Abstract

Title of dissertation: ITEM-ANALYSIS METHODS AND THEIR IMPLICATIONS FOR THE ILTA GUIDELINES FOR PRACTICE: A COMPARISON OF THE EFFECTS OF CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY MODELS ON THE OUTCOME OF A HIGH-STAKES ENTRANCE EXAM

David P. Ellis, Doctor of Philosophy, 2011

Directed by: Professor Steven Ross
Department of Second Language Acquisition and
Center for the Advanced Study of Language

The current version of the International Language Testing Association (ILTA) Guidelines for Practice requires language testers to pretest items before including them on an exam, or when pretesting is not possible, to conduct post-hoc item analysis to ensure any malfunctioning items are excluded from scoring. However, the guidelines are devoid of guidance with respect to which item-analysis method is appropriate for any given examination. The purpose of this study is to determine what influence choice of item-analysis method has on the outcome of a high-stakes university entrance exam. Two types of classical-test-theory (CTT) item analysis and three item-response-theory (IRT) models were applied to responses generated from a single administration of a 70-item dichotomously scored multiple-choice test of English proficiency, administered to 2,320 examinees applying to a prestigious private university in western Japan. Results illustrate that choice of item-analysis method greatly influences the ordinal ranking of examinees. The implications of these findings are discussed and recommendations are made for revising the ILTA Guidelines for Practice to delineate more explicitly how language testers should apply item analysis in their testing practice.

ITEM-ANALYSIS METHODS AND THEIR IMPLICATIONS FOR THE ILTA GUIDELINES FOR PRACTICE: A COMPARISON OF THE EFFECTS OF CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY MODELS ON THE OUTCOME OF A HIGH-STAKES ENTRANCE EXAM

by

David P. Ellis

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor Steven Ross, Committee Chairperson
Professor Robert DeKeyser
Professor Roberta Lavine, Dean's Representative
Professor Robert Lissitz
Professor Michael Long

Acknowledgements

I would like to thank my entire committee for all of their advice and encouragement throughout the completion of my dissertation. Without each member's support, I would not be in the position to graduate that I am today. A special thank you goes to my committee chair, Dr. Steven Ross, for his help in identifying a topic and teaching me how to run the required analyses, and to committee member, Dr. Robert Lissitz, for his help in confirming which analyses to run and how to interpret the subsequent findings. I would also like to thank Drs. Michael Long and Robert DeKeyser, the persons to whom most of the credit goes for my initial application to the PhD program at the University of Maryland, and to whom I relied on most in finding my way through the program. Finally, I would like to thank my wife, the newly minted Dr. Sunyoung Lee-Ellis, whose steadfast persistence throughout her own course of study motivated me to continue marching forward through my own.

Table of Contents

Acknowledgements ii

List of Tables vi

List of Figures viii

1 Introduction 1
- 1.1 Context of the Study 1
- 1.2 Purpose of the Study 3
- 1.3 Research Questions 5
- 1.4 Significance of the Study 6

2 Review of the Literature 7
- 2.1 Item-Analysis Methods 7
  - 2.1.1 Raw-Score Method 8
  - 2.1.2 Classical Test Theory (CTT) 9
  - 2.1.3 Item Response Theory (IRT) 15
    - 2.1.3.1 Three-Parameter Logistic (3PL) Model 16
    - 2.1.3.2 Two-Parameter Logistic (2PL) Model 18
    - 2.1.3.3 One-Parameter Logistic (1PL) Model 18
    - 2.1.3.4 IRT Assumptions 21
- 2.2 Assessing Model Fit 43
  - 2.2.1 Rationale for Model Choice 43
  - 2.2.2 Estimation of Item Parameters 49
  - 2.2.3 Estimation of Classification Accuracy (CA) 53
    - 2.2.3.1 CA Studies outside Language Testing 54
    - 2.2.3.2 CA Studies within Language Testing 58
  - 2.2.4 Other IRT Applications within L2 Testing 59

3 Data Analysis I (Full Data Set) 67
- 3.1 Participants 67
- 3.2 Methods 67
  - 3.2.1 Japanese University Admissions Process 67
  - 3.2.2 Form H of the Entrance Exam 68
  - 3.2.3 Calculation of Examinee Scores 70
  - 3.2.4 Determination of the Cut Score 71
- 3.3 Problems with the Raw-Score Method 73
- 3.4 Alternatives to the Raw-Score Method 78
  - 3.4.1 CTT (Kelley's Discrimination Index) 78
  - 3.4.2 CTT (Point-Biserial Correlation) 80
  - 3.4.3 Rasch/1PL 81
  - 3.4.4 IRT 2PL 86
  - 3.4.5 IRT 3PL 91
- 3.6 Test of IRT Assumptions 99
  - 3.6.1 Test of Unidimensionality 99
  - 3.6.2 Test of Local Item Independence 102

4 Data Analysis II (Random 80%) 103
4.1 Participants...103
4.2 Descriptive Statistics...103
4.3 Cut-Score Determination 105
4.4 Item Analyses..106
4.4.1 CTT (Kelley)..106
4.4.2 CTT (Point Biserial)...107
4.4.3 Rasch/1PL 108
4.4.4 IRT 2PL...109
4.4.5 IRT 3PL...115
4.5 Summary of Analyses 120
4.6 Test of IRT Assumptions 123
4.6.1 Test of Unidimensionality...124
4.6.2 Test of Local Item Independence...126
5 Data Analysis III (Middle 80%).. 127
5.1 Participants...127
5.2 Descriptive Statistics...127
5.3 Cut-Score Determination 129
5.4 Item Analyses..130
5.4.1 CTT (Kelley)..130
5.4.2 CTT (Point Biserial)...131
5.4.3 Rasch/1PL 132
5.4.4 IRT 2PL...133
5.4.5 IRT 3PL...139
5.5 Summary of Analyses 143
5.6 Test of IRT Assumptions 147
5.6.1 Test of Unidimensionality...147
5.6.1 Test of Local Item Independence...150
6 Discussion 151
6.1 Research Question 1..151
6.2 Research Question 2..152
6.3 Research Question 3..155
6.3.1 Raw-Score vs. CTT (Kelley) 157
6.3.2 CTT (Kelley) vs. CTT (Point-Biserial)..158
6.3.3 CTT (Point-Biserial) vs. Rasch/1PL 159
6.3.4 CTT (Point-Biserial) vs. IRT 2PL..160
6.3.5 CTT (Point-Biserial) vs. IRT 3PL..160
6.4 Research Question 4..161
6.5 Research Question 5..162
6.6 Research Question 6..165
7 Conclusion.. 168
Appendix A: Entrance Exam.. 176
Appendix B: Item-Analysis Tables (Full Data Set) 184

Appendix C: Item-Analysis Tables (Random 80%).................................. 194
Appendix D: Item-Analysis Tables (Middle 80%) 202
Appendix E: Xcalibre 4.0 Input & Output Specifications.......................... 210
References 212

List of Tables

Table 1. Example of item statistics using CTT 12

Table 2. Differences between the Rasch-1PL model and the IRT-1PL model....20

Table 3. Excerpt of table from Brown & Bailey (2008), illustrating topics taught in language-testing courses 61

Table 4. Descriptive statistics of the raw score data 70

Table 5. Excerpt of item analysis based on Kelly's D 79

Table 6. Excerpt of item analysis based on point-biserial correlations 80

Table 7. Excerpt of item analysis for the Rasch/1PL model....81

Table 8. Excerpt of item analysis for the 2PL model 87

Table 9. Excerpt of item analysis for the 3PL model 92

Table 10. Summary of flagged items across models 97

Table 11. Classification consistency across models 98

Table 12. Kappa statistics of each pairwise model comparison 99

Table 13. Chi-square comparison of CFA models....101

Table 14. Estimated gamma variances....102

Table 15. Descriptive statistics of the original data and first resampling....103

Table 16. Excerpt of item analysis based on Kelley's D 107

Table 17. Excerpt of item analysis based on point-biserial correlations 108

Table 18. Excerpt of item analysis for the Rasch/1PL model....109

Table 19. Excerpt of item analysis for the 2PL model 110

Table 20. Excerpt of item analysis for the 3PL model 116

Table 21. Summary of the items flagged across models 121

Table 22. Classification consistency across models 122

Table 23. Kappa statistics of each pairwise model comparison 123

Table 24. Chi-square comparison of CFA models....125

Table 25. Estimated gamma variances....126

Table 26. Descriptive statistics of the original data and second resampling 127

Table 27. Excerpt of item analysis based on Kelley's D 130

Table 28. Excerpt of item analysis based on point-biserial correlations 132

Table 29. Excerpt of item analysis for the Rasch/1PL model.......................................133

Table 30. Excerpt of item analysis for the 2PL model 134

Table 31. Excerpt of item analysis for the 3PL model 139

Table 32. Summary of the items flagged across models 143

Table 33. Classification consistency across models 146

Table 34. Kappa statistics of each pairwise model comparison 147

Table 35. Chi-square comparison of CFA models...149

Table 36. Estimated gamma variances..150

Table 37. Cross-tabulation results between/among methods...153

Table 38. Levels of agreement between methods 154

Table 39. Examples of examinee misclassification 155

Table 40. Example of examinee classification variance 156

Table 41. Excerpt of item analysis for TRT 163

Table 42. Classification consistency across models 164

Table 43. Classification consistency between 3PL and TRT 164

Table 44. Displacement as a function of the total number admitted 173

List of Figures

Figure 1. Information for a pair of Math-Content LID items when they are scaled separately or as a testlet (from Yen, 1993, p. 204) 36

Figure 2. Distribution of examinee raw scores 71

Figure 3. Cut score (45) along the distribution of examinee raw scores 73

Figure 4. Item that is functioning well...75

Figure 5. Item not differentiating candidates by ability...76

Figure 6. Item with two apparent correct responses (a double-keyed item)....................77

Figure 7. Test Information Function of the Rasch model...83

Figure 8. Test Information Function of CTT analyses 84

Figure 9. CSEM of the Rasch model 85

Figure 10. Test Response Function of the Rasch model..86

Figure 11. Test Information Function of the 2PL model...88

Figure 12. CSEM of the 2PL model 89

Figure 13. Test Response Function of the 2PL model...89

Figure 14. Scatterplot of difficulty (b) by discrimination (a) 90

Figure 15. Histogram of the b-parameters of the 2PL model 91

Figure 16. Test Information Function of the 3PL model...93

Figure 17. CSEM of the 3PL model 94

Figure 18. Scatterplot of difficulty (b) by discrimination (a) 95

Figure 19. Histogram of the b-parameters of the 3PL model 96

Figure 20. 1-factor SEM of the full data set 100

Figure 21. 2-factor SEM of the full data set 100

Figure 22. Distribution of examinee scores for the first resampling 104

Figure 23. Cut score (44) along the distribution..106

Figure 24. Test Information Function of the 2PL model...111

Figure 25. CSEM of the 2PL model 112

Figure 26. Test Response Function of the 2PL model..113
Figure 27. Scatterplot of difficulty (b) by discrimination (a) 114
Figure 28. Histogram of the b-parameters of the 2PL model 115
Figure 29. Test Information Function of the 3PL model..117
Figure 30. CSEM of the 3PL model 118
Figure 31. Test Response Function of the 3PL model..118
Figure 32. Scatterplot of difficulty (b) by discrimination (a) 119
Figure 33. Histogram of the b-parameters of the 3PL model 120
Figure 34. 1-factor SEM of the first resampling..124
Figure 35. 2-factor SEM of the first resampling..125
Figure 36. Distribution of examinee scores for the second resampling.........................128
Figure 37. Cut score (42) along the distribution 129
Figure 38. Test Information Function of the 2PL model..135
Figure 39. CSEM of the 2PL model 135
Figure 40. Test Response Function for the 2PL model 136
Figure 41. Scatterplot of difficulty (b) by discrimination (a) 137
Figure 42. Histogram of the b-parameters of the 2PL model 138
Figure 43. Test Information Function of the 3PL model..140
Figure 44. CSEM of the 3PL model 141
Figure 45. Test Response Function of the 3PL model..141
Figure 46. Scatterplot of difficulty (b) by discrimination (a) 142
Figure 47. Histogram of the b-parameters of the 3PL model 143
Figure 48. 1-factor SEM of the restricted-range data 148
Figure 49. 2-factor SEM of the restricted-range data 148
Figure 50. Proposed model-choice decision tree 170

1 Introduction

1.1 Context of the Study

Whether due to capitalist ideals, the accuracy of Malthusian predictions, or some other explanation, competitiveness at all levels of society is on the rise across the globe. Simultaneously, or perhaps as a result, the assessment of knowledge, skills, and abilities continues to increase, as does the influence of test-makers and administrators. Because many of these assessments are the primary (or even sole) factor in determining an examinee's fate, the stakes of these exams have never been greater, nor has the importance of the creation, administration, and scoring of valid tests.

One example of this type of all-or-nothing exam is the university entrance exam. As the increasing influence of the Scholastic Aptitude Test (SAT) on U.S. university admissions continues to be lamented in some circles (e.g., Schaeffer, 2010), the SAT is still but one of several factors influencing U.S. university admissions decisions. The same is not true in many other countries around the globe (e.g., Japan), where the outcome of the university entrance exam all but seals examinees' academic fate, and in many cases, the trajectory of their subsequent careers. Quite literally, mere acceptance to one of the country's premier universities can ensure the career success of its would-be students; conversely, admissions failure can have lifelong consequences (Cutts, 1997).

Of course, only a country's most gifted students have any chance of acceptance to an elite university. Nevertheless, in academically oriented societies like Japan and Korea, higher education is the goal of virtually all students, so the decision of whether one will be admitted to a top-tier university, second-tier university, or even *any* university rests solely on the outcome of their admissions exam. For this reason, the validity of these tests is of critical importance to both examinees and their families.

The validity of any test is a function of numerous factors, whether one chooses to view validity theoretically as a unitary construct (e.g., Messick, 1989) or as a multidimensional construct (e.g., Borsboom, Mellenbergh, & van Heerden, 2004). Messick (1980) grouped numerous aspects of validity into three categories – construct validity, content validity, and criterion validity – where *construct validity* refers to the extent a test measures the construct it was designed to measure, *content validity* refers to the extent a test's content covers a representative sample of the domain of interest, and *criterion validity* refers to the extent a test correlates with a variable representative of the construct (e.g., how well a written driving test correlates with a hands-on driving test, if the latter has already been shown to be a valid indicator of actual driving ability).[1] It was in an even earlier paper (1965), however, that Messick addressed the importance of appropriate test use as a component of construct validity. Sometimes termed *consequential validity*, Messick stressed the fact that the psychometric adequacy of a test is a necessary but insufficient condition of overall test validity (c.f., Lissitz & Samuelsen, 2007, p. 445). More specifically, he emphasized that it is imperative test-makers and administrators give due consideration to a test's social consequences, both short-term and long-term (1980, p. 1012). In this light, the consequential validity of university entrance exam outcomes assumes a very important role in the test's overall validity given the short- and long-term consequences of these exams, particularly in countries like Japan where it is the sole determinant of university admission.

