge of the GIP process to indicate probable guesses among the higher-ability students. The item/student interactions highlighted in Yellow indicate the items to which the student indicated they did not know the answer but had some idea – an educated guess. Student G5E204 has a combination of correct and incorrect Yellow (Y) coded responses. However in each case the Pr(1) row – the probability of a correct response for a student of that ability estimate is greater than 0.25, and the residual values are all less than 1.75 and hence these responses do not ‘fail’ the GIP parameters.

Student G5E007 has two self indicated guesses highlighted in Pink (P) which were correct answers. However neither of these are indicated by the GIP procedure due to the ability of the student interaction with items of the calculated difficulty. In both cases the calculated values of the probability of a correct response (Pr(1)) and the item/student residual do not fail the proposed GIP parameters.

C.2.2 Indicating Probable Guessing in the Year 7 data

A similar set of analyses were conducted with the Year 7 data. Table C.4 below provides an extract of the response patterns and GIP outcomes for the lower-ability students.

Table C.4

Response Pattern for Low-ability Year 7 Students Ordered by Item Difficulty Including Response Mode (G, Y, or P) (p = 0.5)

Table with 48 columns (Criteria, Seq, ID, beta, and 45 item IDs) and 20 rows (Criteria, 1. Score, 2. Pr(1), 3. Residual, 4. Mode) for each of the 12 students (188, 189, 28, 12, 21, 63). The table displays scores, probabilities, residuals, and response modes (G, Y, P) for each item, with some cells highlighted in pink or red.

Table C.4 is presented in the same format as the Year 5 data with the first row showing the item difficulties ordered by difficulty from easiest to hardest. The second row records the item label. The third row (1. Score) records the scored responses for student G7E188. The next row 2.(Pr(1)) records the probability that a student of ability estimate -1.54 (the INIT estimate of student G7E188) would achieve a correct response to each item in the test. In all but the first 12 items the probability of a correct response to items of increasing difficulty in this test is less than 0.25; highlighted in red text. The next row (3. Residual) records the student/item residual for the specific response recorded for each item. For Student 188 the values greater than 1.75 have been highlighted in pink. The next row (4. Mode) shows the student's self-indication for each response to each item and it is noted that student 188 knows (G) the answer to every item. The pattern of rows 3 through 6 is then repeated for each student.

Table C.5 is an extract of the six least-able students (by result) in the Year 7 test. The cells shaded in pink show items that were indicated by the GIP procedure (Score = 1 AND Pr(1) < 0.25 AND Residual > 1.75) when the 'p' value for success in an item is set at p = 0.5. The table shows that few cases were indicated in the lower-ability group and the mid-range group and no cases indicated in the most-able group of students.

Table C.5

Response Pattern for Low-ability Year 7 Students Ordered by Item Difficulty Including Response Mode (G, Y, or P) ($p = 0.6$)

Criteria	Seq	ID	δ	-4.15	-2.94	-2.764	-2.407	-1.6	-1.504	-1.345	-1.341	-1.18	-1.075	-0.489	-0.447	-0.395	-0.373	-0.263	-0.226	-0.096	0.014	0.055	0.076	0.096	0.106	0.115	0.279	0.315	0.316	0.491	0.635	0.67	0.69	0.704	0.834	0.953	1.072	1.259	1.287	1.455	1.469	1.507	1.706	1.968	2.154	2.394			
1. Score	188	G7E188	-1.54	-1.941	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0		
2. Pr(1)	188	G7E188	-1.54	-1.941	0.901	0.731	0.695	0.614	0.416	0.392	0.355	0.354	0.318	0.296	0.190	0.183	0.176	0.157	0.153	0.136	0.129	0.124	0.120	0.117	0.115	0.114	0.113	0.098	0.095	0.095	0.081	0.071	0.068	0.067	0.066	0.059	0.052	0.047	0.039	0.038	0.032	0.032	0.031	0.025	0.020	0.016	0.013		
3. Residual	188	G7E188	-1.54	-1.941	-3.018	-1.648	-1.509	0.792	-0.843	-0.804	-0.742	-0.741	-0.684	1.542	-0.484	-0.474	-0.462	-0.457	-0.452	2.357	2.516	2.603	-0.376	-0.369	-0.365	-0.361	-0.359	2.795	-0.330	-0.324	-0.324	3.374	-0.276	-0.271	3.727	-0.266	-0.250	-0.235	-0.222	-0.202	-0.199	-0.183	5.501	-0.178	-0.161	7.060	-0.129	-0.114	
4. mode	188	G7E188	-1.54	-1.941	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G		
1. Score	28	G7E028	-1.37	-1.771	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
2. Pr(1)	28	G7E028	-1.37	-1.771	0.915	0.763	0.730	0.654	0.457	0.434	0.395	0.394	0.356	0.333	0.217	0.210	0.202	0.198	0.181	0.176	0.158	0.149	0.144	0.139	0.136	0.134	0.133	0.132	0.114	0.110	0.110	0.094	0.083	0.080	0.079	0.078	0.069	0.062	0.055	0.046	0.045	0.038	0.038	0.036	0.030	0.023	0.019	0.015	
3. Residual	28	G7E028	-1.37	-1.771	0.304	0.557	0.609	0.728	-0.918	-0.875	-0.808	1.240	1.344	1.416	-0.527	1.939	-5.503	-0.497	-0.470	-0.462	-0.433	2.390	-0.410	2.492	-0.397	-0.393	-0.391	-0.389	-0.359	-0.352	-0.352	-0.323	-0.300	-0.295	-0.292	-0.290	3.678	-0.256	-0.241	-0.220	-0.217	-0.199	-0.198	-0.194	-0.176	-0.154	-0.141	-0.125	
4. mode	28	G7E028	-1.37	-1.771	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	
1. Score	30	G7E030	-1.37	-1.771	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
2. Pr(1)	30	G7E030	-1.37	-1.771	0.915	0.763	0.730	0.654	0.457	0.434	0.395	0.394	0.356	0.333	0.217	0.210	0.202	0.198	0.181	0.176	0.158	0.149	0.144	0.139	0.136	0.134	0.133	0.132	0.114	0.110	0.110	0.094	0.083	0.080	0.079	0.078	0.069	0.062	0.055	0.046	0.045	0.038	0.038	0.036	0.030	0.023	0.019	0.015	
3. Residual	30	G7E030	-1.37	-1.771	0.304	0.557	0.609	0.728	-0.918	-0.875	-0.808	1.240	1.344	1.416	-0.527	1.939	-5.503	-0.497	-0.470	-0.462	-0.433	2.311	2.390	-0.410	-0.401	-0.397	-0.393	-0.391	-0.389	-0.359	-0.352	-0.352	-0.323	-0.300	-0.295	-0.292	-0.290	-0.272	-0.256	-0.241	-0.220	-0.217	-0.199	-0.198	-0.194	-0.176	-0.154	-0.141	-0.125
4. mode	30	G7E030	-1.37	-1.771	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	
1. Score	133	G7E133	-1.37	-1.771	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Pr(1)	133	G7E133	-1.37	-1.771	0.915	0.763	0.730	0.654	0.457	0.434	0.395	0.394	0.356	0.333	0.217	0.210	0.202	0.198	0.181	0.176	0.158	0.149	0.144	0.139	0.136	0.134	0.133	0.132	0.114	0.110	0.110	0.094	0.083	0.080	0.079	0.078	0.069	0.062	0.055	0.046	0.045	0.038	0.038	0.036	0.030	0.023	0.019	0.015	
3. Residual	133	G7E133	-1.37	-1.771	0.304	0.557	0.609	0.728	-0.918	-0.875	-0.808	1.240	1.344	1.416	-0.527	1.939	-5.503	-0.497	-0.470	-0.462	-0.433	2.390	-0.410	2.492	-0.397	-0.393	-0.391	-0.389	-0.359	-0.352	-0.352	-0.323	-0.300	-0.295	-0.292	-0.290	-0.272	-0.256	-0.241	-0.220	-0.217	-0.199	-0.198	-0.194	-0.176	-0.154	-0.141	-0.125	
4. mode	133	G7E133	-1.37	-1.771	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
1. Score	115	G7E115	-1.22	-1.621	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
2. Pr(1)	115	G7E115	-1.22	-1.621	0.926	0.789	0.758	0.687	0.495	0.471	0.431	0.430	0.392	0.367	0.244	0.236	0.227	0.223	0.205	0.199	0.179	0.169	0.163	0.158	0.155	0.152	0.151	0.150	0.130	0.126	0.126	0.108	0.095	0.092	0.090	0.089	0.079	0.071	0.063	0.053	0.044	0.044	0.042	0.035	0.027	0.022	0.018		
3. Residual	115	G7E115	-1.22	-1.621	0.282	0.517	0.565	0.675	1.011	1.060	1.148	1.150	-0.802	-0.761	1.761	-0.556	-0.542	-0.536	-0.507	-0.498	2.144	2.218	-0.442	-0.433	-0.428	-0.424	-0.422	2.392	-0.387	-0.380	-0.380	-0.348	-0.324	-0.318	-0.315	-0.313	-0.293	-0.276	-0.260	-0.237	-0.234	-0.215	-0.213	-0.209	-0.189	-0.166	-0.151	-0.134	
4. mode	115	G7E115	-1.22	-1.621	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
1. Score	175	G7E175	-1.06	-1.461	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
2. Pr(1)	175	G7E175	-1.06	-1.461	0.936	0.814	0.786	0.720	0.535	0.511	0.471	0.470	0.430	0.405	0.274	0.266	0.256	0.252	0.232	0.225	0.203	0.193	0.186	0.180	0.177	0.174	0.173	0.171	0.149	0.145	0.124	0.109	0.106	0.104	0.103	0.092	0.082	0.074	0.062	0.060	0.051	0.051	0.049	0.040	0.031	0.026	0.021		
3. Residual	175	G7E175	-1.06	-1.461	0.261	0.477	0.521	0.623	0.933	0.979	1.060	1.062	1.151	1.213	-0.615	-0.602	-0.587	-0.580	1.800	-0.539	-0.505	-0.488	2.091	-0.469	2.157	-0.459	-0.457	-0.455	-0.419	-0.411	-0.411	-0.377	-0.351	-0.345	-0.341	-0.339	-0.317	-0.299	-0.282	-0.257	-0.253	-0.233	-0.231	-0.227	-0.205	-0.180	-0.164	-0.146	
4. mode	175	G7E175	-1.06	-1.461	G	Y	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G