[1] See also the *Standards for Educational and Psychological Testing* (1999) for a discussion of test validity.

1.2 Purpose of the Study

Whether one considers consequential validity to be an integral component of test validity in a theoretical sense does not diminish its practical importance with respect to high-stakes tests. Evidence of this is reflected in many testing standards/guidelines, including those of at least four language-testing associations:

- International Language Testing Association (ILTA) – Code of Ethics & Guidelines for Practice
- European Association for Language Testing and Assessment (EALTA) – Guidelines for Good Practice in Language Testing and Assessment
- Association of Language Testers in Europe (ALTE) – Principles of Good Practice
- Japanese Language Testing Association (JLTA) – Code of Good Testing Practice

In all of these guidelines, the ethical calculation of scores is attributed paramount importance. For example, Section B.4 of the ILTA Guidelines for Practice states:

> The work of the task and item writers needs to be edited before pretesting. If pretesting is not possible, the tasks and items should be analysed after the test has been administered but before the results are reported. Malfunctioning or misfitting tasks and items should not be included in the calculation of individual test takers' reported scores. (p. 2)

Clearly, language-test designers and administrators have an obligation to examine the quality of their tests before reporting scores, including removing any misfitting items before calculating scores. This requirement is certainly relevant to the testing context of most East Asian countries (Japan, Korea, and China), where the culture often prohibits the pretesting of items due to the fact that exams are typically single use and created in secret, following long-established and well-regarded traditions (Ross, 2011). Unfortunately, neither the ILTA Guidelines (nor any other set of standards) specifies what type of item analysis should be conducted even if administrators acknowledge the need to do so. Stated differently, because there are numerous item-analysis methods in existence, it remains unclear how to determine which method will yield the most valid set of scores for a particular test.

Like U.S. universities, Japanese universities have a target number of applicants they strive to admit each year. More specifically, the Japanese Ministry of Education sets an annual quota for how many students can be admitted to each college or university. As a result, the cut score for admission is sample dependent and norm referenced, meaning it is not tied to a particular standard or mastery level but simply is a function of the annual quota. For this reason, examinees are concerned only with their performance relative to other applicants rather than with demonstrating a particular level of proficiency. In this respect, the admissions fate of examinees is far less within their control than it would be were the entrance exam criterion referenced, so test scorers have an even greater responsibility for ensuring the classification accuracy of examinee scores, particularly around the cut line. The question is whether the item-analysis method chosen by scorers, assuming one is chosen at all, will yield the highest classification accuracy possible. The purpose of this study is to provide some insight into this issue by investigating how different item-analysis methods influence the classification accuracy of examinee scores on a single-use university entrance exam to provide some insight into this issue.

1.3 Research Questions

To fulfill the purpose of this study, the following research questions will be addressed with respect to the examination under investigation:

1. Do the item-analysis methods prescribed by Classical Test Theory (CTT) and Item Response Theory (IRT) identify any test items as faulty/misfitting, and if so, do they identify the same items?
2. Does the identification of misfitting items influence the classification consistency of examinee scores across methods?
3. Which item-analysis method is likely to lead to the highest level of classification accuracy given the nature of the test data?

In addition to the above questions, the following three questions about the assumptions of IRT modeling will be addressed:

4. Can a test designed to measure the construct *English proficiency* satisfy the assumption of unidimensionality?
5. Can a test that contains *testlets* satisfy the assumption of local item independence?
6. If both assumptions can be satisfied, does classification accuracy improve sufficiently to justify the added computational complexity of IRT modeling?

1.4 Significance of the Study

Although the impact of item-analysis method on test scores has been examined in the educational assessment literature on several occasions (e.g., Anderson, 1999; Silva, 1985; Stone, Weissman, & Lane, 2005), it has been addressed in the second language (L2) literature only once (B. Zhang, 2010). As a result, this study will contribute to the L2 literature in a unique and important way. For only the second time, multiple item-analysis methods will be applied to a large data set to determine what influence, if any, different item-analysis methods have on the classification consistency and accuracy of examinee test scores.

Perhaps more importantly, the methods of data analysis described herein will contribute to the field's relatively nascent understanding and application of item response theory (IRT) to second language assessments, particularly with respect to the satisfaction of IRT's two key assumptions – unidimensionality and local item independence. The strengths and weaknesses of each model/method will be discussed in detail, as well as the criteria administrators can use to make an informed decision about which item-analysis method to employ in their particular testing context.

Finally and most importantly, the necessity of item analysis *per se*, especially for high-stakes tests, will be discussed with respect to the standards and guidelines espoused by several language testing associations, most notably the International Language Testing Association (ILTA).

2 Review of the Literature

2.1 Item-Analysis Methods

In the business world, managers routinely track the performance of their subordinates against explicit quantifiable goals to determine whether they are performing successfully or unsuccessfully. This tracking of performance is applicable to most professions, including the field of educational assessment, where test validation can help stakeholders better understand how well their test is capturing the relationship between examinees' observed scores and the construct(s) being measured.

Within educational assessment, there are numerous means of evaluating the validity of a test through the use of item analysis, with these means commonly divided by test type – whether the test contains items that are scored dichotomously (e.g., multiple-choice items with answers keyed as either correct or incorrect) or polytomously (e.g., constructed-response items keyed on a Likert scale or multiple-choice items that permit partial credit scoring depending on which distractor is chosen). On many large-scale tests, dichotomously scored multiple-choice items are preferred because machine scoring of these items is possible, thereby reducing scoring time and cost. This is certainly true of many university entrance exams, where resources for scoring tests are often in short supply. Following is a discussion of the most common scoring method, contrasted with item-analysis methods derived from the two test-model theories most often applied to dichotomously scored multiple-choice tests.

2.1.1 Raw-Score Method

The simplest means of estimating examinee performance on a test containing only dichotomously scored items is the raw-score method, where the answer to each item is either correct or incorrect and the score of every item is equal in weight, such that the total score is simply the sum of all items answered correctly. For interpretive purposes, this number-correct score is often reported as a percentage-correct score, equal to the number of items answered correctly divided by the total number of items on the exam. To this day, the letter-grade system so often employed in U.S. schools has its basis in number-correct/percentage-correct scores, where scores of 90-100% are typically awarded an A, 80-89% a B, 70-79% a C, and so forth.

Because of its simplicity in calculation and familiarity in reporting, the number-correct scoring method is putatively the most pervasive scoring method. Unfortunately, this simplicity/familiarity is not without consequence because the number-correct method requires several assumptions to be true about the test items or the validity of the resulting scores greatly diminishes. For example, this method assumes each item on the test is a fair measure of examinee ability. More specifically, it assumes each item contains only one unambiguously correct answer, all other answer options (*distracters*) appear equally plausible to all examinees, and examinees of higher ability will get each item correct more often than will examinees of lower ability.

A second assumption of this scoring method is that each item contributes equally to the overall estimate of examinee ability regardless of its relative difficulty. In other words, because each correct response is assigned an equal value (usually 1), the assumption is that every item contributes equally to estimating an examinee's ability despite the fact certain items on the test will be relatively more difficult than others.

A third assumption of the raw-score method is that the distribution of answers across the full range of examinee ability is equal across all items. Put another way, it is assumed that the distance (*discrimination*) between, e.g., examinees of low ability and high ability will remain constant across all items.

A fourth and final assumption of the raw-score method is that guessing does not influence the overall outcome of an examinee's score in a meaningful way. Stated differently, the likelihood examinees will guess correctly the answer to any one item is equal across all items and therefore negligible in its effect on overall scores.

2.1.2 Classical Test Theory (CTT)

Because the assumptions inherent in the raw-score method are considered unreasonable for many exams, test theory has been evolving for more than a century to mitigate the need to subscribe to such assumptions. For example, in the first half of the 20th century, psychometricians began evaluating the reliability of test scores, meaning they began to evaluate the internal consistency of tests. Various coefficients were created to measure a test's internal consistency, either for single administrations of a test (e.g., Cronbach's α) or multiple administrations (e.g., Pearson's r). In principle, the higher the correlation between test scores (or test-half scores[2]), the more likely the test is measuring a construct consistently/reliably. In other words, the higher the correlation, the lower the amount of measurement error present in the observed scores.

[2] When there is only one administration of an exam, the test is sometimes split in half to create two "alternate" forms of a test so its internal consistency can be estimated.

Such correlation coefficients led to the genesis of a measurement theory eventually coined *Classical Test Theory* (Novick, 1966). Based on the premise that observed scores are a function of only two factors – true scores and measurement error – the theoretical basis for CTT resides in the following formula:

$$X = T + E \qquad (1)$$

where the observed score (X) is the score an examinee achieves on a test, the true score (T) is the score that represents the theoretical (but unobservable) ability of that examinee, and measurement error (E) is the byproduct of an imperfect measure of that examinee's true ability.

In vogue for much of the 20th century, CTT led psychometricians to develop numerous statistics designed to evaluate test/item characteristics, including correlation coefficients that examine the reliability of tests containing two continuous variables (Cronbach's α, Pearson's r), two dichotomous variables (Phi, ϕ), or one of each (Point Biserial, r_{pb}). Other components of CTT analysis included estimates of an item's difficulty and discrimination. For example, on a dichotomously scored multiple-choice test, the *difficulty* of an item (p) is equal to the number of examinees who answer the item correctly divided by the total number of examinees. In other words, its difficulty is reflected in the proportion of examinees who answer the item correctly.[3] The lower the proportion of correct responses, the more difficult the item is presumed to be.

In contrast, an item's *discrimination* (D) is equal to the difference in proportion-correct scores between examinee groups of differing abilities. Kelley (1939) was one of

[3] In this sense, the difficulty of an item is inversely related to its p-value, meaning easier items have higher p-values because more examinees are successful in answering the item correctly. For this reason, some researchers prefer the term "facility" to "difficulty" because of the positive correlation between p-values and relative ease of the item.

the first psychometricians to offer a means of calculating an item's discrimination, suggesting use of the top 27% and bottom 27% of examinees (based on overall test score) as the basis for how well an item differentiates examinees of differing ability. An item that discriminates maximally would have D = 1.0, where all of the examinees in the upper-ability group answer the item correctly and all of the examinees in the lower-ability group answer the item incorrectly, as shown below in Equation 2:

$$D = p_{upper} - p_{lower} \qquad (2)$$

Within the CTT framework, the ideal item has a mean difficulty $p = .50$ and a discrimination D = 1.0. Mathematically speaking, discrimination is a function of an item's difficulty across the pool of examinees, meaning the two statistics are interrelated, with discrimination limited by difficulty. Consider an item that is infinitely easy or impossibly difficult. In the former case, $p = 1.0$, where all examinees, regardless of overall ability, are able to answer the item correctly. In the latter case, $p = 0$, where no examinees, regardless of ability, are able to answer the item correctly. In both cases, the discriminating power of the item is D = 0; all examinees get the item correct or incorrect, regardless of ability, so there is no discrimination among examinees across ability levels. On the other hand, items with a difficulty $p = 0.50$ create the most opportunity for discrimination among examinees. For example, if 100 individuals record a response to an item with a difficulty $p = 0.50$, 50 of the examinees will have answered the item correctly and 50 will have answered it incorrectly. This allows up to 2500 (50 x 50) potential differentiations among the 100 examinees. On the other hand, if an item's difficulty is, e.g., $p = .25$, it would indicate 25 examinees answered the item correctly and 75 incorrectly, for a maximum of 1875 possible differentiations (25 x 75).

While item discrimination is a function of item difficulty, it is important to note that the possible number of differentiations among examinees is not always equal to the actual discrimination of an item because the value of D is dependent on how the upper- and lower-ability groups are formed. For example, Kelley's (1939) suggestion of 27% as the basis for extreme group formation is somewhat arbitrary,[4] so the discriminating power of any given item will vary depending on the percentage chosen for group formation. In most cases, the percentage should be between 25% and 33%, depending on several factors, including the number of examinees who take the exam and the distribution of total scores across the range of abilities. Whatever the case, maximum discrimination can be realized on items with difficulty at $p = 0.50$ for the reasons explained above. Table 1 below illustrates an example of what item difficulty and discrimination might look like for four items of a test:

Table 1. Example of item statistics using CTT

Item #	Difficulty (*p*) All Examinees	Difficulty (*p*) Upper 27%	Difficulty (*p*) Lower 27%	Discrimination (D)
1	0.50	0.90	0.10	0.80
2	0.95	1.00	0.90	0.10
3	0.65	0.80	0.30	0.50
4	0.50	0.30	0.70	-0.40

As shown, Item 1 appears to be functioning nearly ideally. Its difficulty is at the target 0.50, thereby maximizing its discriminatory potential, and its actual discrimination is quite high (D = 0.90 – 0.10 = 0.80). In contrast, Item 2 is functioning very poorly because virtually every examinee was able to answer the item correctly ($p = 0.95$), resulting in a very low discrimination between upper- and lower-ability groups (D = 0.10). In other

[4] Kelley actually based the choice of 27% on where the tails of a normal distribution form, which has rationale but is still somewhat arbitrary from a mathematical standpoint.

words, the item contributes little to discerning examinee ability. Item 3 is an example of a solid item that is not necessarily ideal because it is relatively easy (p = 0.65) but is nevertheless serving the test well because it is discriminating well between extreme group ability levels (D = 0.50). A test containing items with similar characteristics would be considered a well-designed test, at least with respect to these aspects of reliability/validity.

Item 4 is interesting because of it has the ideal p-value (0.50), but it has a negative D-value. This negative value (D = -0.40) indicates the lower-ability group actually outperformed the upper-group on this item, suggesting a problem with the item. While there are many possible reasons an item would exhibit a negative D-value, a mistake in the answer key or a distractor that misleads examinees of higher-ability but does not have the same effect on examinees of lower ability are two possible explanations. Whatever the reason, items with negative D-values should be examined closely because they misfit the model and lower the reliability of the test. If the answer key is not mistaken, a decision will need to be made about whether to keep the item in the scoring analysis.

When item analysis using difficulty and discrimination as the parameters is performed on all of the items of a test, items that contribute little (or worse, negatively) to the discriminatory power of a test overall can be removed, both for reliability analysis and subsequent test administrations. In most cases, removal of the misfitting items will improve the reliability of the test provided enough items remain.[5]

While CTT is a considerable improvement over the raw-score method due to its focus on the internal consistency of a test, it is not without its own limitations. First, CTT item statistics are population dependent, meaning they are applicable only to the group of examinees who took the test at that particular time. Second, CTT person-ability estimates

[5] Generally speaking, the reliability of a test correlates with the number of items on the test.

are item dependent, meaning they are due solely to the particular group of items the examinees answered on that particular test administration. This item-person co-dependence not only creates a circular reference, it also precludes the generalizability of CTT statistics to other examinee populations and test administrations. In other words, aside from the theoretical problem of circularity, there is a very practical problem in that any change to the composition of a test or to the examinee population necessitates a reanalysis of the item/test characteristics.

Another limitation of CTT is the assumption of *item equivalence*, meaning total scores are the sum of equally-weighted item scores. Recall that in the case of the raw-score method, all items are scored as either correct (value = 1) or incorrect (value = 0), so each item carries equal weight in contributing to the final score. While CTT item analysis can help identify and eliminate misfitting items from the analysis, thereby improving the overall reliability of the test score, it still assumes item equivalence in that all dichotomously scored items receive a value of 1 or 0. In other words, it assumes the test scale is an interval scale, which may not necessarily be the case.

A third limitation of CTT is its treatment of measurement error. In CTT, it is assumed that the standard error of measurement (SEM) is constant across all ability levels. However, this assumption is often unreasonable. Recall that the ideal item difficulty value is $p = 0.50$. At this level of difficulty, item discrimination can be maximized, so the majority of test items on a well-designed test will typically have *p*-values at/around 0.50. The consequence of this design is that there are relatively few items that are extremely difficult (e.g., p = 0.10) or extremely easy (e.g., p = 0.90), so examinees whose abilities are at the extremes (relative to those in their exam cohort) confront very few items at their level of ability. This is problematic because reliability is positively correlated with the number of items, so fewer items equates to lower reliability, which in turn equates to greater measurement error in examinee scores that are in the tails of the distribution. In short, SEM in most cases will not be constant across ability levels, so the assumption of a fixed SEM is problematic.

2.1.3 Item Response Theory (IRT)

In response to the limitations of CTT item analysis, psychometricians began looking for means to identify test/item characteristics that could be generalized across both test administrations and examinee populations. These identification and calculation of these characteristics, collectively known as *Item Response Theory* (IRT), fulfill this need and have been in development since the 1950s, with several seminal publications emerging in the 1960s (e.g., Birnbaum, 1968; Lord & Novick, 1968; Rasch, 1960). It was not until the 1980s, however, that IRT realized its full potential, when computers became powerful enough to execute the complex calculations required (e.g., BILOG: Mislevy & Bock, 1982). To this day, IRT models are the preferred choice for large-scale, high-stakes test administrations because of their strong theoretical underpinnings and their practical benefits, including, e.g., sample-free item calibration, item-free person measurement, misfitting item and person identification, and test equating and linking (Henning, 1987).

The prominence of IRT is evident throughout psychometric research, being the topic of several book chapters (e.g., W. Yen & Fitzpatrick, 2006) and monographs (e.g., Baker, 2001; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985), as well as publications on the scholarly periphery (e.g., Chong, 2011; Partchev, 2004). It has been applied to numerous contexts as well, including subscale scores (e.g., Kolen, Zeng, & Hanson, 1992; Skorupski & Carvajal, 2010), differential item functioning (e.g., Wyse & Mapuranga, 2009; Zenisky, Hambleton, & Robin, 2004) and growth/change modeling (Reise & Haviland, 2005).

In contrast to CTT, which focuses primarily on test-level concerns like reliability, IRT focuses primarily on the factors that influence the observed scores of each individual item (i.e., *item-pattern scoring* vs. *number-correct scoring*). Common to all three IRT models is a set of three parameters, with the models varying only in the assumptions they make about each of the parameters. Below is a brief description of each model, in reverse order of model complexity for explanatory clarity.

2.1.3.1 Three-Parameter Logistic (3PL) Model

As mentioned, all three IRT models have three parameters:

1. Item discrimination (a)
2. Item difficulty (b)
3. Pseudo-guessing[6] (c)

[6] Following Yen & Fitzpatrick (2006), the term pseudo-guessing is used here to differentiate it from random guessing, in that low ability learners likely will at least try to make an educated guess at the correct answer rather than choosing an answer at random. In the latter case, c_i = .25 on a 4-choice multiple-choice item, whereas c_i could be greater than or less than .25 depending on the nature of the distracters and assuming examinees are trying to answer the item to the best of their ability.

The IRT-3PL model is expressed mathematically in the following formula, known as the *item response function*:

$$P_i(\theta) = c_i + \frac{1-c_i}{1+\exp[-Da_i(\theta-b_i)]} \qquad (3)$$

In Equation (3), the three parameters of the model are represented by a_i, the discrimination power of item i, b_i, the difficulty of item i, and c_i, the likelihood of guessing item i correctly given no ability to answer the item correctly. The function $1/[1+\exp(-t)]$ is a logistic function, with exp(–t) denoting e, the natural exponent. Within this logistic function, D is a multiplicative constant, typically set to 1.7 or 1.702, because this value helps the 2PL model approximate the normal ogive model (Yen & Fitzpatrick, 2006, p. 114).

Unlike the CTT model, where person abilities are dependent on the particular item set, IRT permits the calculation of the probability a person at a given level of ability will be able to answer a locally-independent item correctly, represented in Equation (3) by the term $P_i(\theta)$. In sum then, the IRT-3PL model states that the probability (P) a person of a given ability (θ) will answer item i correctly is a (logistic) function of the item's discrimination power, a_i, its difficulty, b_i, and the likelihood the correct answer can be guessed in the absence of any ability, c_i.

2.1.3.2 Two-Parameter Logistic (2PL) Model

The IRT-2PL model differs from the 3PL model only in that it assumes pseudo-guessing is not a meaningful contributor to item fit, or more typically, that it is not applicable to the data (e.g., in the case of rater-scored constructed-response items). Hence, its item response function is very similar in form:

$$P_i(\theta) = \frac{1}{1+\exp[-Da_i(\theta-b_i)]} \qquad (4)$$

Note the only difference is the absence of the guessing parameter, c_i. Implicit in this model then is the assumption that the probability an examinee of very low ability will answer item *i* correctly approaches 0 ($c_i = 0$). Stated differently, the lower asymptote intercepts the Y-axis (person ability) near 0 in the 2PL model.

2.1.3.3 One-Parameter Logistic (1PL) Model

In the IRT-1PL model, the only parameter that is estimated is the difficulty parameter, b_i, as shown below in Equation (5):

$$P_i(\theta) = \frac{1}{1+\exp(\theta-b_i)} \qquad (5)$$

Implicit in this model is the assumption that item discrimination is constant across item difficulty and person ability. As with CTT, this assumption is unreasonable in many circumstances. Nevertheless, the IRT-1PL model is employed in many circumstances because a large amount of item information can be captured by this one parameter alone (Yen & Fitzpatrick, 2006, p. 114).

2.1.3.3.1 IRT-1PL model vs. Rasch-1PL model

Even though IRT has been in existence for over 50 years, there is a common misconception to this day that the IRT-1PL model is synonymous with the Rasch-1PL model (e.g., Taylor & Lee, 2010). While it is true they employ the same item response function to examine fit (including only the difficulty parameter, b_i), they have diametrically opposed philosophies of fit evaluation: whereas the IRT-1PL model (like all IRT models) assumes all items are sound, the Rasch-1PL model assumes the model is sound such that any item misfit is a function of the item, not the model. In other words, employment of the Rasch-1PL model excludes items (and persons) from the analysis to improve the fit of the model, whereas the IRT-1PL model does not – the model is either accepted or rejected depending on its initial global fit. Table 2 is a summary description of the differences between the two 1PL models (adapted from Linacre, 2005):

Table 2. Differences between the Rasch-1PL model and the IRT-1PL model

Aspect	Rasch Dichotomous Model	Item Response Theory
Item Response Function	1PL	1PL
Context	When each individual in the person sample is parameterized for item estimation.	When the person sample is parameterized by a mean and standard deviation for item estimation.
Motivation	Prescriptive: Distribution-free ability estimates and distribution-free item-difficulty estimates on an additive latent variable	Descriptive: Computationally simpler approximation to the Normal Ogive Model of L.L. Thurstone, D.N. Lawley, F.M. Lord
Persons, objects, subjects, cases, etc.	Person n of ability Bn, or Person ν (Greek nu) of ability βn in logits	Normally-distributed person sample of ability distribution θ, conceptualized as N(0,1), in probits: incidental parameters
Items, prompts, etc.: structural parameters	Item i of difficulty Di, or Item ι (Greek iota) of difficulty δi in logits	Item i of difficulty bi (the "one parameter") in probits
Nature of binary data	1 = success (presence of property) 0 = failure (absence of property)	1 = success (presence of property) 0 = failure (absence of property)
Probability of binary data	Pni = probability that person n is observed to have the requisite property, "succeeds", when encountering item i	Pi(θ) = overall probability of "success" by person distribution θ on item i
Local origin of scale: zero of parameter estimates	Average item difficulty, or difficulty of specified item (criterion-referenced)	Average person ability (norm-referenced)
Item discrimination	Item characteristic curves (ICCs) modeled to be parallel with a slope of 1 (the natural logistic ogive)	ICCs modeled to be parallel with a slope of 1.7 (approximating the slope of the cumulative normal ogive)
Fit evaluation	Fit of the data to the model; Local, one parameter at a time	Fit of the model to the data; Global, accept or reject the model
Data-model mismatch	Defective data do not support parameter separability in an additive framework. Consider editing the data.	Defective model does not adequately describe the data. Consider adding discrimination (2-PL), lower asymptote (guessability, 3-PL) parameters.

In short, the Rasch-1PL model is confirmatory and the IRT-1PL model is exploratory. In other words, with Rasch, the data must fit the model, while with IRT, the model must fit the data. This distinction is important when classification accuracy as a function of item fit is being examined because model fit will depend on the philosophy adhered to – either initial fit (in the case of the IRT-1PL model) or upon subsequent fit after misfitting items and persons are removed from the data (in the case of the Rasch-1PL model). In this study, the Rasch-1PL model will be employed because the IRT-1PL model is not theoretically justifiable given the nature of the test items, as explained later.

2.1.3.4 IRT Assumptions

Like classical test theory, IRT relies on a set of assumptions. First is the assumption of *unidimensionality*. Central to all IRT models, unidimensionality is the assumption that all of the items on a test (or test section) measure only one latent trait/ability (e.g., reading comprehension). This assumption is the basis of all measurement theory to the extent the sum of item scores is used to assign some overall value of ability to an examinee, as is the case on most tests. A second related but separate assumption central to IRT is *local item independence* (LII), meaning the response to each item is not influenced by the response to any other item. In other words, LII is achieved if examinees' respective ability value (θ) explains fully their performance on all items.

Because the assumptions of unidimensionality and local item independence are very strong assumptions, whether they can be satisfied in practice has been the source of much debate in both the educational assessment and second language testing literature. Broadly speaking, three approaches have been advocated for dealing with assumption violations: *overcome them* (using more advanced item-analysis methods like multidimensional IRT modeling), *mitigate them* (through modification of existing IRT models like Bejar's (1980) method), or *disregard them* (by relaxing the requirements of assumption satisfaction through claims of "essential unidimensionality" or "psychometric unidimensionality”). Following is a review of the means of testing IRT assumptions and the proposed solutions to managing violations.

2.1.3.4.1 Unidimensionality

According to Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978), testing the assumption of unidimensionality takes precedence over all other goodness of fit tests because the results of all other tests will be difficult to interpret if the assumption of unidimensionality is untenable (p. 487). To date, the most common test of unidimensionality has been some variant of factor analysis.

Stout (1984) was an early implementer of factor analysis in assessing unidimensionality, comparing classical and modern methods of factor analysis. However, Stout was not alone in suggesting a means of testing unidimensionality. In a quasi-meta-analysis of the existing indices at the time, Hattie (1984) found a staggering 87 different tests of unidimensionality (Table 1, pp. 51-4). To test the efficacy of these indices, Hattie employed multivariate 3PL modeling in a simulation study involving 36 models (2 levels of difficulty x 3 levels of guessing x 6 model combinations of dimensions and levels of discrimination) and found that indices based on answer patterns, reliability, component analysis, linear analysis and non-linear factor analysis were all ineffective. He therefore concluded the only reasonable means of assessing dimensionality is by summing the absolute residuals from the 2PL latent-trait estimation procedure.

The following year, Hattie (1985) once again conducted a simulation study, this time with 30 indices, and he found similar results to his previous study, concluding that researchers were incorrectly equating unidimensionality with other terms like reliability, internal consistency, and homogeneity. Reckase, Carlson, and Ackerman (1985) echoed this claim, stating factor analysis does not recover the underlying structure of dichotomous data, so it should not be used to demonstrate unidimensionality (due to its failure to meet the assumptions of tetrachoric correlations).

Alternatives to factor analysis have been proposed over the years. For example, Bejar (1980) created a method that tests unidimensionality via item parameter estimate comparison, where one set of estimates is obtained using all of the items on the test and the other using only the items contained within a particular subsection. When violations of unidimensionality are apparent, a decision must be made whether to accept the content-area-based estimates or total-test-based estimates. If total-test-based estimates are accepted, the implicit assumption is that the entire latent space is unidimensional and everything outside that space is "error," i.e., sources of variation are of no concern. On the other hand, if the content-area-based estimates are accepted, then there is implicit acknowledgement of a multidimensional latent space. The question then is whether the multidimensionality found is important for practical purposes.

One application of the Bejar (1980) method was conducted by Henning (1988), which was a follow-up study to Henning et al (1985), who examined the effects of a violation of unidimensionality on Rasch-1PL item and person estimates. In the 1988 study, Henning conducted a simulation study using a two-dimensional set of 60 items taken by a simulated 120-person sample and a two-dimensional set of 60 participants taking a 120-item test. Using the Bejar method to test for a violation of the assumption, Henning demonstrated that IRT models are robust to violations of item unidimensionality but not person unidimensionality. As a result, Henning concluded it is appropriate to use the Bejar method to eliminate items/persons responsible for causing the violation. In other words, because the Bejar method is sensitive to item-by-item and person-by-person changes, it is useful to employ because violations of the unidimensionality assumption diminish person/item separability.