Table C.6

Response Pattern for Mid-range Ability Year 7 Students Ordered by Item Difficulty Including Response Mode (G, Y, or P)

Criteria	Seq	ID	δ	-4.15	-2.94	-2.764	-2.41	-1.6	-1.5	-1.35	-1.34	-1.18	-1.08	-0.49	-0.45	-0.4	-0.37	-0.26	-0.23	-0.1	-0.03	0.014	0.055	0.076	0.096	0.106	0.115	0.279	0.315	0.316	0.491	0.635	0.67	0.69	0.704	0.834	0.953	1.072	1.259	1.287	1.455	1.469	1.507	1.706	1.968	2.154	2.394	
1. Score	140	G7E140	-0.03	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Pr(1)	140	G7E140	-0.03	0.984	0.948	0.939	0.915	0.82																																								

In reviewing the items indicated by the GIP process when a 'p' value of 0.5 was applied to the student ability estimate, a total of 101 cases were identified as probable guesses. Of these, 36 cases were identified by the student as 'knowing' the response (Green). Deconstructing this statistics by ability group, 25 of the 46 (54%) cases indicated in the lowest quartile were **not** self-identified by coloured markers as guesses. Of the 31 (35%) cases indicated by the GIP process, 11 were **not** self identified as guesses in the next lowest ability quartile. In the next most-able quartile all of the 23 (0%) GIP indicated probable guesses were also self-identified guesses. In the most-able ability group no cases were indicated as probable guesses by the GIP procedure.

Table C.6 shows the impact on the lower-ability students with an additional 14 item/student interactions being indicated by the GIP process as probable guesses when a 'p' value of 0.6 was applied to the ability estimate of the initial analysis. It was noted that in only two cases were items indicated by the GIP procedure which the student had not indicated as a random guess (7 instances) or a partial knowledge guess (5 cases).

The implementation of the $p = 0.6$ constraint on the student ability estimate increased the number of items indicated by the protocol by 141 (223 vs 82) cases. In a pattern similar to that shown from the $p = 0.5$ analysis the majority of indicated guessing cases were determined in the lowest ability quartile with 131 cases, of which 81 (62%) did **not** self identify as a guess, the GIP indicated item/student result as a guess. In the next least-able quartile 16 of the 62 (25%) GIP indicated probable guesses were not self-identified. In the next most-able ability quartile the GIP procedure indicated 29 cases of whom ALL were self-identified either a partial guess OR a random guess. In the most-able ability group no cases were indicated as probable guesses.

C.3 Arabic Versions of the Tests

C.3.1 Analyses of Facility Rates of Year 5 Items

Table C.7 highlights the relationship between the facility rates of the items in which there was a high correlation between the facilities of the Arabic and English versions of the test items for the various analyses. Table C.7 also annotates a plausible reason why these relationships may or may not be apparent.

Table C.7 Year 5 Comparison of Item Facility by Analysis Phase

Year 5 Mathematics		English			Arabic			Language demand	
Seq	Item Label	Original*	INIT	SIG	GIP _{p0.6}	INIT	SIG	GIP _{p0.6}	
1	Math5Q01	92%	99%	99%	99%	77%	26%	78%	
2	Math5Q02	89%	91%	91%	91%	32%	23%	32%	
3	Math5Q03	85%	96%	96%	96%	96%	97%	96%	Algorithm – Non-language dependent
4	Math5Q04	84%	93%	93%	92%	27%	1%	27%	
5	Math5Q05	76%	92%	91%	91%	11%	1%	8%	
6	Math5Q06	75%	92%	92%	92%	75%	71%	74%	Image interpretation
7	Math5Q07	71%	99%	99%	99%	58%	26%	60%	
8	Math5Q08	68%	95%	95%	95%	30%	2%	31%	
9	Math5Q09	64%	81%	81%	80%	48%	30%	45%	
10	Math5Q10	60%	93%	93%	92%	33%	16%	32%	
11	Math5Q11	60%	87%	87%	87%	41%	10%	39%	
12	Math5Q12	59%	77%	74%	75%	37%	2%	37%	
13	Math5Q13	59%	76%	75%	75%	20%	10%	18%	
14	Math5Q14	58%	90%	90%	89%	37%	1%	32%	
15	Math5Q15	58%	73%	72%	73%	57%	46%	55%	Image interpretation
16	Math5Q16	57%	85%	84%	84%	55%	17%	53%	Image interpretation
17	Math5Q17	54%	72%	68%	70%	57%	39%	57%	
18	Math5Q18	54%	57%	55%	53%	24%	2%	23%	
19	Math5Q19	52%	65%	63%	62%	15%	2%	13%	
20	Math5Q20	51%	88%	87%	87%	12%	1%	10%	
21	Math5Q21	50%	70%	69%	68%	9%	*	8%	
22	Math5Q22	49%	74%	73%	73%	32%	1%	34%	
23	Math5Q23	46%	79%	78%	78%	28%	17%	26%	
24	Math5Q24	45%	73%	73%	72%	65%	63%	63%	Image interpretation
25	Math5Q25	45%	88%	88%	87%	15%	2%	13%	
26	Math5Q26	44%	85%	85%	85%	43%	22%	38%	
27	Math5Q27	43%	74%	73%	72%	33%	1%	32%	
28	Math5Q28	43%	74%	72%	72%	28%	3%	25%	
29	Math5Q29	42%	56%	55%	53%	53%	39%	50%	
30	Math5Q30	41%	55%	50%	51%	32%	13%	32%	Image interpretation
31	Math5Q31	38%	59%	58%	57%	40%	26%	38%	
32	Math5Q32	37%	64%	62%	62%	58%	55%	57%	Algorithm – Non-language dependent
33	Math5Q33	36%	56%	54%	51%	33%	3%	33%	
34	Math5Q34	36%	72%	71%	71%	8%	*	7%	
35	Math5Q35	35%	58%	55%	53%	66%	52%	64%	Equation interpretation
36	Math5Q36	34%	70%	68%	68%	7%	*	7%	
37	Math5Q37	32%	39%	37%	32%	13%	1%	13%	
38	Math5Q38	31%	72%	71%	70%	35%	10%	33%	
39	Math5Q39	29%	61%	58%	57%	31%	2%	30%	
40	Math5Q40	26%	47%	43%	41%	42%	26%	40%	Image interpretation

C.3.2 Relative Performance of Selected Items – Arabic and English Facility Rates

Table C.8 shows a breakdown of the results of students have responded to the Arabic version of the test together with the facility rates achieved in both the Arabic and the English implementations of the test.