Despite such applications, the Bejar method is not without critics. For example, Hambleton and Rovinelli (1986) compared four methods of testing dimensionality: 1) linear factor analysis 2) non-linear FA 3) residual analysis and 4) the Bejar method. Using five simulated data sets (each with 40 items and 1500 participants) with unidimensional and two-dimensional latent ability, two variables were manipulated: 1) the correlation between traits; and 2) the percent of items measuring the trait (fixed to either 50-50 or 25-75). Results show that the linear factor analysis method overestimated the number of dimensions in all simulations, the non-linear FA successfully estimated the number of dimensions in three of the five cases, and the residual analysis and Bejar methods failed across all conditions. The authors therefore concluded non-linear factor analysis is most promising method of assessing unidimensionality. Bejar (1988) later claimed Hambleton and Rovinelli misinterpreted the Bejar method and that it should still be considered valid, particularly for achievement tests that appear to examine multiple domains/content, but there are few recent references to the Bejar method in the literature.

Another proposed alternative to factor analysis is multidimensional scaling (MDS). Ayala and Herzog (1991), for example, proposed its use because they claimed it is easier to apply and less expensive than conducting factor analysis. In a comparison of MDS, confirmatory factor analysis (CFA), and exploratory factor analysis (EFA), Ayala and Herzog found both MDS and CFA successful in identifying all of the dimensions of the data. Despite these findings, however, their suggestion for use of MDS instead of factor analysis seems to have gained little traction in the field, at least with respect to assessing unidimensionality. One exception can be found in Meara, Robin, and Sireci (2000), who evaluated the suitability of multidimensional scaling (MDS) to dichotomous data due to the fact that if a test is multidimensional, it is unclear whether a composite score can be used to summarize examinee performance. Their results show MDS was successful in identifying multiple dimensions when the correlation between the dimensions was low but not when it was high.

Despite these proposed alternatives, factor analysis has persisted over the years as the preferred means of testing unidimensionality. A recent example of this can be found in Cook, Kallen, and Amtmann (2009), who conducted a confirmatory factor analysis of a pain test to examine the impact of n-size and non-normality on unidimensionality. In particular, the authors explored how CFA fit criteria were affected by two characteristics of item banks developed to measure health outcomes: a large number of items and non-normal data. Analyses were conducted using both observed data and simulated data. Results show that CFA fit values were sensitive to both data distribution and the number of items. As a result, the authors concluded that using traditional cutoffs and standards for CFA fit statistics is not recommended for establishing unidimensionality of large item banks but that it is the preferred method under other testing conditions.

Interestingly, most early analyses of unidimensionality rarely found clear satisfaction of the assumption. As a result, it was not long before researchers started proposing relaxed interpretations of the assumption. For example, Drasgow and Parsons (1983) conducted a simulation study to examine the effects of violating the unidimensional assumption. Using maximum likelihood to estimate item and person parameters for the 1PL, 2PL, and 3PL models, the authors specifically examined whether an item set was “sufficiently unidimensional” such that the use of IRT modeling could be justified. Using LOGIST software (Wood & Lord, 1976) with a simulated population of 1000 and responses designed to represent the normal ogive curve of the 2PL model, results show violations of unidimensionality did affect model fit adversely. They therefore argued unidimensionality must be achieved with IRT applications, but they also stressed the fact that if the first factor is “prepotent,” then unidimensional models do in fact provide adequate descriptions of multidimensional data (p. 198).

The concept of “sufficient unidimensionality” was first suggested by Reckase (1979), who examined the robustness of IRT models when applied to achievement tests because, according to Reckase, many achievement tests are by design multidimensional. Applying the IRT-1PL and -3PL models in particular, Reckase analyzed their fit of 10 data sets (five real and five simulated) and found that in cases where more than one dimension was present, the 3PL model fit one factor and disregarded the others while the 1PL summed the loadings. As a result, Reckase claimed that both person and item parameters are stable provided at least 20% of the variance can be explained by the primary factor (i.e., 20% of the variance explained would result in a 3PL-discrimination parameter of about .60). Reckase went further to state that even if less than 20% of the variance could be explained by the first factor, the person ability estimates would still likely remain stable, although the same is not true for the item parameter estimates (p. 227). In short, Reckase made the case, perhaps for the first time, that unidimensionality could be seen as relative and not necessarily absolute for IRT model applications.

Stout (1987, 1990) furthered this claim by creating a test of what he called *essential unidimensionality*. Based not on factor analysis per se but on the principle that unidimensionality should hold when sampling from a subpopulation of examinees of approximately equal ability, Stout found that a non-parametric IRT model is a more straightforward means of testing unidimensionality than factor analysis, and that only if a multidimensional latent space appears to exist should factor analysis be employed.

In the field of second language testing (SLT), Henning, Hudson, and Turner (1985) claimed that unidimensionality is “mythical” in almost all cases, so the real question is to what degree items must appear to be measuring the same trait in order for IRT analysis to hold (p. 142). They also stated that CTT assumes unidimensionality insofar as a summative test score is reported as a reflection of ability, so violations of unidimensionality are problematic for both CTT and IRT analyses.

To support their claims, Henning et al (1985) examined data derived from 312 examinees who took the UCLA English as a Second Language Placement Exam (ESLPE), which consisted of 150 four-option multiple-choice items (30 each in five subsections – listening comprehension, reading comprehension, grammar, vocabulary, and error detection). Conducting a Rasch analysis of the test as a whole and each of its five subtests, Henning et al employed the Bejar (1980) method to examine unidimensionality. Results demonstrate a linear relationship between each test subsection and the test as a whole, so the authors claimed unidimensionality was not violated. They found further support for their claim in the form of an independent factor analysis conducted on the same test (Davidson, 1985), as well as its very high internal consistency (KR-20 = .96). In short, and although not explicitly stated, Henning et al appear to have implied there is a general language proficiency trait that is sufficiently unidimensional for use with IRT analysis, at least as reflected by the ESLPE.

Lynch, Davidson, and Henning (1988) also used the ESLPE to measure the impact of person dimensionality on differential item functioning (DIF) estimates. With a sample size of 678 participants, Lynch et al found no real difference between the top and bottom 27% other than ability, although there was some clustering by major (more engineering & science majors in the top 27% and more arts & humanities majors in the bottom 27%). The authors therefore concluded unidimensionality was not violated with respect to person parameter estimates.

Henning (1992) further refined his previous ideas regarding dimensionality by introducing the concepts of *psychological dimensionality* and *psychometric dimensionality*. Taking a relatively strong position, Henning stated:

> Item response theory (IRT) models are…considered by some to be suspect as appropriate measurement tools with language assessment data…since communicative language performance is assumed to be by nature multidimensional…However, it is the position of this paper that these criticisms may be based on insufficient awareness of the nature and constraints of unidimensionality and multidimensionality and on inadequate appreciation of the distinction between what may be termed 'psychological' dimensionality and 'psychometric' dimensionality. (p. 2)

Stating that psychometric unidimensionality is similar to psychological unidimensionality in that both refer to the capacity of a test to measure some primary dimension or trait, Henning went on to state that psychometric unidimensionality can be present when the test measures a variety of correlated underlying psychological dimensions.

An earlier study by Reckase, Ackerman, and Carlson (1988) reinforces Henning's (1992) claim that psychometric unidimensionality can be demonstrated even when psychological multidimensionality is being measured. In this study, the authors used the 2PL model to analyze two datasets – one real and one simulated – to demonstrate that multidimensional ability can be considered psychometrically unidimensional provided each item in the set is measuring the same composite ability (e.g., FLP). In their conclusion, Reckase et al state that:

> Most items require more than one ability to obtain a correct answer. This would seem to preclude the use of unidimensional IRT procedures with such items…Rather than specifying that items need to measure only a single trait, the results presented here show that the unidimensionality assumption implies that items need only measure the same composite of abilities as indicated by multidimensional IRT analysis. (p. 203)

Further arguing the case that most academic constructs are multidimensional by nature, Dawadi (1999) examined the robustness of IRT modeling in the face of deliberate violations of unidimensionality. Comparing the root mean squared errors (RMSEs) of both unidimensional and simulated two-dimensional data, Dawadi found that minor violations of unidimensionality were tolerable (provided the correlations between all factor pairs are greater than .80). Dawadi also reiterated the fact that even, e.g., vocabulary ability can be shown to be multidimensional if analyzed at a granular-enough level, so unidimensionality should be seen as a relative rather than absolute concept, despite the fact other studies have shown IRT models are not valid when unidimensionality is violated (e.g., Dorans & Kingston, 1985; Downing & Haladyna, 1996; Folk & Green, 1989; Oshima & Miller, 1990; Walker & Beretvas, 2000).

Zhang (2008) provides a relatively recent summary of the issue of dimensionality, stating that when a unidimensional model is applied to tests with two ability traits, the unidimensional ability estimate shows each examinee’s original standing on two traits by one statistic and that this statistic will probably reflect the stronger trait more than the weaker one. The question then is to decide whether the influence from the weaker trait can be ignored, or to the extent it cannot, whether items that measure multiple ability dimensions do so to the same degree. If this is the case, unidimensionality can be assumed to not be violated, as was the case in two studies of dimensionality on two large-scale standardized tests.

Two relevant applications of IRT support this argument for essential unidimensionality. In the first study, Childs & Oppler (1999) conducted an IRT analysis of the Medical College Admission Test (MCAT) based on the presumed multidimensionality of each of the three test sections – verbal reasoning, physical sciences, and biological sciences. Results showed that while some items in each section were not completely homogenous, violation of unidimensionality was not a particular concern; i.e., essential unidimensionality was achieved. In the second, Schedl, Gordon, Carey, and Tang (1996) examined the dimensionality of the TOEFL reading test in an effort to determine whether reasoning skill – a construct putatively tested in four item types appearing in the ETS test specifications – was a separate dimension from general reading ability. Using Stout’s (1987) procedure for assessing essential unidimensionality and McDonald's (1982) nonlinear factor analysis procedure, Schedl et al found a two-factor solution but no evidence of a reasoning-skill factor. Instead, it appeared the second factor was related either to passage content or passage position (the final two passages had the highest second-factor loadings). As a result, the authors claimed support for the finding of Lunzer, Waite, & Dolan (1979) that reading comprehension is a single construct (c.f., Freedle & Kostin, 1993; Grabe & Stoller, 2002), or rather, that the psychological construct is multidimensional but the psychometric construct is unidimensional (op. cit., Henning, 1992; Reckase et al., 1988). As will be illustrated, it could be the case the second order factor is what is now known as a testlet effect.

2.1.3.4.1.1 Multidimensional IRT (MIRT) vs. Unidimensional IRT (UIRT)

For all the attempts to relax the unidimensional assumption, there is still a wealth of evidence multidimensional data negatively impact the model fit of unidimensional IRT models. This has been demonstrated in many areas, including bias detection (Oshima & Miller, 1992), test equating (Bolt, 1999; De Champlain, 1996), and test score and subscale score interpretation (Tate, 2004), among others. As a result, the development of multidimensional IRT (MIRT) models began to emerge in parallel with the research on so-called essential unidimensionality.

As Hartig and Hohler (2009) explained, the advantage of MIRT models over unidimensional IRT (UIRT) models is strongest under three circumstances: 1) when unintended multidimensionality is uncovered in a presumed unidimensional construct; 2) when modeling latent covariance structures between ability dimensions; and 3) when modeling interactions of multiple abilities that are required to solve specific test items.

Other researchers (e.g., Gibbons, Immekus, & Bock, 2007) have also demonstrated the added value of MIRT modeling with computerized adaptive tests (CATs), and still others have shown how MIRT modeling can improve parameter estimates when the number of items is small (J. Zhang, 2004).

Despite the positive results MIRT models have generated, many researchers have argued against their use because of their complexity and their need for massive numbers of data. Zhang (2008), for example, defended the use of standard IRT models with many types of multidimensional data, citing the computer programs MULTILOG (Thissen, 1991) and BILOG (Mislevy & Bock, 1990) as examples of relatively straightforward IRT modeling programs that can estimate both ability and item parameters, while existing MIRT modeling programs like TESTFACT (Wood et al., 2003) and NOHARM (Fraser & McDonald, 1988) can estimate only item parameters. Zhang also argued unidimensional models are more parsimonious and are therefore preferred, particularly given the fact their estimates are satisfactory provided they either meet Stout's (1987, 1990) requirements for essential unidimensionality or if the multiple dimensions uncovered are highly correlated. In brief, Zhang stated that:

> In educational testing, construct-relevant secondary dimensions are usually correlated with the main construct, thus the essential unidimensionality assumption probably holds and the unidimensional ability estimates adequately represent the ability level of students on the main construct. Regarding the construct-irrelevant dimensions, such as cultural background, curricular emphasis, and speed of work, although they may not be related to the studied cognitive trait, the chance that any of them will

> affect a multitude of items in a well-designed and well-administered test is also small, thus the unidimensional model may still fit. (p. 164)

Ip (2010) is another advocate of the standard IRT model, claiming MIRT is empirically indistinguishable from locally dependent IRT, so multidimensional response data do not necessarily require the use of MIRT models. Ip further argues that "…unidimensionality is more of an abstract ideal than a reality…In fact, it is hard to argue that truly valid unidimensional tests exist in any subject matter area" (p. 397). As a result, Ip states that if there is a predominant general factor in the data and if the dimensions beyond that major dimension are relatively small, the presence of multidimensionality has little effect on item parameter estimates and the associated ability estimates. On the other hand, if the data are multidimensional with strong factors beyond the first one, utilizing a locally dependent IRT model will improve fit and avoid the need to employ one of the considerably more complex MIRT models in existence.

Stone, Ye, Zhu, and Lane (2010) conducted one of the more recent studies that examined the cost-benefit tradeoff of MIRT models. In their study, the authors compared three means of improving the reliability of subscale scores. Analyzing data collected from 10,545 eighth-graders who took the Delaware Student Testing Program (DSTP) test, the authors conducted an assessment of essential unidimensionality (Stout, 1987, 1990) via exploratory factor analysis, using Mplus to calculate eigenvalues, fit statistics, and residuals.[7] For the 48 dichotomously scored items, a MIRT model equivalent to Reckase's (1997) MIRT-3PL model was employed along with two IRT models. Results show that the MIRT model led to greater reliability of the subscale scores, but there were

[7] A confirmatory factor analysis was not conducted because eigenvalues were unavailable and the chi-square goodness-of-fit index was not calculated because of its sensitivity to large n-sizes.

practical issues that limit its routine application. For example, it was significantly more complex to employ than the IRT models commonly employed. Moreover, higher inter-factor correlations were found with the MIRT approach, indicating that less unique variance existed among the subscale scores. As a result, the authors concluded that is would make sense to either increase the multidimensionality of tests to maximize the utility of the MIRT model or continue using the unidimensional IRT models because essential unidimensionality can in fact be demonstrated, at least with these types of data.

Yen & Walker (2007) also examined the impact of MIRT modeling on a test with subsections, each of which presumably measured a different trait (listening, speaking, reading, and writing). Running multiple analyses of data collected from 12,008 elementary school students in second to fifth grade, the authors found that MIRT models could better model the composite scores, but that, citing Davey and Hirsch (1990), MIRT analysis may not be as capable of discriminating among examinees traits because of the increase in parameters needed to obtain estimates (at least for tests of fewer than 100 items). The authors went on to say that these types of problems have hindered the development of practical applications involving MIRT models and that they are essentially ignored in favor of the less-complex UIRT models.

2.1.3.4.2 Local Item Independence

Differing from dimensionality but an equally important assumption of IRT models is local item independence (LII). Whereas dimensionality is concerned with whether each item on the test is measuring one or more dimensions of ability, LII is concerned with whether the response to any one item influences the response to any other.

Yen (1993) provided a good summary of the causes of local item dependence (LID; i.e., a violation of LII), including test-external factors like assistance, interference, speededness, fatigue, practice, the explanation of a previous answer, scoring rubrics, and raters, as well as test-internal factors like item/response format, passage dependence, and item chaining.

In her study, Yen described how performance assessments are susceptible to violations of LII because multiple items often are based on a single setting (e.g., on a test in language arts, a setting might be established with a short story and then the student is asked to contrast two characters in the story, provide and defend an alternative ending, and relate events in the story to a personal experience). Analyzing three LID contexts – the extent of within-passage LID on performance assessments vs. multiple-choice tests, the extent of LID on follow-up math item sets, and the effect of inappropriate LII assumptions on model statistics – Yen compared data from the Comprehensive Test of Basic Skills, Fourth Edition (CTBS/4), to data from the Maryland School Performance Assessment Program (MSPAP), and found, perhaps surprisingly, that the reading comprehension component of the CTBS/4 exhibited little LID, while the MSPAP did. Unfortunately, Yen did not describe the reading comprehension part of the test in detail (see p. 194), so it is not possible to discern why multiple items related to the same passage did not exhibit LID. Nevertheless, Yen concluded that violation of LII typically leads to overestimates of test information and reliability while underestimating the standard error of measurement, suggesting the need to use *testlets* (Wainer & Kiely, 1986, 1987) to offset this effect. However, Yen warned a lot of information is lost when analyzing at the testlet level, as demonstrated in Figure 1.

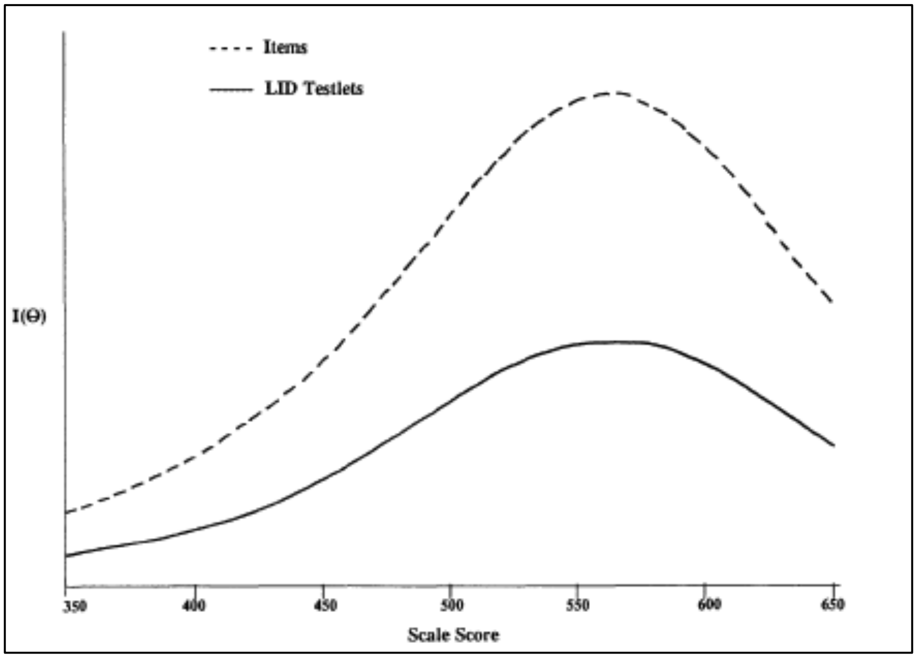


Figure 1. Information for a pair of Math-Content LID items when they are scaled separately or as a testlet (from Yen, 1993, p. 204)

In conclusion, Yen identified six procedures that can be employed to reduce LID, or when not feasible, to analyze the data in a way that ensures LID has a minimal impact on parameter estimation:

1. Create independent items.
2. Administer tests under favorable conditions (e.g., eliminate likelihood of fatigue).
3. Combine the grading of LID items.
4. Review tests to identify LID items *a priori*.
5. Create separate scales to grade items.
6. Use testlets.

Yen went on to state that managing LID later in the testing process has the added advantage of having less impact on test design, administration, and scoring, so using the latter procedures is desirable when possible.

As with the assumption of unidimensionality, Henning spent considerable time discussing the assumption of local item independence (LII). In particular, Henning, Hudson, and Turner (1985) made the relatively bold claim that LII does not appear to be violated on most language tests, and that the unitary vs. divisible-trait hypothesis (see, e.g., Farhady, 1983; Oller, 1976, 1983) is not applicable because tests can have items regrouped according to factors to run multiple IRT analyses. Moreover, they claimed cloze tests do not violate LII inasmuch as internal consistency estimates are unaffected by such violations. Nevertheless, these claims are questionable given the findings of other research on the dimensionality of language proficiency (e.g., Bachman & Palmer, 1981b, 1982, 1989; Fouly, Bachman, & Cziko, 1990; Llosa, 2007, 2008; Sawaki, 2003, 2007; Sawaki, Stricker, & Oranje, 2009a; D. Shin, 1999; S. Shin, 2005).

From a theoretical perspective, Henning (1989) claimed that at least a dozen different definitions of LII were found in the literature, not all of which are compatible or sufficient in their level of detail. As a result, Henning attempted to synthesize the definitions and further delineate the concept, stating three conditions must be satisfied in order for LII to be achieved: 1) unidimensionality; 2) uncorrelated local items, meaning responses to any given item are uncorrelated with any other item *at a fixed ability level*; and 3) non-invasiveness, meaning the performance on one item does not influence the performance on any other item *for any given individual*. It is the distinction of these two latter conditions that differentiates Henning's definition of LII from other more widely known definitions (e.g., Lord & Novick, 1968), which Henning claimed inadvertently equates unidimensionality and what he termed *classical item independence*. Perhaps more importantly, Henning reiterated his claim in an earlier paper (Henning, 1988) that

LII is required of CTT as well as IRT in that CTT relies on internal consistency and difficulty estimates, both of which would be distorted if LII were violated. As a result, it is no more reasonable to violate LII when employing CTT than it is when employing IRT. In this sense, Henning claimed LII is the central assumption of all testing theory (p. 95).

Because LII is such a stringent assumption, researchers have proposed several modifications to IRT models in an attempt to overcome its requirements (e.g., Bradlow, Wainer, & Wang, 1999; Braeken, Tuerlinckx, & De Boeck, 2007; Hoskens & De Boeck, 1997; Ip, 2000, 2002; Ip, Smits, & De Boeck, 2009; W. Wang & Wilson, 2005). In particular, Keller, Swaminathan, and Sireci (2003) proposed two strategies for dealing with context-dependent items (locally dependent) items on tests designed to be scored dichotomously: 1) ignore the LID and proceed as planned; or 2) model the LID through polytomous (testlet) scoring. Citing concerns from previous research about the need to account for a loss of information through testlet formation (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; W. Yen, 1993), Keller et al concluded that, while ignoring LID does lead to an overestimation of statistics, trying to offset the overestimation through polytomous scoring may cause an underestimation and may even lead to improper classification decisions. In other words, "…the way test specialists approach the scoring of context-dependent item sets will not only affect the estimates of test characteristics such as reliability and information, it will also affect the outcome of the test for many examinees" (p. 218). This is an important finding for this research given that the focus is on the classification accuracy of various item-analysis methods. In other words, whether LID is present in the data and how it is managed is an important factor when trying to maximize classification accuracy.

2.1.3.4.2.1 Testlet Response Theory

Just as MIRT models were created in response to violations of the assumption of unidimensionality, *Testlet Response Theory* (TRT) models were created in response to violations of local item independence. Wainer and Kiely (1986, 1987) are credited with coining the term *testlet*, defined as “a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow” (p. 190). Interestingly, the motivation for testlet response theory has its genesis in trying to maintain the viability of computerized adaptive tests (CATs), which have IRT as their basis and therefore require the assumption of LII to be met, particularly because the sequence of item presentation is based directly on whether the previous item had been answered correctly. Stated differently, “context effects” could be seen as irrelevant when participants take an identical test, but because CATs by design create a novel test for each examinee, context effects can heavily influence the outcome and therefore must be avoided. Wainer and Kiely further explained that context effects can be due to three factors: 1) *item location* (e.g., item difficulty has been shown to vary depending on an item’s location on a test (e.g., W. Yen, 1980)); 2) *cross information* (i.e., the answer to one item contributes to the answer of another); and 3) *unbalanced content* (i.e., repeated emphasis on the same content or theme, as is the case when several items are generated from the same prompt).

In their results, Wainer and Kiely acknowledged a loss of information but stated that testlets remain a better option for scoring when full information is not required. In short, they consider TRT a good compromise between losing some information and having IRT models better fit the data:

> In both the simulations and the analysis of real data, we have shown how this model can be used to score such tests and provide estimates of the tests' precision that are neither as optimistic as models that incorrectly assume conditional independence nor as pessimistic as those that only use total score. (p. 124)

In two later studies about the use of TRT in response to violations of local item independence – Bradlow, Wainer, and Wang (1999), and Wang, Bradlow, and Wainer (2002) – the authors explained that standard IRT models fit to dichotomously scored responses ignore the fact that item sets often are based on a single prompt (e.g., a reading comprehension passage). In such cases, at least some items are unlikely to be locally independent, so standard IRT models that assume LII is satisfied will overestimate the precision with which examinee proficiency is measured. This in turn may lead to inaccurate inferences, such as prematurely ending an examination in which the stopping rule is based on the estimated standard error of examinee proficiency (i.e., on a CAT).

To model examinations that contain a combination of independent items and testlets, Bradlow et al (1999) modified a standard IRT model to include an additional random effect for items nested within a testlet. Using a Bayesian framework, the authors applied their modified IRT model and a standard IRT model to SAT data. Among other results, the authors concluded, importantly, that violations of LII appear to be more problematic on CATs than paper-based tests (p. 167). Moreover, even though IRT models fit each item of the testlet as if it were an independent item (which led to an overestimation of the parameters), the amount of error introduced was considered acceptable given the testlets were short (4-6 items). Conversely, and as Keller et al (2003) found, treating the testlet as a single item resolved the LID problem but had two shortcomings. First, it resulted in a loss of information, and second, all of the items in the testlet had to be calibrated together to identify the parameters of the testlet, which limited the item-selection algorithm to a much smaller item bank from which to select items.

More recently, Rijmen (2009, 2010) analyzed the impact of testlets ("item bundles") using three different models: a bi-factor model (Gibbons & Hedeker, 1992), a testlet model (Bradlow et al., 1999; Wainer, Bradlow, & Wang, 2007), and a second-order model. As Rijmen explained, the latter two are formally equivalent and can be structured as restricted bi-factor models, where the bi-factor and second-order models originate out of the factor-analysis tradition and the testlet model of out educational measurement. That is, the testlet model is a special case of the bi-factor model, obtained by constraining the loadings on a specific dimension to be proportional to the loadings on the general dimension (p. 364).

Fitting the three models to the 20-item reading comprehension section of an international English-proficiency test (four testlets with five items each) taken by 13,508 participants, the bi-factor model fit the data best on a variety of estimates, including the likelihood-ratio test statistic. Subsequently, Rijmen fit the IRT-2PL model to the data and found the standard errors "too optimistic" (i.e., underestimated) when LII did not hold. He therefore concluded LII cannot be assumed when testlets are part of an exam.

2.1.3.4.3 Unidimensionality vs. Local Item Independence

Within the context of IRT modeling, an abundant literature has emerged regarding the relationship between unidimensionality and local item independence. While some researchers conflated the two concepts (e.g., Lord & Novick, 1968), others have claimed they are separate but related (e.g., Henning, 1989), while others still claim they are independent (Ip, 2010; Meara et al., 2000). It is argued here they are independent because local item independence is concerned only with the relationship *across all items* while unidimensionality is concerned only with the relationship *within each item*. Consider a crossword puzzle. It is easy to argue there is a lack of local item independence in that as each word is correctly filled in, responses to adjacent words become easier to identify. On the other hand, the dimension(s) being measured by the items in the puzzle, whether general content knowledge or some combination of content knowledge and spatial ability, for example, is a completely separate matter. In other words, the dimension(s) being tested by any one item, whether one or many, has no relationship to whether the response to one item influences the response to another. Any relationship that is demonstrated between the two would be merely coincidental.

Items based on a single reading passage are a good example of this potentially coincidental relationship: the item bundle presumably lacks local item independence because all of the items are based on the same prompt; it could also be the case each item is measuring multiple dimensions of ability (e.g., reading comprehension and subject matter expertise). In this case, both assumptions are violated, but for coincidental (independent) reasons. In other words, in any given context, both assumptions can be satisfied, one can be violated and the other satisfied, or both violated. Even in the latter case, though, the violations would be independent. Ip's (2010) mathematical analysis of the two assumptions reinforces this theoretical argument, demonstrating the concepts are "unequivocally distinct mathematical entities" (p. 396).

2.2 Assessing Model Fit

Whatever model is chosen, there are numerous means of assessing model fit. Below are the three components of assessing model fit that will be considered in this dissertation.

2.2.1 Rationale for Model Choice

The first criterion for satisfying model selection is the rationale on which the choice of model is based. As illustrated throughout the literature review, there are several components of each item-analysis method that can help determine which is most appropriate for a test, and at first blush, it would appear method/model selection is a straightforward endeavor. For example, if a single-use 20-item multiple-choice algebra exam were administered to 50 students at a local school, CTT item analysis would probably be appropriate. In fact, IRT modeling requires far more data than available in this example, so it would not even be an option. Alternatively, if a 10-item constructed-response exam was administered to 10,000 students and each was exam scored by two raters selected randomly from a pool of 200 raters, a polytomous IRT model might be most appropriate because it could take into account both the relative difficulty of each item and rater severity/leniency such that the resulting scores would put all participants on equal footing because of the true interval scale that would result from the analysis.