Table C.8

Year 5 Comparison of Item Facility INIT Analysis of English and Arabic Versions of Test

Maths5	Baseline stats		INIT analysis outcomes			Response mode correct Arabic*		
Grade 5 Item no	Facility Rate	Disc	Language Dependency	English Version	Arabic Version	Known Answer	Gessed Informed	Gessed Random
1	88.7%	0.32	Arabic	99.0%	78.2%	9.3%	54.9%	35.9%
2	85.5%	0.33	Arabic Clued	91.0%	32.0%	60.8%	25.8%	13.4%
3	84.4%	0.33	English Equivalent	96.0%	95.4%	95.5%	3.5%	1.0%
4	92.2%	0.27	Arabic	92.0%	26.4%	3.8%	13.8%	82.5%
5	75.4%	0.42	Arabic	91.0%	9.6%	6.9%	6.9%	86.2%
6	76.5%	0.29	Arabic Clued	92.0%	75.6%	76.0%	19.2%	4.8%
7	64.0%	0.44	Arabic Clued	99.0%	59.1%	22.9%	24.0%	53.1%
8	67.9%	0.47	Arabic	95.0%	30.7%	5.4%	34.4%	60.2%
9	63.5%	0.41	Arabic	80.0%	47.2%	46.2%	34.3%	19.6%
10	60.3%	0.53	Arabic Clued	92.0%	32.3%	37.8%	40.8%	21.4%
11	60.0%	0.51	Arabic	87.0%	40.6%	15.4%	61.8%	22.8%
12	59.3%	0.43	Arabic	76.0%	38.0%	2.6%	18.3%	79.1%
13	59.0%	0.45	Arabic Clued	75.0%	19.8%	45.0%	41.7%	13.3%
14	58.0%	0.45	Arabic	89.0%	36.3%	0.9%	14.5%	84.5%
15	57.9%	0.42	Arabic Clued	73.0%	55.8%	63.9%	19.5%	16.6%
16	57.2%	0.49	Arabic Clued	84.0%	52.8%	17.5%	37.5%	45.0%
17	54.4%	0.20	Arabic	71.0%	56.8%	45.9%	31.4%	22.7%
18	54.2%	0.39	Arabic	56.0%	22.8%	7.2%	15.9%	76.8%
19	50.9%	0.43	Arabic	64.0%	14.5%	13.6%	38.6%	47.7%
20	52.1%	0.52	Arabic	87.0%	10.6%	6.3%	18.8%	75.0%
21	50.1%	0.44	Arabic	69.0%	8.9%	3.7%	14.8%	81.5%
22	48.8%	0.45	Arabic Clued	73.0%	34.0%	1.0%	17.5%	81.6%
23	45.6%	0.38	Arabic	78.0%	27.7%	51.2%	31.0%	17.9%
24	45.4%	0.46	English Equivalent	72.0%	66.0%	84.5%	13.5%	2.0%
25	44.8%	0.06	Arabic	87.0%	12.5%	10.5%	21.1%	68.4%
26	28.8%	0.39	Arabic Clued	85.0%	41.9%	37.8%	20.5%	41.7%
27	43.5%	0.36	Arabic	73.0%	32.0%	2.1%	6.2%	91.8%
28	46.5%	0.40	Arabic	73.0%	27.7%	7.1%	25.0%	67.9%
29	41.2%	0.42	Arabic Clued	54.0%	50.8%	58.4%	26.0%	15.6%
30	42.0%	0.38	English Equivalent	53.0%	31.0%	30.9%	46.8%	22.3%
31	37.1%	0.45	Arabic	58.0%	38.9%	52.5%	27.1%	20.3%
32	36.1%	0.32	English Equivalent	63.0%	57.1%	83.8%	8.7%	7.5%
33	37.8%	0.22	Arabic	55.0%	33.0%	5.0%	20.0%	75.0%
34	35.8%	0.41	Arabic	71.0%	7.6%	4.3%	21.7%	73.9%
35	35.3%	0.39	Arabic Clued	56.0%	65.0%	56.3%	25.4%	18.3%
36	28.8%	0.11	Arabic Clued	69.0%	6.6%	5.0%	0.0%	95.0%
37	30.5%	0.33	Arabic	36.0%	12.9%	5.1%	15.4%	79.5%
38	34.7%	0.31	Arabic	70.0%	35.0%	18.9%	33.0%	48.1%
39	28.8%	0.11	Arabic Clued	60.0%	31.4%	4.2%	32.6%	63.2%
40	25.8%	0.32	Arabic Clued	45.0%	39.9%	49.6%	37.2%	13.2%

- The statistics shown reflect the response mode for correct answers in the Arabic version of the test

Table C.8 shows that in the items that were not totally language dependent the facility rate was due mainly to ‘known’ responses or ‘informed guess’ responses. This observation supported the conclusion that students made genuine attempts at each of the items encountered in the Arabic version of the test.

Figure C.2 shows an item with a familiar algorithm where the language does not appear to impede students’ ability to identify the problem and solve it irrespective of the language issue.

The items below have been selected to show the relative facility rates (percent correct) of the cohorts in the Arabic versions of the test session (completed first in all cases) and the English versions of the test. The purpose of showing these items is to demonstrate the relationship between the capacity of the students to attempt the Arabic items and to gauge the impact of language in students’ ability to interpret the content and correctly answer the English version of the item.

C.3.3 Observations of Responses for Selected Year 5 Items

These observations provided confidence that the analyses conducted on these data were grounded in genuine attempts to complete the test, not simply spurious random data.

Figure C.2

Year 5 Math5Q03

Year 5 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
3	Math5Q03	96%	96%	96%	96%	97%	96%

3 ما هو الرقم الناقص؟

$3 \times 8 = \boxed{?}$

(A) 11
(B) 16
(C) 24
(D) 32

Figure C.3 shows an item (Item Math5Q08) in which the responses were significantly impacted by the language barrier. The difference between the English and Arabic facility rates was noticeable. Figure C.3 shows the Arabic version of the same item (Item Math5Q08) in which the English version had a facility rate of 95%. However the Arabic version is uninterpretable by the assessed students and results with a 30% success rate; about the expected value of a random guess in the INIT and GIP analyses. When the self-identified guesses were excluded in the SIG analysis only 2% of the candidature have answered the item correctly. Six students indicated that they knew the answer. All six of these students had some Arabic language knowledge. This item highlights the impact of language on the student’s capacity to interpret the demands of a relatively simple mathematics concept as demonstrated by the English version facility.

Figure C.3

Year 5 Math5Q08 – English Version and Arabic Version

Year 5 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
8	Math5Q08	95%	95%	95%	30%	2%	31%

<p>8 The chart shows the number of visitors to a Sports Centre during four months.</p> <p>Which month had the most visitors?</p> <table border="1"> <thead> <tr> <th>Month</th> <th>Number of visitors</th> </tr> </thead> <tbody> <tr> <td>January</td> <td>6055</td> </tr> <tr> <td>February</td> <td>6505</td> </tr> <tr> <td>March</td> <td>6500</td> </tr> <tr> <td>April</td> <td>6550</td> </tr> </tbody> </table> <p> <input type="radio"/> January <input type="radio"/> February <input type="radio"/> March <input type="radio"/> April </p>	Month	Number of visitors	January	6055	February	6505	March	6500	April	6550	<p>8 بيّن الجدول أدناه عدد الزائرين إلى مركز رياضي في خلال أربعة أشهر.</p> <p>في أي شهر جاء العدد الأكبر من الزائرين؟</p> <table border="1"> <thead> <tr> <th>عدد الزائرين</th> <th>الشهر</th> </tr> </thead> <tbody> <tr> <td>6055</td> <td>يناير</td> </tr> <tr> <td>6505</td> <td>فبراير</td> </tr> <tr> <td>6500</td> <td>مارس</td> </tr> <tr> <td>6550</td> <td>أبريل</td> </tr> </tbody> </table> <p> <input type="radio"/> يناير <input type="radio"/> فبراير <input type="radio"/> مارس <input type="radio"/> أبريل </p>	عدد الزائرين	الشهر	6055	يناير	6505	فبراير	6500	مارس	6550	أبريل
Month	Number of visitors																				
January	6055																				
February	6505																				
March	6500																				
April	6550																				
عدد الزائرين	الشهر																				
6055	يناير																				
6505	فبراير																				
6500	مارس																				
6550	أبريل																				

The differences between the facility rates in the Arabic versions due to self-identified guessing are more explicit in Tables 6.21 in the Year 5 results – and Table 6.28 – in the case of the Year 7 results. In the section that follows a variety of items have been selected to demonstrate those in which language appears less of an issue.

Year 5 Mathematics Math5Q06 asked students which shape filled the image to complete the square.