Unfortunately, model choice is not as clear as it may seem in these examples. Consider the composition of the test analyzed in this study (see Appendix A). On the surface, it would appear the IRT-3PL model is most appropriate because the test consists of 70 dichotomously scored multiple-choice items that presumably will discriminate among participants differently and result in pseudo-guessing in some cases, particularly for some of the lower-ability examinees. On the other hand, it is possible the assumption of unidimensionality will be violated because the test contains three different item formats – reading comprehension, cloze, and grammaticality judgment – which, taken collectively, purportedly measure general English proficiency but could in fact measure multiple dimensions.

Whether the test as a whole is measuring a single dimension is an empirical matter and will be tested via factor analysis, but even if it is found to be "essentially unidimensional," it is very likely local item independence is violated because of the presence of three long testlets – two reading passages with 15 and 20 reading-comprehension items, respectively, and a 15-item cloze test based on a third reading passage. As a result, it could be argued testlet response theory (TRT) is most appropriate for this exam, or perhaps an even more sophisticated MIRT model if in fact unidimensionality has been violated as well.

One could also argue for application of the Rasch-1PL model. Because this test was designed to be administered only once with no pilot testing, it could be argued any misfitting items should be removed from the analysis before scoring, regardless of whether any of the IRT models fits the data well. Stated differently, it could be argued the exploratory "model must fit the data" approach of IRT is too lenient for this test given that its items have not been piloted in advance, so employing the confirmatory "data must fit the model" approach of Rasch is more appropriate.

Of course, one could also rationalize the use of CTT analysis. Because this is a single-use test that will never be linked or equated to any other version or administration, even Rasch analysis could be considered unnecessarily complex. As long as internal consistency can be established, there may be no real advantage to employing one of the more complex IRT approaches. In fact, even though many researchers across a wide range of subjects have espoused the virtues of IRT, claiming that it can improve the precision and validity of psychological research (e.g., Reeve, Hays, Chang, & Perfetto, 2007; Reise, Ainsworth, & Haviland, 2005), other researchers have been left to explain why it has not been more widely applied outside large testing firms, state agencies, and a few school districts (e.g., De Champlain, 2010; Reise & Henson, 2003; Scherbaum, Finlinson, Barden, & Tamanini, 2006).

Acknowledging the theoretical superiority of IRT models to CTT models, the fundamental question this latter group poses is whether IRT modeling is sufficiently superior to justify the added complexity and cost associated with IRT analyses. In the end, they conclude IRT should play a significant role in future assessments, but they should be seen more as a complement to CTT analysis and not necessarily a wholesale replacement of it (De Champlain, 2010, p. 117), particularly when, e.g., a participant's relative standing will experience little change whether CTT or IRT item-analysis methods are employed (Reise & Henson, 2003, pp. 99-100).

Linn (1990) took this argument one step further, arguing IRT use is sometimes unsupported, particularly with achievement test data. Acknowledging that "IRT is the most important technical development in measurement in recent years (pre-1990)," Linn argued test use and score interpretation are more important concerns and should not be forsaken in favor of the somewhat indiscriminate use of IRT, as had been the case at the time of Linn's publication. Linn concluded by stating: "If items that are found to be most sensitive to instruction are eliminated so that the IRT assumptions are better satisfied, there is a real danger that IRT will do more to decrease than to increase the validity of achievement test scores" (p. 136).

Wainer and Thissen (1985, 1987) among others (e.g., Barnes & Wise, 1991; Dinero & Haertel, 1977; Hambleton & Traub, 1971, 1973; van de Vijver, 1986) discussed this impact of estimating ability with the wrong item-analysis method. Using simulated data and examining the extent accuracy was influenced by model misfit, Wainer and Thissen found the IRT-3PL model to be superior for long tests (more than 40 items) but disadvantageous with shorter tests. As a result, they concluded that item-analysis-method selection should at least in part be dictated by test length and sample size.

Sireci (1991) found similar results for small sample populations. Examining data collected from the administration of a 28-item reading comprehension test with a sample population of 428 (the combination of three administrations over three years, Sireci performed CTT analysis (*p*-values, KR-20, and Point Biserial) and IRT analysis (1PL, 2PL and 3PL) using chi-square difference testing of the -2loglikelihoods. Results show that the IRT-2PL model fit the data best, but none of the models exhibited item-parameter stability because of the small sample size and small number of items. As a result, Sireci concluded it is possible to use IRT in some small-scale testing contexts, but its benefit over CTT analysis is minimal, so it is probably unnecessary and perhaps even undesirable.

This debate among researchers continues not only over the presumed merits of IRT over CTT, but also over the relative merits of IRT models with respect to each other. As mentioned, the amount of additional data required to maximize an IRT model's utility is considerable, so many researchers have questioned the benefit of adding, e.g., the guessing parameter to the already-complex IRT-2PL model. For example, Hernandez (2009), citing Hambleton, Crocker, Masters, van der Linden, and Wright (1992), stated:

> The inclination to guess is an idiosyncratic characteristic of particular low ability examinees. Lucky guessing is a random event. Neither feature contributes to valid measurement of a latent trait. Parameterizing guessing penalizes the low performer with advanced special knowledge and also the non-guesser. Rasch flags lucky guesses as unexpected responses. They can either be left intact which inflates the ability estimates of the guessers, or removed which provides a better estimate of the guessers' abilities on the intended latent trait. In practice, 3-P guessing parameter estimation is so awkward that values are either pre-set or pre-constrained to a narrow range. (p. 217)

Not all researchers agree. Barnes & Wise (1991) proposed a modified IRT model to account for guessing even when a sample size is small because they claimed the impact of guessing is even more dramatic under such circumstances. As a result, they suggested a modified 1PL model with a fixed nonzero lower asymptote due to the fact the Rasch-1PL model is robust to violations of equal discrimination but not to the presence of nonzero asymptotes (Dinero & Haertel, 1977; Hambleton & Traub, 1971; van de Vijver, 1986). Using their model, results were found to be comparable for simulations run with 50, 100, and 200 examinees and 25 & 50 test items, respectively, and in fact much better compared to the Rasch-1PL model with a zero lower asymptote. They therefore concluded their modified 1PL model should be used for multiple-choice tests when the sample size and/or number of items is insufficient for the data-hungry 3PL model.

Overall, von Davier (2009) perhaps summarized this debate best:

> If questioned about their beliefs, psychometricians in one camp would argue the firm conviction that the Rasch model is mathematically elegant and intuitive as well as plausible for practitioners, pointing out the advantages of a simple model that "counts" every item in the same way. Psychometricians of another camp would argue that the 3PL is much more flexible and is suitable to take into account that some item types have a nonzero probability to be solved by guessing and other random response strategies. This leads us to ask: which of these models is appropriate for test data of a certain type? Or better: is there a correct answer to this question? Unfortunately, choosing between the 3PL and the Rasch model or other variants of item response theory (IRT) does not become easier even after it is understood that these models are closely related. If an extraneous principle such as Occam's razor is used, one may argue in favor of the simpler model; if the goal is to be more flexible in terms of the ability of the item function to fit different trace lines, a model with more parameters may seem appropriate. To make matters worse, there are alternatives that can be substituted for the 3PL model when the issue is to account for random response strategies or guessing. (p. 111)

In the end, von Davier argued that models that account for guessing are not necessarily superior and that practical concerns may dictate which model is chosen rather than considerations about how guessing should be conceptualized within a model (p. 114).

As demonstrated, a rational argument can be made for the employment of virtually any item-analysis method, which only reinforces the need to examine multiple models on a single set of data to determine whether/how model choice impacts the validity of the resulting test scores.[8] In short, ethics standards governing international testing specialists mandate more responsible model-choice decisions than basing it on rationale alone, especially when the fates of thousands of examinees hang in the balance.

2.2.2 Estimation of Item Parameters

Although trying to match model design with the characteristics of the test is an important first step in identifying the model that will lead to the most valid scores, as shown there is no guarantee one particular model will fit the data best, or even well. As a result, actually running a set of item analyses can help clarify which item-analysis method is most suitable for a particular set of data.

Over the years, numerous parameter estimates, both within CTT and IRT, have been proposed. Although others exist, those most commonly utilized statistics for dichotomously scored multiple-choice items include the following:

[8] In addition to the models discussed thus far, there has also been research into whether a polytomous scoring model should be used for single-select multiple-choice questions, the rationale being that not all distracters are created equal and there may be some systematic differences between the wrong answers chosen by high-ability examinees and low-ability examinees. Hakstian and Kansup (1975) summarized some of the earliest work on this issue, explaining the concepts of *elimination testing* (Coombs, Milholland, & Womer, 1956) and *confidence/probabilistic testing* (Dressel & Schmid, 1953). They concluded that because both methods require special training for examinees and considerably more testing time, considerable improvement in reliability and validity must be achieved in order to justify their use. In their study, neither reliability nor validity was consistently increased by the experimental methods, so there was little reason to recommend either over traditional dichotomous scoring. Kansup and Hakstian (1975) published a second article on the same topic nearly concurrently, this time framing it as *differential weighting* but coming to the same conclusion – there is very little reliability or validity to be gained by polytomous scoring of multiple-choice items, so such methods should not be employed given the substantial time and cost associated with them (c.f., Adams, Griffin, & Martin, 1987).

CTT

Item difficulty (p)

Item discrimination (D)

Internal consistency (r_{pb}), and

Global reliability (Cronbach's $\alpha,$ KR-20/21)

IRT

Item difficulty (b),

Item discrimination (a)

Pseudo-guessing (c)

Model fit (TIF, Test Information Function)

Model misfit (SEM, Standard Error of Measurement)

Because these statistics are derived differently for each theory, it is important to be able to create equitable comparisons across the two theories. Fan (1998) provided a good summary of how this can be accomplished. In a quasi-replication of Lawson (1991), Fan evaluated two issues: 1) the empirical relationship between IRT and CTT item and person statistics; and 2) the extent to which IRT and CTT item statistics are invariant across participant samples. As noted elsewhere, Fan explained that the weak theoretical assumptions of CTT are one of the theory's strengths, but it suffers from a circular dependency: person statistics are item dependent and item statistics are sample dependent. IRT, on the other hand, generates item and person statistics that are sample and item independent, respectively, but its strong theoretical assumptions are sometimes difficult to satisfy, as illustrated throughout this literature review.

Using data from the Texas Assessment of Academic Skills (TAAS), Fan generated three different sampling plans (random, gender, high/low ability) so CTT and IRT statistics could be compared multiple times. With each sample population equal to 1000 (chosen from a pool of 193,000 eleventh-graders in Texas public schools), Fan analyzed the responses to 48 reading and 60 math multiple-choice items. Person statistics were compared by correlating IRT ability values with observed scores; item statistic comparisons included item difficulty (*b* with *p*) and item discrimination (*a* with bias-corrected r_{pb}). Overall, the two sets of statistics were compared based on the degree of invariance within their respective ordinal rankings.

Interestingly, the correlations of CTT and IRT person statistics across all sampling plans and model comparisons were very high (range = 0.966-0.997). Moreover, correlations for item difficulty (*p)* were very high, ranging from 0.862 to 0.999. In particular, the Rasch model correlated almost perfectly with CTT item difficulty (.999 for all sampling plans), leading Fan to conclude that, "the results here would suggest that the Rasch model might not offer any empirical advantage over the much simpler CTT framework" (p. 371).

With respect to item discrimination, the correlations between r_{pb} and *a* were also high, but less so than the other statistics, and with some notable exceptions (e.g., primarily with reading: random (.264), females (.358), and males (.199)). Interestingly, Fan stated that, "…CTT and IRT may yield noticeable discrepancies with regard to which items have more discrimination power, which, in turn, *may lead to the selection of different items for a test*" (p. 373, emphasis added), a very important conclusion considering the focus of this dissertation.

With respect to invariance, the correlations again were very high, ranging from 0.862 to 0.993, and with nearly all above 0.90. In this case, Fan measured the invariance of item statistics within samples and found that all models tracked similarly, with random reading having the highest correlation, random math the second highest, and so forth. In short, Fan stated that the overall findings failed to support the IRT framework as superior enough to CTT to justify its application for these data and reiterated the prediction put forth by Thorndike (1982) regarding IRT:

> For the large bulk of testing, both with locally developed and with standardized tests, I doubt there will be a great deal of change. The items that we will select for a test will not be much different from those we would have selected with earlier procedures, and the resulting tests will continue to have much the same properties. (p. 12)

In contrast to Fan (1998), Adedoyin (2010) did find differences in invariance between CTT and IRT. Examining data collected from a random sample of 5000 participants on a 40-item math exam to determine whether invariance across person ability estimates could be achieved with both CTT and IRT modeling, results show that CTT estimates were not invariant (and in fact exhibited great variation) while the IRT-2PL model was invariant. As a result, Adedoyin concluded IRT is superior to CTT for similar testing contexts.

Anderson (1999) also examined the benefits of moving from CTT to IRT in a study involving 6000 students who took the 50-item Mathematics 12 exam, which is administered to prospective high school graduates in British Columbia, Canada. Because the IRT 3PL seemed to be the most rationale choice given the nature of the test, it was chosen as a means of comparison with traditional CTT indices. Strikingly, Anderson found that results were virtually indistinguishable for all letter-grade levels. As a result, Anderson concluded application of the IRT-3PL model was not justified given the substantial increase in complexity and cost.

2.2.3 Estimation of Classification Accuracy (CA)

Although parameter estimation can illuminate how well a particular model fits a particular set of data, more specific reliability indices are often required. For example, classification consistency and classification accuracy are of supreme importance for tests with cut scores. Emerging in the 1970s as an alternative measure of reliability for criterion-referenced tests, *classification consistency* is the degree to which classifications across parallel-form test administrations align; *classification accuracy* is defined as the extent to which examinee classifications based on observed scores match the true scores of the examinees (Livingston & Lewis, 1995, p. 180).

Throughout the decade and into the 1980s, numerous researchers posited measures of classification consistency and accuracy, including Huynh (1976), Livingston and Wingersky (1979), Subkoviak (1976a, 1976b, 1988), and Wilcox (1981). Common to all of these measures are the assumptions that test items are scored dichotomously, weighted equally, and summed to calculate observed scores. Recognizing the limitation of these assumptions, Livingston and Lewis (1995) suggested a broader method of estimating classification accuracy, applicable not only to tests like those described above, but also to tests that include partial credit scoring (e.g., essay tests or tests with free response items) and tests that have items (or subtests) of unequal weight. In order to apply Livingston & Lewis' (1995) method, four kinds of input are required: 1) the distribution of observed scores; 2) a reliability coefficient of the scores; 3) the minimum and maximum possible scores; and 4) the cut point(s) separating classifications/categories. From this input, decisions about classification accuracy and consistency can be derived. In addition, an estimate of the *effective test length* can also be derived, which represents the minimum number of items required to produce a total score of the same reliability.