Figure C.4

Year 5 Math5Q06

Year 5 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
6	Math5Q06	92%	92%	92%	75%	71%	74%






<p>6 قامت مريم بقص جزء من المربع.</p>  <p>أي من الأشكال التالية يمثل الجزء المقصوص؟</p> <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>(A)</p> </div> <div style="text-align: center;">  <p>(B)</p> </div> <div style="text-align: center;">  <p>(C)</p> </div> <div style="text-align: center;">  <p>(D)</p> </div> </div>

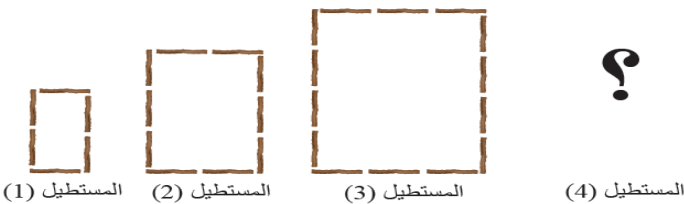
Figure C.4 shows the Arabic version of the item which was administered before the English version. When presented with the Arabic version, 75% of the students responded correctly indicating that for this item, language may be a barrier to some, but not the majority when responding to an item with a familiar context. When presented with the English equivalent, 92% of the same cohort responded correctly.

Figure C.5

Year 5 Math5Q15

Year 5 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
15	Math5Q15	73%	72%	73%	57%	46%	55%

15 استخدم سيف أعواداً صغيرة | لعمل مستطيلات كما في النمط التالي.



كم عوداً سيحتاج لتشكيل المستطيل (4) في هذا النمط؟

(A) 5 (B) 16 (C) 18 (D) 20

Figure C.5 shows relatively similar proportions of students successfully answering the item in both the Arabic and English versions. For those more-able students the language does not appear to be an issue. The variation in the facility of English INIT analysis with Arabic INIT analysis is a function of the increased number of non-attempts in the Arabic version relative to the English version.

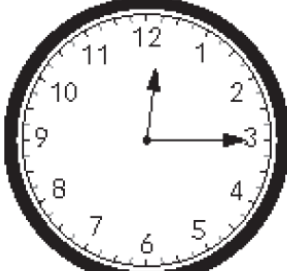
Figure C.6 shows a similar pattern with the task demand relatively intuitive and language independent.

Figure C.6

Year 5 Math5Q24

Year 5 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
24	Math5Q24	73%	73%	73%	65%	63%	63%

24 إلى أي وقت تشير الساعة أدناه؟



(A) 3:00
(B) 3:12
(C) 12:03
(D) 12:15

Figures C.7 and C.8 show varying degrees of the impact of language on the capacity of students to respond to these two items. For Item Math5Q28 the English version is relatively easy. The vast majority of students responded to the Arabic version with a self-identified random guess. The Arabic version of the item is presented below in Figure C.8.

Item Math5Q29 is interesting as there are few clues in the Arabic version, although the format is probably familiar to Year 5 Students. The English and Arabic INIT facility rates are surprisingly similar.

Figure C.7

Year 5 Math5Q28 and Math5Q29- English Versions

Year 5 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
28	Math5Q28	73%	73%	73%	28%	3%	25%
29	Math5Q29	56%	55%	53%	53%	39%	50%


28 Kate's family arrive at Global Village on Friday evening at 5:45 pm. How long can the family stay at the Village before it closes for the evening?

A 7 hours, 15 minutes

B 6 hours, 15 minutes

C 8 hours, 45 minutes

D 7 hours, 45 minutes



29 Which set of numbers divides exactly into 26 430 ?

A 2, 4, 10

B 2, 5, 9

C 3, 5, 6

D 4, 5, 10

Figure C.8

Year 5 Math5Q28 and Math5Q29 – Arabic versions



28 وصلت أسرة خديجة إلى القرية العالمية مساء يوم الجمعة في تمام الساعة 5:45 مساءً. كم من الوقت تبقى للعائلة قبل أن تغلق القرية أبوابها؟

A 7 ساعات و 15 دقيقة

B 6 ساعات و 15 دقيقة

C 8 ساعات و 45 دقيقة

D 7 ساعات و 45 دقيقة

29 ما هي مجموعة الأعداد التي تقسم العدد 26 430 تماماً؟

A 2, 4, 10

B 2, 5, 9

C 3, 5, 6

D 2, 5, 10


In the two figures below only the Arabic versions of the items are displayed. This is because the context is relatively familiar and the English version easily interpretable from the graphics.

Figure C.9

Year 5 Math5Q30

Year 5 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
30	Math5Q30	55%	50%	51%	32%	13%	32%

30 إذا كان الجزء المظلل من الشكل المجاور يمثل $\frac{1}{3}$:



فماذا يمثل الجزء المظلل في الشكل التالي؟

A $\frac{2}{3}$
 B $\frac{1}{2}$
 C $\frac{3}{2}$
 D $\frac{2}{6}$

Figure C.10

Year 5 Math5Q35

Year 5 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
35	Math5Q35	58%	55%	53%	66%	52%	64%

35 يعرف راشد أن $24 \div 4 = 6$

أي عملية حسابية يمكنه أن يحسب باستخدام هذه الحقيقة؟

$4 \times 6 =$ B
 $6 \div 4 =$ A

$6 \times 24 =$ D
 $24 \times 4 =$ C

Math5Q35 is an anomaly with a higher facility rate in the Arabic version than the English.

However this may be explained in part by the position of the correct answer for English medium students and the fact that the first position is the correct answer.

When considering the data and examples above, the relationships between the capacity to interpret and respond to the demands of these language independent questions suggested that responses were not random and had some association with mathematical ability. One aspect of these examples, and the results shown in Table 6.18 is the similarity in the proportion of correct responses in the INIT analysis and the GIP analysis. In most cases these facility rates are within one or two percent of each other with the GIP result typically being marginally lower than the INIT result. In the items in which language is a factor, the rates tend to be approximately a random guessing rate of 20% to 30%. This may be a function of the Arabic test being far too 'hard' of the students due to the language factor.

This 'too hard' attribute of the test may result in these data being effectively random data. Scholars have noted that the Rasch model is appropriate when data fit the model: i.e the model produces reliable statistics. In cases where the fit is poor it is recommended that the Rasch model not be used to analyse the data.

C.4 Analysis of the Year 7 Arabic versions

C.4.1 Investigation of Year 7 Items

Table C.9 shows the overall facility rates achieved for each language version of the test and the three analyses. The table also provides an annotation regarding the nature and form of the question and its impact on the variation due to language dependence in the item.

Table C.9

Year 7 Comparison of Item Facility by Analysis Phase

Year 7 Mathematics		English				Arabic			Language demand
Seq	Item Label	ORIGINAL	INIT	SIG	GIP _{p0.6}	INIT	SIG9	GIP _{p0.6}	
1	Math7Q01	92.7%	98%	98%	98%	91%	89%	91%	Image interpretation
2	Math7Q02	86.7%	99%	99%	99%	99%	88%	89%	Algebraic Equation (simple)
3	Math7Q03	76.5%	93%	93%	93%	73%	30%	73%	
4	Math7Q04	84.7%	98%	98%	98%	88%	83%	86%	Image interpretation
5	Math7Q05	72.6%	83%	82%	83%	79%	73%	79%	Evaluate indices - no language
6	Math7Q06	60.3%	90%	89%	90%	37%	*	26%	
7	Math7Q07	70.3%	80%	79%	79%	35%	13%	21%	
8	Math7Q08	68.1%	98%	98%	98%	89%	88%	88%	Arithmetic sequence
9	Math7Q09	64.3%	73%	70%	72%	32%	2%	14%	
10	Math7Q10	85.5%	73%	71%	72%	57%	46%	56%	Image interpretation
11	Math7Q11	70.6%	94%	93%	94%	82%	80%	82%	Image interpretation
12	Math7Q12	43.4%	94%	94%	94%	23%	*	6%	
13	Math7Q13	59.1%	63%	60%	60%	43%	15%	33%	
14	Math7Q14	57.6%	75%	74%	74%	38%	4%	23%	
15	Math7Q15	57.3%	84%	84%	84%	17%	2%	1%	
16	Math7Q16	56.2%	75%	73%	74%	46%	34%	41%	Image interpretation
17	Math7Q17	54.7%	79%	78%	77%	23%	*	7%	
18	Math7Q18	53.9%	82%	80%	82%	51%	9%	48%	
19	Math7Q19	52.9%	8%	5%	78%	18%	39%	33%	
20	Math7Q20	52.7%	93%	92%	93%	73%	66%	69%	Unit conversion km to m
21	Math7Q21	50.6%	75%	74%	75%	11%	*	2%	
22	Math7Q22	48.6%	84%	83%	84%	14%	*	2%	
23	Math7Q23	41.2%	90%	89%	90%	73%	64%	71%	Image interpretation
24	Math7Q24	48.1%	76%	74%	75%	65%	52%	62%	Image interpretation
25	Math7Q25	45.6%	76%	73%	75%	77%	67%	73%	Image interpretation
26	Math7Q26	44.2%	69%	60%	67%	25%	4%	6%	
27	Math7Q27	43.9%	83%	82%	83%	77%	69%	70%	Add fractions - algorithm
28	Math7Q28	41.3%	31%	27%	12%	38%	1%	5%	
29	Math7Q29	40.1%	35%	26%	20%	30%	*	13%	
30	Math7Q30	39.5%	68%	61%	67%	54%	28%	53%	Image interpretation
31	Math7Q31	37.9%	52%	49%	50%	17%	1%	*	
32	Math7Q32		56%	45%	53%	44%	*	35%	
33	Math7Q33	30.8%	83%	83%	83%	17%	2%	*	
34	Math7Q34	17.7%	51%	49%	45%	37%	23%	27%	Image interpretation
35	Math7Q35	33.0%	68%	64%	67%	9%	2%	1%	
36	Math7Q36	25.8%	63%	61%	63%	54%	26%	41%	Image interpretation
37	Math7Q37	31.4%	62%	52%	62%	46%	20%	39%	Image interpretation
38	Math7Q38	17.8%	20%	10%	80%	30%	*	15%	
39	Math7Q39	30.8%	77%	75%	77%	23%	2%	*	
40	Math7Q40	21.0%	71%	68%	70%	49%	20%	23%	Image interpretation
41	Math7Q41		50%	40%	41%	37%	2%	10%	
42	Math7Q42	20.1%	42%	39%	35%	9%	*	*	
43	Math7Q43	14.8%	47%	45%	40%	36%	17%	8%	Sequence
44	Math7Q44	19.2%	57%	47%	55%	50%	29%	35%	Image interpretation

Table C.10 shows the breakdown of the self-identified guessing patterns for each item.

Table C.10

Year 7 Comparison of Item Facility INIT Analysis of English and Arabic Versions of Test

Year 7 Item no	Maths Facility Rate	Original source Language Dependency	INIT responses		Response mode correct Arabic*		
			English Version	Arabic Version	Known Answer	Guessed Informed	Guessed Random
1	92.7%	English Equiv	98.0%	91.0%	62.5%	34.7%	2.8%
2	86.7%	English Equiv	99.0%	99.0%	85.9%	14.1%	0.0%
3	76.5%	Arabic	93.0%	73.0%	16.8%	32.7%	50.4%
4	84.7%	Arabic clued	98.0%	88.0%	67.9%	27.0%	5.1%
5	72.6%	English Equiv	83.0%	79.0%	85.4%	9.8%	4.9%
6	60.3%	Arabic	90.0%	37.0%	0.0%	32.1%	67.9%
7	70.3%	Arabic	80.0%	35.0%	29.6%	55.6%	14.8%
8	68.1%	Arabic clued	98.0%	89.0%	85.7%	10.0%	4.3%
9	64.3%	Arabic	73.0%	32.0%	4.2%	10.4%	85.4%
10	85.5%	Arabic clued	73.0%	57.0%	71.9%	12.4%	15.7%
11	70.6%	Arabic clued	94.0%	82.0%	91.5%	2.3%	6.2%
12	43.4%	Arabic	94.0%	23.0%	0.0%	2.9%	97.1%
13	59.1%	Arabic	63.0%	43.0%	25.0%	30.9%	44.1%
14	57.6%	Arabic	75.0%	38.0%	6.9%	5.2%	87.9%
15	57.3%	Arabic	84.0%	17.0%	7.7%	3.8%	88.5%
16	56.2%	Arabic clued	75.0%	46.0%	62.5%	30.6%	6.9%
17	54.7%	Arabic	79.0%	23.0%	0.0%	8.6%	91.4%
18	53.9%	Arabic	82.0%	51.0%	11.4%	53.2%	35.4%
19	52.9%	Arabic clued	79.0%	44.0%	79.7%	4.3%	15.9%
20	52.7%	English Equiv	93.0%	73.0%	70.8%	24.8%	4.4%
21	50.6%	Arabic	75.0%	11.0%	0.0%	12.5%	87.5%
22	48.6%	Arabic	84.0%	14.0%	0.0%	40.9%	59.1%
23	41.2%	Arabic clued	90.0%	73.0%	62.6%	26.1%	11.3%
24	48.1%	Arabic clued	76.0%	65.0%	65.0%	22.0%	13.0%
25	45.6%	English Equiv	76.0%	77.0%	70.1%	25.6%	4.3%
26	44.2%	Arabic	69.0%	25.0%	13.5%	18.9%	67.6%
27	43.9%	English Equiv	83.0%	77.0%	87.6%	8.8%	3.5%
28	41.3%	Arabic	31.0%	20.0%	6.7%	0.0%	93.3%
29	40.1%	Arabic	35.0%	30.0%	0.0%	22.7%	77.3%
30	39.5%	Arabic clued	68.0%	54.0%	38.8%	18.8%	42.5%
31	37.9%	Arabic	52.0%	17.0%	8.3%	25.0%	66.7%
32		Arabic	56.0%	44.0%	0.0%	3.3%	96.7%
33	30.8%	Arabic	83.0%	17.0%	8.3%	33.3%	58.3%
34	17.7%	Arabic	51.0%	37.0%	51.9%	25.9%	22.2%
35	33.0%	Arabic	68.0%	9.0%	15.4%	46.2%	38.5%
36	25.8%	Arabic clued	63.0%	54.0%	41.9%	47.3%	10.8%
37	31.4%	Arabic clued	62.0%	46.0%	46.6%	15.5%	37.9%
38	17.8%	Arabic	70.0%	25.0%	5.6%	8.3%	86.1%
39	30.8%	Arabic clued	77.0%	23.0%	7.1%	21.4%	71.4%
40	21.0%	Arabic clued	71.0%	49.0%	43.3%	48.3%	8.3%
41		Arabic clued	50.0%	37.0%	4.3%	4.3%	91.3%
42	20.1%	Arabic	42.0%	9.0%	0.0%	10.0%	90.0%
43	14.8%	Arabic clued	47.0%	36.0%	55.8%	30.2%	14.0%
44	19.2%	Arabic clued	57.0%	50.0%	38.7%	32.3%	29.0%

* The statistics shown reflect the response mode for correct answers in the Arabic version of the test

As observed in the Year 5 sample the facility rates and response modes for, the Arabic versions of the tests supported the contention that the students have made serious attempts in answering the items during both versions of the tests.

C.4.2 Observations of Responses for Selected Year 7 Items

As for the Year 5 section a few items have been selected to demonstrate various response patterns by the cohort to items of varying language dependency.


Figure C.11

Math7Q12– Arabic Version

Year 7 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
12	Math7Q12	94%	94%	94%	23%	N/A	6%

12

لدى لطيفة الأشكال التالية المقتطعة من بطاقة.



استخدمت هذه الأشكال لتحصل على نموذج ثلاثي الأبعاد.
ما هو اسم النموذج الذي حصلت عليه؟

(A) مخروط
 (B) اسطوانة
 (C) منشور رباعي
 (D) كرة

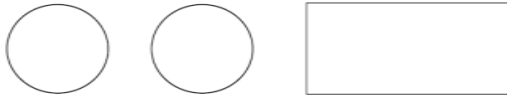
In the English version of this item only two students omitted the question, with 94% of the remaining 187 students responding correctly. In the Arabic version 23% ‘guessed’ correctly – approximately $\frac{1}{k}$. Only four students omitted the item in the Arabic version which may be an indicator of the propensity to guess by students in an assessment regime with no downside from an incorrect guess.

No students answered this item correctly without an indicated guess in the Arabic version of the SIG analysis. As a consequence, the item was removed from the SIG analysis as a result of the recoding of all the guessed answers to have ‘missing’ values. In the Rasch analysis model an item with no correct responses is ‘extreme’ and omitted from the analysis. Hence there are no statistics for this item in the SIG analysis.

Figure C.12

Math7Q12– English Version

12 Latifa has these shapes cut out of card.



She uses them to make a three-dimensional model.
What is the name of the model she makes?

(A) cone
(B) cylinder
(C) rectangular prism
(D) sphere

Figure C.13

Math7Q01– Arabic Version

Year 7 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
1	Math7Q01	98%	98%	98%	91%	89%	91%

1 يخطط حمدان للذهاب في نزهة يوم غد.
تبين التوقعات الجوية أن درجة الحرارة غدا ستكون 24°C .
أي مما يلي تبين درجة حرارة 24°C ؟



(A) (B) (C) (D)

Figure C.13 also shows the relatively minor impact of language in an item that is common in its presentation and where the task demands of the item are conveyed independent of language. The minor increase in the facility rate observed by the SIG analysis reflects the reduction in the number of students attempting the item when self-identified guesses were suppressed in the SIG analysis.