Classification accuracy and consistency measures have been utilized in a multitude of ways over the years, including to simplify estimates of pass/fail classification accuracy (Breyer & Lewis, 1994); classify students with a modified Guttman scale (Schulz, Kolen, & Nicewander, 1999); base classification decisions on measurement decision theory for dichotomous data (Rudner, 2001, 2002, 2003), polytomous data (Rudner, 2005) and raw scores (Li & Sireci, 2005); obtain greater accuracy with small data sets (Guo, 2006); merge common classification accuracy measures with randomized response designs (Betebenner, Shang, Xiang, & Zhao, 2008); estimate the accuracy and consistency of complex assessments (W. Lee, Brennan, & Wan, 2009); derive measures of classification accuracy from measures of consistency (Newton, 2009) – or the problem with doing so (Bramley, 2010); estimate the classification accuracy of a single decision based on multiple measures (Douglas & Mislevy, 2010), and even why cut scores should not be employed at all (Dwyer, 1996).

2.2.3.1 CA Studies outside Language Testing

In addition to the applications cited above, there are several applications directly relevant to this dissertation, both within and outside second language testing. For example, Silva (1985) compared CTT measures of classification consistency to IRT measures of consistency for cut scores. Results show that the IRT indices were superior to the CTT indices. Hoffman and Wise (2000) also examined the accuracy of decisions near the cut score but for single administration exams. Claiming such exams “up the stakes” for classification accuracy, the authors defined error as the difference between true and observed scores on a single administration of the exam and stated that the classification accuracy of a true score can be determined by looking at the proportion of the conditional (normal) distribution of observed scores falling within the same category as the true score. On this basis, they examined the standard errors of measurement (SEMs) to determine classification accuracy within IRT. The primary conclusion is that classification-accuracy functions based on true scores vary considerably from those based on observed scores, which is very problematic on high-stakes tests.

Lee, Hanson, and Brennan (2000, 2002) also examined the classification accuracy and consistency of a single administration test but with multiple-category classifications. They found that the IRT-3PL model was superior to the two beta binomial models (2-parameter and 4-parameter) they examined, but they reiterated the now-common theme that in practice model choice should be based on model fit, assumption satisfaction, and computational feasibility.

Ercikan and Julian (2002) made several important clarifications in their study regarding classification accuracy, most notably that: 1) classification accuracy is a measure of the accuracy of decisions, not scores; 2) the level of classification accuracy will vary with changes in measurement accuracy across ability levels; and 3) the effect of measurement accuracy on classification accuracy will be most observable near cut scores, a claim suggested by several researchers (e.g., Hambleton & Slater, 1997; W. Lee et al., 2000; Livingston & Lewis, 1995; Schulz et al., 1999).

Further expounding on this latter point, Ercikan and Julian argued that the most common indicators of measurement accuracy (e.g., KR-20, Cronbach's α) are inappropriate for assessing the accuracy of proficiency scores because they provide an indication of overall measurement accuracy, not the measurement accuracy at cut scores. As an alternative, Ercikan and Julian advocated for the amount of measurement error near the cut scores to be used for the estimation of classification accuracy (based on the likelihood of misclassification errors). In their study, the authors examined the impact of the number of proficiency levels on classification accuracy near the cut scores (as a function of the measurement error). The authors found that, perhaps unsurprisingly, as the number of levels increased, the overall classification accuracy decreased. They therefore concluded that classification accuracy is sensitive to measurement accuracy, particularly when larger numbers of proficiency levels are included. Moreover, they stated that even though higher measurement accuracy tends to imply higher classification accuracy, higher reliability as indicated by, e.g., KR-20 does not imply higher classification accuracy. Therefore, in designing (or choosing among) tests, it is very important to examine the measurement accuracy provided by the test at cut-score points rather than relying on more common measures of classification accuracy.

Nystrom (2004) examined the classification accuracy of a Swedish national test in mathematics, trying to derive a single decision regarding mastery/non-mastery from two subtests previously scored independently (algebra and differential equations). The author found there was a significant reduction in classification accuracy when mastery was based on the composite score. Perhaps more importantly, Nystrom found that accuracy near the cut score depended on where the cut score fell along the ability-level distribution, indicating it would be highest near the middle ability level and far less accurate toward the extremes, which supports (and further clarifies) Ercikan and Julian's (2002) finding.

In a study very similar to the proposed dissertation, Stone, Weissman, and Lane (2005) examined the consistency of classifications based on competing IRT models using data from a state assessment program. Examining data collected from 13,621 eleventh-grade students who took a test comprising 60 multiple-choice items and four constructed-response items, the authors found the 3PL model a better fit of the data than the 1PL model, stating that there were significant and systematic differences between the 1PL and 3PL model classifications despite a high level of agreement between classifications (kappa, κ = 0.92). Stone et al concluded that this difference is particularly relevant when classifications are used for high-stakes purposes and indicated the importance of identifying the best-fit model for each data set.

Kalohn and Spray (1999) examined the effects of a misfitting IRT model on classification accuracy. In a simulation study of 623 items taken by a randomly generated sample of 2000 examinees, the 3PL model misfit no items, while the (Rasch) 1PL model misfit nearly half of the items, resulting in the elimination of nearly half the items for model fit analysis. This item elimination led to a substantial increase in false negative errors, false positive errors, and the percentage of misclassifications for the 1PL model. In other words, applying a misfitting model (in this case, the Rasch-1PL model) had a dramatic effect on the outcome of the classification accuracy of the test. The authors therefore posited a warning against the misuse of models:

> There has been a trend recently of some practitioners recommending the use of only one IRT model. A blanket recommendation of a particular model, regardless of fit, could have serious repercussions in the certification and licensure industry and may impact the protection of the general public. (p. 59)

Stated differently, it appears some practitioners have employed the models they know how to employ, whether or not there was sufficient rationale to justify the application, a problem that will be addressed in this dissertation.

2.2.3.2 CA Studies within Language Testing

As illustrated, there have been numerous classification-accuracy studies outside the field of second language testing (SLT). Within SLT, however, there appears to be only one study that has examined the impact of item-analysis method on classification consistency and accuracy (B. Zhang, 2010). In that study, the author examined the classification of 5000 examinees who took a large-scale language certification exam. Comparing the outcomes of CTT, IRT, Polytomous IRT, and TRT models, the author found that the TRT model fit the language-proficiency data best because of the clear violation of local independence among items (i.e., a strong testlet effect was present).

While Zhang's finding is important, it is noteworthy he did not explicitly investigate violations of the two assumptions of IRT – unidimensionality and local item independence. While he indirectly investigated LII by citing the presence of a strong testlet effect (using the criteria suggested in Bradlow, Wainer, and Wang (1999)), he apparently did not investigate the possibility of multidimensionality despite the fact it would appear likely given what is known about measures of general language proficiency. At the very least, a discussion of why the clear violation of LII was not problematic such that it still made sense to include IRT models in his comparison should have been included. Put another way, it is questionable whether standard unidimensional IRT models should have been included at all in his comparison. In fact, to the extent CTT modeling requires satisfaction of these assumptions as well (see, e.g., Henning et al., 1985), these issues should have been addressed but were not. As a result, they remain open questions with respect to the choice of item-analysis method in SLT.

2.2.4 Other IRT Applications within L2 Testing

Zhang's (2010) multi-model comparison notwithstanding, the majority of studies in second language testing that have examined measurement models have employed Rasch analysis, including many recent studies in the areas of pragmatics (e.g., Brown & Ahn, 2011), lexicon (e.g., Beglar, 2010), writing (e.g., di Gennaro, 2009), cut-score setting (e.g., Kozaki, 2010), rater judgment (e.g., Kim, 2009), oral discourse (e.g., Davis, 2009), and C-Test validation (e.g., Lee-Ellis, 2009). On the other hand, far fewer studies have employed IRT modeling despite their widespread use in other contexts.

One reason IRT modeling has not gained more footing in language testing may be due to the skepticism surrounding its appropriateness to the domain, as mentioned earlier. Choi (1989, 1992) and Choi & Bachman (1992), for example, highlighted several possible theoretical and practical issues associated with the use of IRT in language testing, the primary one being the assumption of unidimensionality. As Choi & Bachman (1992) pointed out, L2 research has shown on numerous occasions that language proficiency is likely a multidimensional construct (e.g., Bachman, 1982; Bachman & Palmer, 1981a, 1981b; Sang, Schmitz, Vollmer, Baumert, & Roeder, 1986). As such, it is equally likely many tests purportedly measuring language proficiency are multidimensional. Another concern pertains only to the Rasch model, which assumes equal discrimination among items and no guessing by examinees. Because it is likely unreasonable to assume all items on a test discriminate equally and that examinees will not guess at all, especially when there is no penalty for doing so, the legitimacy of these Rasch model assumptions has been called into question.

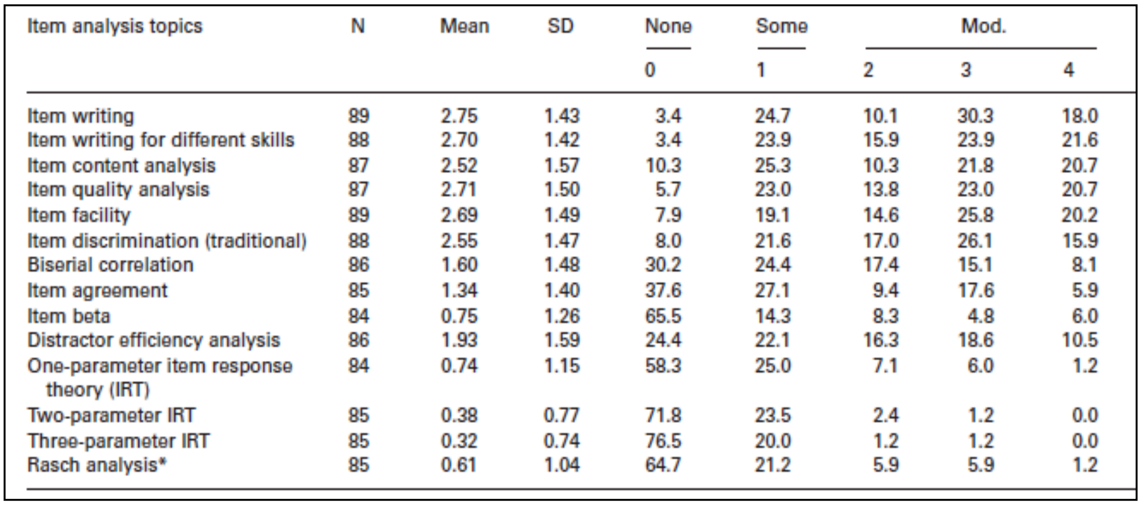
To test these assumptions, Choi & Bachman (1992) examined data collected from 1400 participants who took the reading comprehension sections of the First Certificate of English (FCE) and 1000 participants who took the reading comprehension section of the Test of English as a Foreign Language (TOEFL). Results regarding unidimensionality were conflicting, with the most stringent test (a factor analysis of inter-item tetrachoric correlations) indicating clear violations of the assumption, while the least stringent test indicating support for unidimensionality (greater than 20% of the variance accounted for by the first factor – see Reckase, 1979).

Results regarding the testing of Rasch-model assumptions were much clearer, with several measures indicating the IRT-2PL and -3PL models being superior fits of the data. Choi and Bachman (1992) therefore concluded the more sophisticated IRT models are more appropriate for language tests than the Rasch model, despite its widespread use in SLT (c.f., McNamara, 1990, 1991).

Another reason IRT may not have yet gained a foothold in SLT is merely because not enough second language testing specialists understand it conceptually and/or know how to apply it in practice. Evidence of this supposition can be inferred from the data in Brown and Bailey (2008). As shown in Table 3 below, CTT item analysis was taught in many language testing courses at the time of the survey (Item facility: mean = 2.69, and Item discrimination: mean = 2.55, on a 0-5 Likert scale), but IRT concepts were taught far less often (IRT and Rasch: means = 0.32-0.74, on the same 0-5 scale).

Stated differently, CTT item analysis was taught in 92% of the courses (100% - 8.0% in the column “None”), but IRT concepts were taught in only 24-42% of the courses (e.g., 100% - 76.5% in the column “None” for IRT 3PL). While Brown & Bailey explain this is a self-identified, self-selected sample and cannot be assumed to be a representative sample of all language testing courses, it is clear IRT is taught far less often and in much less detail than CTT, which could help explain the relative dearth of IRT studies in the SLT literature.

Table 3. Excerpt of table from Brown & Bailey (2008), illustrating topics taught in language-testing courses



Item analysis topics	N	Mean	SD	None	Some	Mod.		
				0	1	2	3	4
Item writing	89	2.75	1.43	3.4	24.7	10.1	30.3	18.0
Item writing for different skills	88	2.70	1.42	3.4	23.9	15.9	23.9	21.6
Item content analysis	87	2.52	1.57	10.3	25.3	10.3	21.8	20.7
Item quality analysis	87	2.71	1.50	5.7	23.0	13.8	23.0	20.7
Item facility	89	2.69	1.49	7.9	19.1	14.6	25.8	20.2
Item discrimination (traditional)	88	2.55	1.47	8.0	21.6	17.0	26.1	15.9
Biserial correlation	86	1.60	1.48	30.2	24.4	17.4	15.1	8.1
Item agreement	85	1.34	1.40	37.6	27.1	9.4	17.6	5.9
Item beta	84	0.75	1.26	65.5	14.3	8.3	4.8	6.0
Distractor efficiency analysis	86	1.93	1.59	24.4	22.1	16.3	18.6	10.5
One-parameter item response theory (IRT)	84	0.74	1.15	58.3	25.0	7.1	6.0	1.2
Two-parameter IRT	85	0.38	0.77	71.8	23.5	2.4	1.2	0.0
Three-parameter IRT	85	0.32	0.74	76.5	20.0	1.2	1.2	0.0
Rasch analysis*	85	0.61	1.04	64.7	21.2	5.9	5.9	1.2

The field of SLT is of course not alone in its lack of widespread adoption of IRT. As mentioned earlier, Streiner's (2010) discussion about the lack of IRT adoption within medical education highlights its still relative underutilization in many contexts (e.g., Edelen & Reeve, 2007; Hall, Reise, & Haviland, 2007; Prieto, Delgado, Perea, & Ladera, 2010; Reeve et al., 2007; Scherbaum et al., 2006; Unick & Stone, 2010). In a search for reasons, Streiner posited that it is a reflection of the fact that: 1) only a handful of graduate students are taught IRT, a finding corroborated by Brown and Bailey; 2) CTT is much easier to understand conceptually and its parameters are easier to calculate; and 3) IRT software has a steep learning curve and is expensive compared to common statistical packages that have means of calculating CTT statistics.