Figure C.14 also shows the negligible impact of language on an item that is common in its algorithm and the task demands are conveyed in a mode relatively independent of language.

Figure C.14*Math7Q02– Arabic Version*

Year 7 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
2	Math7Q02	99%	99%	99%	99%	88%	89%

2 إذا كان $7x = 21$ فما قيمة x ؟

A 2 B 3 C 14 D 28

Figures C.15 and C.16 show two examples of the various item representations that are familiar to Year 7 students and consequently the impact of language is reduced with a subsequent decrease in the proportion of random guessing for these items.

Figure C.15*Math7Q05– Arabic Version*

Year 7 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
5	Math7Q05	83%	82%	83%	79%	73%	79%

5 أي الأعداد التالية يمكن التعبير عنه بـ 3×2^3 ؟

A 15
 B 18
 C 24
 D 48

Figure C.16

Math7Q04 – Arabic Version

Year 7 Mathematics		English			Arabic		
Seq	Item Label	INIT	SIG	GIP	INIT	SIG	GIP
4	Math7Q04	98%	98%	98%	88%	83%	86%



4 لدى طارق التذكرة الموضحة بالشكل المجاور لحضور عرض فيلم سينمائي.



ويوضح المخطط مقعد كل من أحمد وطارق.

أحمد طارق

5	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	A	B	C	D	E	F	G	H	I	J

أي شكل مما يلي يمثل تذكرة أحمد؟

A  B 

C  D 

The facility rates shown in the items displayed throughout Appendix C suggested that students did not generally randomly guess the answer when there were clues to the demands of familiar items. However, the differences in the facility rates of the English and Arabic versions of these items indicate that random guessing was occurring. Further the relative similarities between the English SIG and English GIP facility rates would suggest that these processes were functioning similarly and the scale of differences between the INIT analysis facility rates and the conditioned rates of the SIG and GIP were relatively consistent.

APPENDIX D

Investigation of Alternative p Values

D.1 Frequency Analysis – GIP Indicated Guesses by Item by p Value

D.1.1 SIM1

The tables below provide an analysis of the proportion of items, sorted by item difficulty, indicated by the GIP procedure as probable guesses for SIM1. In each case a supplementary analysis of the comparison of initial defined guesses, presented in Table D.1, with each of the GIP tables developed, there were no cases of items which were not defined as a guess being indicated as a guess by the GIP procedure. This validates the rigour of the procedure as an arithmetic protocol however as indicated in Chapter 11.6, issues apart from the simple arithmetic calculations should be considered.

Table D.1 SIM 1 Table of Defined Guesses by Ability Quartile and Item Difficulty

δ	-2.66	-1.69	-1.73	-1.75	-1.35	-1.47	-1.13	-1.22	-1.01	-1.03	-1.1	-0.68	-0.8	-0.72	-0.6	-0.67	-0.58	-0.43	-0.27	-0.04	0.123	0.233	0.477	0.665	0.801	0.899	0.842	0.898	0.941	0.928	1.026	1.233	1.153	1.142	1.527	1.283	1.58	1.594	1.697	1.874	
Quartile	Q01	Q04	Q03	Q02	Q06	Q05	Q08	Q07	Q11	Q10	Q09	Q13	Q14	Q12	Q16	Q15	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q27	Q28	Q26	Q30	Q29	Q31	Q34	Q33	Q32	Q36	Q35	Q37	Q38	Q39	Q40	Total
Q4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	1	11	4	5	9	14	12	60	
Q3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4	8	15	21	14	13	24	22	23	30	22	28	19	23	24	23	18	24	357	
Q2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	6	6	12	13	16	17	27	33	18	21	23	24	27	26	23	26	31	20	33	21	18	28	27	20	519
Q1	4	12	5	3	14	9	12	6	22	16	11	18	23	12	31	23	26	31	22	26	26	29	25	19	25	27	33	23	29	29	28	29	27	23	26	23	26	20	24	18	835
ALL def.	4	12	5	3	14	9	12	6	22	16	11	18	23	12	32	25	32	37	34	39	42	48	56	60	58	69	70	60	80	77	74	87	82	72	89	71	73	80	83	74	1771

Table D.1 shows the Guttman-like pattern of the defined guesses imbedded in the SIM1 data. The marginal variations in the number of defined guesses compared to the item difficulty is a function of the randomisation process by which the guesses were embedded in these data.

Table D.2 SIM 1 Table of GIP Indicated Guesses for p = 0.6 by Ability Quartile and Item Difficulty

δ	-2.66	-1.69	-1.73	-1.75	-1.35	-1.47	-1.13	-1.22	-1.01	-1.03	-1.1	-0.68	-0.8	-0.72	-0.6	-0.67	-0.58	-0.43	-0.27	-0.04	0.123	0.233	0.477	0.665	0.801	0.899	0.842	0.898	0.941	0.928	1.026	1.233	1.153	1.142	1.527	1.283	1.58	1.594	1.697	1.874			
Quartile	Q01	Q04	Q03	Q02	Q06	Q05	Q08	Q07	Q11	Q10	Q09	Q13	Q14	Q12	Q16	Q15	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q27	Q28	Q26	Q30	Q29	Q31	Q34	Q33	Q32	Q36	Q35	Q37	Q38	Q39	Q40	Total		
Q4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	1	3	0	5	9	9	12	17	59		
Q2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	2	1	6	3	6	7	7	14	14	9	19	17	15	25	25	18	195				
Q1	0	0	0	0	0	0	0	0	1	1	1	3	2	0	4	1	1	7	9	16	20	23	23	17	24	26	33	23	29	29	28	29	27	23	26	23	26	20	24	18	537		
ALL p0.60	0	0	0	0	0	0	0	0	1	1	1	3	2	0	4	1	1	7	9	16	20	23	25	22	26	27	39	28	35	36	35	44	42	35	45	45	50	54	61	53	791		

As indicated in the body of the thesis the difference between the ability estimate of the student and the difficulty location of the item needs to be greater than 1.1 logits to indicate a probability of success on the items less than 0.25. In cases where this threshold is not reached the GIP procedure does not indicate a probable guess. Table D.2 highlights that in the easy items (up to item Q07) there are no instances in which the ability estimates as adjusted by the $p = 0.6$ condition are more than 1.1 logits below the item difficulty. Hence no items are indicated in any quartile as likely guessed items. As the item difficulties increase (student ability estimates remain stable) the occurrence in which the 1.1 logit difference threshold was reached increases with increasing numbers of items indicated as likely guesses. The indication of guesses also increased linearly as ability reduces as indicated by the Guttman-like pattern for each quartile in Table D.2.

Table D.3 SIM 1 Table of GIP Indicated Guesses for $p = 0.62$ by Ability Quartile and Item Difficulty

δ	-2.66	-1.69	-1.73	-1.75	-1.35	-1.47	-1.13	-1.22	-1.01	-1.03	-1.1	-0.68	-0.8	-0.72	-0.6	-0.67	-0.58	-0.43	-0.27	-0.04	0.123	0.233	0.477	0.665	0.801	0.899	0.842	0.898	0.941	0.928	1.026	1.233	1.153	1.142	1.527	1.283	1.58	1.594	1.697	1.874			
Quartile	Q01	Q04	Q03	Q02	Q06	Q05	Q08	Q07	Q11	Q10	Q09	Q13	Q14	Q12	Q16	Q15	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q27	Q28	Q26	Q30	Q29	Q31	Q34	Q33	Q32	Q36	Q35	Q37	Q38	Q39	Q40	Total		
Q4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	1	1	5	0	7	9	12	12	21	71		
Q2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	5	3	3	6	3	11	10	7	14	14	16	19	18	15	26	25	20	218			
Q1	0	0	0	1	0	0	0	0	3	2	2	3	4	0	7	1	4	10	12	16	23	23	23	17	25	27	33	23	29	29	28	29	27	23	26	23	26	20	24	18	561		
All p0.62	0	0	0	1	0	0	0	0	3	2	2	3	4	0	7	1	4	10	12	16	24	23	25	22	28	30	39	28	41	39	35	44	42	44	45	48	50	58	61	59	850		

Table D.3 shows the impact of increasing the p value to a value of 0.62 compared to 0.60. The effect is to adjust the ability estimate required to indicate success on the item by a value of 0.490 logits. Table D.3 displays a marginal increase in the number of item/student interactions which are indicated as guesses by the GIP procedure. The pattern of increasing indication was relatively uniform across items and quartiles with some randomness in the number of interactions. This was a function of; no change in the outcome for items previously indicated and the threshold being reached for some item/ability interactions at different situations in the matrix. The general pattern of; the lower the student ability estimate the greater the number of items indicated by the GIP procedure was maintained.

Table D.4 SIM1 Table of GIP Indicated Guesses for $p = 0.65$ by Ability Quartile and Item Difficulty

δ	-2.66	-1.69	-1.73	-1.75	-1.35	-1.47	-1.13	-1.22	-1.01	-1.03	-1.1	-0.68	-0.8	-0.72	-0.6	-0.67	-0.58	-0.43	-0.27	-0.04	0.123	0.233	0.477	0.665	0.801	0.899	0.842	0.898	0.941	0.928	1.026	1.233	1.153	1.142	1.527	1.283	1.58	1.594	1.697	1.874			
Quartile	Q01	Q04	Q03	Q02	Q06	Q05	Q08	Q07	Q11	Q10	Q09	Q13	Q14	Q12	Q16	Q15	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q27	Q28	Q26	Q30	Q29	Q31	Q34	Q33	Q32	Q36	Q35	Q37	Q38	Q39	Q40	Total		
Q4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	4	2	0	0	4	2	7	2	14	12	12	13	21	94		
Q2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	7	5	5	8	4	13	12	8	19	24	17	25	20	17	26	25	20	258			
Q1	0	0	0	1	0	0	1	0	3	2	3	5	9	3	13	5	4	10	14	22	23	25	24	19	25	27	33	23	29	29	28	29	27	23	26	23	26	20	24	18	596		
All p0.65	0	0	0	1	0	0	1	0	3	2	3	5	9	3	13	5	4	10	14	22	24	25	26	26	31	32	41	31	44	41	36	52	53	47	53	57	55	58	62	59	948		

Table D.5 SIM1 Table of GIP Indicated Guesses for $p = 0.70$ by Ability Quartile and Item Difficulty

δ	-2.66	-1.69	-1.73	-1.75	-1.35	-1.47	-1.13	-1.22	-1.01	-1.03	-1.1	-0.68	-0.8	-0.72	-0.6	-0.67	-0.58	-0.43	-0.27	-0.04	0.123	0.233	0.477	0.665	0.801	0.899	0.842	0.898	0.941	0.928	1.026	1.233	1.153	1.142	1.527	1.283	1.58	1.594	1.697	1.874			
Quartile	Q01	Q04	Q03	Q02	Q06	Q05	Q08	Q07	Q11	Q10	Q09	Q13	Q14	Q12	Q16	Q15	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q27	Q28	Q26	Q30	Q29	Q31	Q34	Q33	Q32	Q36	Q35	Q37	Q38	Q39	Q40	Total		
Q4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	6	4	4	6	1	6	9	7	11	14	18	18	17	17	24	165		
Q2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	6	11	7	12	15	15	17	20	16	23	27	19	31	21	18	28	27	20	338		
Q1	0	0	0	2	3	1	2	0	4	3	4	7	13	4	16	10	11	18	18	23	23	28	25	19	25	27	33	23	29	29	28	29	27	23	26	23	26	20	24	18	644		
All $p=0.70$	0	0	0	2	3	1	2	0	4	3	4	7	13	4	16	10	11	18	18	24	25	31	31	30	34	45	52	42	52	50	50	61	61	53	71	62	62	65	68	62	1147		

Tables D.4 and D.5 above display the outcomes of increasing the p value to 0.65 and 0.7 respectively. The impact on the effective ability estimates of the students is to adjust the $p = 0.5$ ability estimate by 0.619 and 0.847 logits respectively. Both tables show the increasing numbers of items indicated by the protocol as expected and maintain the pattern of increasing indications as item difficulties increase and student ability is relatively lower.

D.1.2 SIM3

The tables (Tables D.6 to D.10) below follow the same pattern as presented for alternative RP calculations of the SIM1 data but provide greater differentiation due to the increased sample size (1000 students) and the consequent increase in the number of deciles represented in the tables. Table D.6 follows the same design as Table D.1 with the embedded defined guesses in a Guttman pattern with some variations due to the impact of the randomising algorithm employed to embed the guessed responses.

Table D.6 SIM 3 Table of Defined Guesses by Ability Decile and Item Difficulty

δ	-2.23	-1.91	-1.9	-1.88	-1.7	-1.68	-1.68	-1.4	-1.27	-1.06	-1.06	-0.98	-0.89	-0.78	-0.59	-0.37	-0.37	-0.36	-0.22	0.078	0.369	0.375	0.542	0.567	0.841	0.876	0.88	0.953	1.005	1.051	1.07	1.085	1.337	1.426	1.51	1.531	1.56	1.636	1.795	1.84	Total			
Decile	Q03	Q02	Q07	Q05	Q01	Q04	Q09	Q11	Q06	Q10	Q13	Q08	Q12	Q15	Q17	Q18	Q16	Q14	Q19	Q20	Q21	Q22	Q24	Q23	Q26	Q25	Q27	Q28	Q30	Q31	Q29	Q32	Q34	Q33	Q38	Q35	Q36	Q37	Q40	Q39				
Decile 10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Decile 9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	6	13	22	15	74		
Decile 8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	5	0	11	20	14	21	18	29	26	28	26	201			
Decile 7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	11	19	23	28	23	24	28	26	26	30	24	16	26	22	330			
Decile 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	26	25	22	21	27	12	29	32	33	20	27	29	26	20	24	21	402			
Decile 5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	13	25	13	24	24	27	21	28	30	24	28	31	22	26	23	28	28	18	24	459				
Decile 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	12	21	24	22	23	24	19	24	24	25	28	24	17	25	20	27	25	23	24	17	27	23	501				
Decile 3	0	0	0	0	0	0	0	0	0	1	0	0	10	24	25	19	8	27	24	18	26	32	30	22	33	32	23	24	23	37	32	28	20	30	20	21	28	24	18	659				
Decile 2	0	0	1	0	0	0	6	14	0	12	29	6	23	25	24	21	30	31	22	23	29	23	33	16	26	17	28	26	30	30	19	26	18	23	33	22	23	24	21	18	752			
Decile 1	10	10	20	22	6	9	31	20	22	33	25	24	22	33	22	25	16	27	23	20	17	23	29	31	35	19	21	29	24	30	20	31	27	22	27	27	23	29	19	23	926			
ALL defined	10	10	21	22	6	9	37	34	22	45	55	30	45	68	70	74	65	66	84	88	90	107	150	114	156	142	165	164	187	182	169	209	205	174	233	192	204	201	209	190	4304			

Table D.7 is a replicate of Table 5.14 showing the outcomes of the application of the RP of 0.6 used throughout the study compared to the potential outcomes of alternative values. As observed in SIM1 the table shows that the GIP procedure indicates guessed items in a relatively linear pattern increasingly as ability estimates trend lower and items difficulties trend higher. The protocol is shown to be most efficient in the lower ability groups and more commonly in the more difficult items which is the anticipated outcome and follows the logic of the study which assumes that as items become harder, less-able students will employ guessing techniques to account for the lack of certain knowledge or skills.

Table D.7 Table of GIP Indicated Guesses for $p = 0.6$ by Ability Decile and Item Difficulty

δ	-2.23	-1.91	-1.9	-1.88	-1.7	-1.68	-1.68	-1.4	-1.27	-1.06	-1.06	-0.98	-0.89	-0.78	-0.59	-0.37	-0.37	-0.36	-0.22	0.078	0.369	0.375	0.542	0.567	0.841	0.876	0.88	0.953	1.005	1.051	1.07	1.085	1.337	1.426	1.51	1.531	1.56	1.636	1.795	1.84	Total				
Decile	Q03	Q02	Q07	Q05	Q01	Q04	Q09	Q11	Q06	Q10	Q13	Q08	Q12	Q15	Q17	Q18	Q16	Q14	Q19	Q20	Q21	Q22	Q24	Q23	Q26	Q25	Q27	Q28	Q30	Q31	Q29	Q32	Q34	Q33	Q38	Q35	Q36	Q37	Q40	Q39					
Decile 10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Decile 9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Decile 8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Decile 7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	12	28	
Decile 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	19	17	19	24	25	19	128				
Decile 5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	25	17	26	23	24	19	18	177				
Decile 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	1	4	9	7	17	16	21	22	22	25	19	14	18	19	18	237					
Decile 3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	9	19	25	31	20	30	20	32	30	23	18	27	17	24	27	21	21	407					
Decile 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	8	28	25	31	18	23	19	23	28	23	31	24	28	22	20	34	26	26	21	24	18	502					
Decile 1	0	0	0	0	1	1	1	0	3	3	3	4	2	9	7	21	14	23	15	20	17	20	24	27	34	17	22	24	23	27	15	26	22	22	25	19	17	25	19	20	572				
ALL p = 0.60	0	0	0	0	1	1	1	0	3	3	3	4	2	9	7	21	14	23	17	28	45	45	68	54	81	62	80	81	83	95	87	105	116	110	147	124	123	139	143	126	2051				

Tables D.8, D.9 and D.10 below provide the outcomes of the application of the RP values of 0.62, 0.65 and 0.70 respectively. As anticipated by the hypothesis they show an increasing number of item/student interactions being indicated as a probable guess as item difficulty increases and student ability reduces.

Table D.8 Table of GIP Indicated Guesses for $p = 0.62$ by Ability Quartile and Item Difficulty

δ	-2.23	-1.91	-1.9	-1.88	-1.7	-1.68	-1.68	-1.4	-1.27	-1.06	-1.06	-0.98	-0.89	-0.78	-0.59	-0.37	-0.37	-0.36	-0.22	0.078	0.369	0.375	0.542	0.567	0.841	0.876	0.88	0.953	1.005	1.051	1.07	1.085	1.337	1.426	1.51	1.531	1.56	1.636	1.795	1.84	Total			
Decile	Q03	Q02	Q07	Q05	Q01	Q04	Q09	Q11	Q06	Q10	Q13	Q08	Q12	Q15	Q17	Q18	Q16	Q14	Q19	Q20	Q21	Q22	Q24	Q23	Q26	Q25	Q27	Q28	Q30	Q31	Q29	Q32	Q34	Q33	Q38	Q35	Q36	Q37	Q40	Q39				
Decile 10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Decile 9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Decile 8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	4		
Decile 7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	2	3	8	9	6	15	20	18	83			
Decile 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	2	1	0	0	5	4	8	14	18	23	26	21	23	17	165			
Decile 5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	1	1	4	5	5	2	6	13	17	21	24	15	18	25	22	23	206			
Decile 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	13	15	12	14	15	13	19	17	22	20	31	26	20	20	19	21	302			
Decile 3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	6	5	22	15	23	21	23	22	25	23	26	32	23	19	23	19	26	22	16	19	413			
Decile 2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	7	4	6	6	11	23	22	30	23	25	18	23	31	24	27	23	25	19	22	34	23	21	18	29	21	517			
Decile 1	0	0	0	0	1	1	1	0	5	3	3	4	2	14	12	13	10	17	11	16	16	18	27	24	25	16	24	19	24	30	18	28	24	21	24	20	19	31	19	18	558			
ALL $p = 0.62$	0	0	0	0	1	1	1	0	5	3	3	4	3	15	13	21	14	23	17	28	45	45	83	68	88	72	88	92	94	95	97	120	116	120	163	135	136	153	149	137	2248			

Table D.9 Table of GIP Indicated Guesses for $p = 0.65$ by Ability Quartile and Item Difficulty

δ	-2.23	-1.91	-1.9	-1.88	-1.7	-1.68	-1.68	-1.4	-1.27	-1.06	-1.06	-0.98	-0.89	-0.78	-0.59	-0.37	-0.37	-0.36	-0.22	0.078	0.369	0.375	0.542	0.567	0.841	0.876	0.88	0.953	1.005	1.051	1.07	1.085	1.337	1.426	1.51	1.531	1.56	1.636	1.795	1.84	Total			
Decile	Q03	Q02	Q07	Q05	Q01	Q04	Q09	Q11	Q06	Q10	Q13	Q08	Q12	Q15	Q17	Q18	Q16	Q14	Q19	Q20	Q21	Q22	Q24	Q23	Q26	Q25	Q27	Q28	Q30	Q31	Q29	Q32	Q34	Q33	Q38	Q35	Q36	Q37	Q40	Q39				
Decile 10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Decile 9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Decile 8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	3	6	12			
Decile 7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	0	2	0	0	2	5	5	9	16	10	18	26	26	123			
Decile 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	1	3	2	2	1	5	10	15	21	22	30	33	22	23	17	210				
Decile 5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	3	4	7	10	11	11	6	25	23	21	27	17	19	25	22	23	258			
Decile 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	6	5	16	20	15	19	17	13	19	18	25	21	31	26	20	20	19	21	332			
Decile 3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	3	6	8	26	17	23	22	23	22	25	24	26	33	23	19	23	19	26	22	16	19	427			
Decile 2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	9	7	9	8	13	25	24	31	24	25	18	23	31	24	27	23	25	19	22	34	23	21	18	29	21	537			
Decile 1	0	1	1	0	1	1	2	1	5	7	3	7	5	14	17	15	12	18	14	16	16	18	27	24	25	16	24	19	24	30	18	28	24	21	24	20	19	31	19	18	585			
ALL $p = 0.65$	0	1	1	0	1	1	2	1	5	7	3	7	6	15	20	25	19	27	22	32	48	50	92	74	94	82	97	103	105	106	97	141	135	130	171	151	148	157	157	151	2484			

Table D.10 Table of GIP Indicated Guesses for $p = 0.70$ by Ability Quartile and Item Difficulty

δ	-2.23	-1.91	-1.9	-1.88	-1.7	-1.68	-1.68	-1.4	-1.27	-1.06	-1.06	-0.98	-0.89	-0.78	-0.59	-0.37	-0.37	-0.36	-0.22	0.078	0.369	0.375	0.542	0.567	0.841	0.876	0.88	0.953	1.005	1.051	1.07	1.085	1.337	1.426	1.51	1.531	1.56	1.636	1.795	1.84	Total			
Decile	Q03	Q02	Q07	Q05	Q01	Q04	Q09	Q11	Q06	Q10	Q13	Q08	Q12	Q15	Q17	Q18	Q16	Q14	Q19	Q20	Q21	Q22	Q24	Q23	Q26	Q25	Q27	Q28	Q30	Q31	Q29	Q32	Q34	Q33	Q38	Q35	Q36	Q37	Q40	Q39				
Decile 10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Decile 9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Decile 8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	1	3	4	9	10	31		
Decile 7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	4	1	3	6	5	3	16	20	15	23	19	22	30	26	198		
Decile 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	4	10	10	9	12	11	11	14	31	25	23	31	36	22	23	17	292			
Decile 5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	2	4	10	15	20	18	29	19	21	29	26	22	27	18	20	25	22	23	353			
Decile 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	12	7	20	29	17	19	24	19	22	19	25	21	31	26	20	20	19	21	378			
Decile 3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	2	9	14	24	28	18	23	22	23	22	26	24	26	33	23	19	23	19	26	22	16	19	466			
Decile 2	0	0	0	0	0	0	0	0	0	0	2	1	2	4	3	16	17	18	15	17	27	25	31	24	25	18	23	31	24	27	23	25	19	22	34	23	21	18	29	21	585			
Decile 1	0	2	2	1	7	6	4	2	10	18	6	15	7	22	19	17	15	24	17	16	16	18	27	24	25	16	24	19	24	30	18	28	24	21	24	20	19	31	19	18	655			
ALL $p=0.7$	0	2	2	1	7	6	4	2	10	18	8	17	9	26	23	34	33	43	34	42	64	71	100	79	109	113	121	119	142	136	126	151	165	151	179	161	164	164	167	155	2958			

As mentioned in the Conclusion, Chapter 11 Section 11. 6 the arithmetic adjustment of the RP value is not the sole consideration in the development of an effective and efficient protocol. It is considered that although the increasing value of the RP provides higher recovery rates of the defined guesses in the simulated data in authentic live data, other psychometric factors also should be considered. Section D.2 below addresses some of these issues.

D.2 Consideration of Alternate Response Probabilities

Given that this study is grounded in an assumption that a probability of success on an item with a four-distractor structure is 25% (0.25) the difference between item difficulty location and person ability estimate that is the limiting value for that threshold is 1.1 logits in a Rasch measurement scale. This difference represents the equivalent of approximately two years and two months of learning in a student's learning progression (Cohen, 1985).

Table D11 shows the interaction between the adjustment of the RP and the residual difference in learning age for different p value. It can be seen that the impact of increasing the p value reduces the contribution to the response that reflects learning of the student.

Table D.11 Consideration of Impact of Alternative RP values

RP	Ability Adjustment	Other factors	Approximate Effective learning difference (yrs, mths)
0.50	0	1.100	2 years, 2 months
0.60	0.405	0.695	1 year, 6 months
0.62	0.490	0.610	1 year, 4 months
0.65	0.619	0.481	11 months
0.70	0.847	0.253	6 months

The issue that requires qualitative research beyond the scope of this study would be to investigate the optimal RP that provides consistent and reliable indicators of guessing in a student population. This study as submitted provides a conceptual outcome that shows definite changes in outcomes with an arbitrary difference in learning age of approximately one-and-a-half years between the anticipated response for an item of a given difficulty and the observed ability of the student.