Despite this relative dearth of IRT studies in SLT, there are a few language researchers who have examined the IRT assumptions. For example, Matthews (1992) examined the local item independence of foreign language proficiency cloze tests and found they do exhibit sufficient LII to justify the use of IRT models despite the theoretical assumption cloze items are locally dependent (Bachman, 1982; Turner, 1989 - Sorry, no Overdrive). Lee (2004) alternatively examined the LII of a 40-item EFL reading comprehension test and found clear violation of LII among passage-related items but questioned whether violation of the assumption could not be absorbed by the test as a whole. Referring to Reckase, Ackerman, and Carlson (1988), Lee noted that even if the ability trait putatively being measured is not necessarily a single trait, it could be that the composite of abilities being measured across items is sufficiently unidimensional such that application of IRT models is reasonable (p. 79), the same conclusion reached by other researchers (e.g., Henning, 1992).

With respect to unidimensionality, Yen and Walker (2007) examined the dimensionality of composite scores on an English language proficiency test. In an analysis of data from 12,008 elementary school students who took a State English Proficiency Assessment (SEPA) as part of the No Child Left Behind Act, the authors examined the dimensionality of *oral language proficiency* as measured by the oral and listening comprehension sections of the SEPA. Results indicate the MIRT model best fit the data of the upper elementary students but the UIRT model best fit the lower elementary students, at least with respect to the χ^2-difference test. These mixed results led Yen and Walker (2007) to conclude it is necessary to conduct simulation studies to examine further this discrepancy in model fit. Moreover, the authors stated these findings may reflect the hypothesis put forth by Davey and Hirsch (1990), who stated that MIRT models may be less able to discriminate examinees on tests with fewer than 100 items because of the increase in parameterization. In other words, the increase in model complexity dramatically increased the need for both larger sample populations and item counts in order to discriminate among examinees more effectively.

von Davier (2008) also examined dimensionality, proposing a general diagnostic model (GDM) to examine multidimensional data collected from the reading and listening sections of the TOEFL Internet-Based Test (iBT). Analyzing the model fit of the unidimensional IRT-2PL model, the two-dimensional IRT 2PL, and an eight-skill GMD (among other things), von Davier (2008) found the unidimensional model to fit nearly as well as the multidimensional model and better than the far more complex eight-skill GDM, therefore reiterating the need for considerably more data to accommodate the increase in parameters of the more complex models.

In another examination of the dimensionality of the TOEFL iBT, Sawaki, Stricker, and Oranje (2009b) found that a single higher-order factor model fit the sample data well, thereby providing support for the case of a unidimensional language proficiency measure. However, this finding is tempered by the fact the authors found the model with a higher-order factor and four first-order factors to be the best-fitting model. This finding led the authors to conclude that not only should composite test scores be reported, but also section sub-scores.

Aside from those studies testing the assumptions of IRT, there are a few other studies relevant to the focus of this dissertation. One is Henning (1984), who compared CTT and the Rasch-1PL model to demonstrate the advantages of latent-trait (IRT) models over CTT models.[9] Using a sample population of 108 university applicants who scored lower than 500 on the TOEFL, a 48-item reading comprehension test containing eight passages with six 4-option multiple-choice items each was administered. Analysis of the data yielded the finding that the KR-20 and KR-21 values were higher for the Rasch model than CTT, thereby supporting Henning's thesis that latent-trait models are superior to CTT analysis. Unfortunately, Henning did not address the fact that local item independence was probably violated (i.e., there was a testlet effect), so it is unclear what role, if any, violation of this assumption should have played in Henning's analysis.

Another interesting finding resulting from the Rasch analysis is that the responses of 14 participants misfit the model, leading Henning to proclaim these test scores should not be used for decision-making purposes. While excluding such data from analysis for research purposes is reasonable, an interesting ethical question arises when the scores are

[9] Interestingly, Henning acknowledged that his choice of Rasch over the IRT-2PL and -3PL models was made for practical reasons: it requires fewer data, does not require software, and does not try to identify systematicity in measurement error (p. 125).

the result of a high-stakes test where retesting or discarding data is not an option. That is, how should administrators deal with scores around the cut that misfit the chosen model?

Perkins and Miller (1984) also conducted a CTT-Rasch comparison study using the same 48-item reading comprehension test to determine whether CTT and Rasch would identify the same misfitting items. Interestingly, each model found a very different set of items, indicating examinee scores likely would vary depending on which item-analysis method were employed. This is another important finding regarding this focus of this dissertation.

De Jong and Glas (1987) also employed the Rasch-1PL model to examine the construct validity of a Dutch standardized national test of foreign language listening comprehension and found it was indeed valid. Boldly, the authors claimed Rasch (1960) "proved" that the Rasch model is the only valid model for dichotomous data when summing scores (p. 170), based on the fact that summed scores imply unidimensionality, and that the Rasch-1PL model is appropriate for similar language proficiency tests because foreign-language proficiency can be measured along a single dimension (p. 191). This latter statement is of particular interest to the focus of this dissertation because of the ongoing question of whether language proficiency tests are by design multidimensional or whether they can be treated as "essentially unidimensional" (Stout, 1987) or "psychometrically unidimensional" (Henning, 1992) for IRT modeling purposes.

A fourth relevant study was conducted by McCall (2002), who conducted an IRT analysis of a multiple-choice reading-comprehension test. Unlike most other language testing researchers, McCall actually examined model-fit for all three IRT models and examined the assumption of unidimensionality using factor analysis. Perhaps unsurprisingly, McCall found the 3PL model fit the data best and that the test was unidimensional because items clustered on the passages rather than the content. However, this latter conclusion is a misinterpretation of the findings in that the clustering of items on passages, while not a violation of unidimensionality, would be a violation of local item independence, the other key assumption required of standard IRT modeling. Therefore, the question of whether LII was violated remains unanswered and should have been considered in the analysis.

Finally, Carr (2006) examined the impact of reading-passage task characteristics on examinee performance. Using data obtained from a random sample of 9,000 participants who took the reading-comprehension section of the TOEFL, Carr found that the 3PL model best fit the data, both in terms of log-likelihood values and the least number of misfitting items. Interestingly, however, Carr found no significant relationship between guessing by low-ability examinees and empirically salient characteristics of passages and key sentences, which could be interpreted to mean guessing is not an important parameter to measure on reading-comprehension tests given the complexity it adds to the IRT model, an issue to be investigated in this dissertation.

The remainder of this dissertation is organized as follows: Chapters 3, 4, and 5 contain the results of three sets of data analysis – one with the full data set and two with resampled data sets – to see whether the outcomes of the full-data analysis hold across different population samples. Chapter 6 is a discussion of the synthesized findings from the three sets of analysis, and Chapter 7 is a discussion of the implications of the findings with respect to the ILTA Guidelines for Practice and the language-testing field at large.

3 Data Analysis I (Full Data Set)

3.1 Participants

The participants in this study were 2,320 high school students applying to a prestigious private university in western Japan. Almost all applicants attended mainstream or "academic" (college-preparatory) high schools, and most fell within the 82nd percentile in terms of relative academic ability among all Japanese high school students at the time of admission (2006). This restricted range in aptitude is a function of the university entrance system in Japan, where students can take numerous practice exams that are comprised of former exam items. Using publicly available data, college-prep institutes ("cram schools") create, administer, and score these practice exams for would-be students, thereby helping them determine the university departments to which they are most likely to be admitted.

3.2 Methods

3.2.1 Japanese University Admissions Process

The university admissions process in Japan involves a complex series of steps that starts literally years before a student takes a university department examination like the one being examined in this study. After the sixth grade, students who wish to attend a selective public or private middle school must take a school-specific exam that helps determine which track they will follow through their middle-school years. For example, if they exhibit aptitude in the humanities, they will take the H track, in sciences the S track, and so forth. For all other students, they attend a local middle school.

Following middle school, schooling splits into three tiers: elite academic, mainstream academic, and vocational. For those who will go on to university (those who attend elite academic high schools and some who attend mainstream academic high schools), university entrance exams follow. Public universities, considered more prestigious than private universities on the whole, require applicants to take a standard national exam. Created in the late-1980s and endorsed by the Japanese Ministry of Education (MOE), the exam comprises a series of constructed-response items developed by a revolving committee of college faculty. To ensure equity from year to year, the format and content of the exam are standardized and rarely change, with any proposed changes requiring a forewarning of at least two years. In total, there are exams in 34 subjects, and each public university and its departments require a particular combination of subjects depending on their academic focus.

Because of the extremely rigorous nature of the public national exam, only those most academically gifted will take it. To attend private universities, on the other hand, examinees are required to take only the specific university department's exam for which they strive to gain admission. In this particular case, the data under investigation were generated from the Form-H exam of the policy studies department at a prestigious private university in western Japan.

3.2.2 Form H of the Entrance Exam

For the policy studies department of this particular university, several forms of the entrance examination are administered depending on the characteristics of the applicant. For example, Form F is administered to legacy applicants, Form R to graduates of foreign high schools, Form S to applicants in the early admissions process, Forms T to athletes, and so on. The form under investigation in this study is Form H (Appendix A), which is the most competitive form because it is open to any high-school graduate and therefore has the largest candidate pool. Designed to be an exam of general English language proficiency, Form H is comprised of 70 dichotomously scored multiple-choice items grouped into four sections:

Part I: 15 reading-comprehension items based on a single reading passage

Part II: 20 cloze-test items based on a single reading passage

Part III: 15 synonym-choice items based on a single reading passage

Part IV: 20 error-identification items based on 20 unrelated sentences

As mentioned in the literature review, each of the forms is administered only once per year without any pilot testing and then made part of the public domain after scores have been calculated and admission decisions are complete.

The creation of a new version of Form H is a highly structured event with a long tradition and is the product of many hours of test construction and internal moderation by a team of sequestered foreign-language specialists. Every summer, around 20 faculty members from the various departments within the school gather to evaluate the tests created by each department and to make changes as necessary. Because consensus on the correct answer is reached for every item on every test, there is the implicit belief all items are fair, reliable, valid measures of examinee ability. As a result, scoring is based on the number-correct method, where each correct response is given a value of 1 and the total score is the sum of all correct responses. Student scores are then ordered from highest scoring to lowest scoring, with a maximum score of 70 and a minimum score of 0. Offers of acceptance are based on this ordinal ranking of student scores.

No test specifications exist at the item level for Form H. Test designers follow an approximate blueprint by reviewing previous versions of the exam and for the most part attempt to replicate the format used in earlier versions. After a process involving many rounds of item drafting, moderation, and internal critique, the 2005 version of Form H, the one used to collect data for this study, was administered in a single administration. Because justification for this "moderation model" of test development is based on the belief that expert opinion and careful editing of test content is sufficient to create items that can measure candidates' relative abilities, no pretesting was performed and no item analysis post administration was conducted. This 'pre-scientific' approach to language testing (Spolsky, 1978, 1981) predates both modern psychometric methods and the formulation of the ILTA Guidelines for Practice, yet arguably remains the most commonly used model around the world in language testing.

3.2.3 Calculation of Examinee Scores

After confirmation the scoring key contains no clerical errors, examinee scores are calculated by summing the total number of items answered correctly by each examinee. The validity of this summative score is predicated on one very strong assumption: the moderation process used to create the exam ensures only one answer per item is correct, but the distracters contained in each item are plausible options that can help discriminate among examinees of varying ability. Table 4 displays the descriptive statistics of the examinee scores for this particular administration of Form H.

Table 4. Descriptive statistics of the raw score data

N	Min	Max	Mean	Std Dev	Skewness		Kurtosis	
					Statistic	SE	Statistic	SE
2320	5	65	37.303	8.254	-.064	.051	-.080	.102

As shown, examinee scores range from 5 to 65, with 70 being the maximum attainable score. The mean score is 37.303, and as shown in Figure 2, the distribution of scores approximates a normal distribution, although the distribution skews slightly to the left (Skewness = -.064) and is slightly platykurtic (Kurtosis = -.080). Given its near-normal distribution, it can be estimated about two-thirds of examinee scores fall between the raw scores of 29 and 46 (± 1 s.d. from the mean) while 95% of the scores fall between 21 and 54 (± 2 s.d. from the mean).

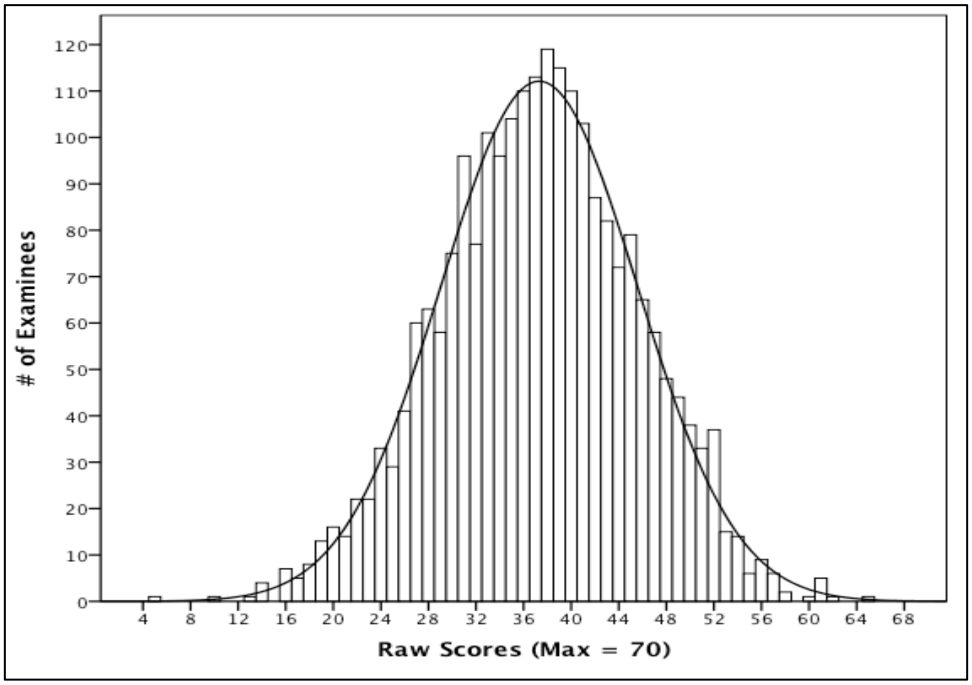


Figure 2. Distribution of examinee raw scores

3.2.4 Determination of the Cut Score

Recall that the cut score for this exam is a function not of an attained level of mastery but of a quota set by the Japanese Ministry of Education. In other words, the cut score is not criterion referenced and can fall anywhere along the distribution of scores depending on the target number of enrollees for the given year.

Derived from a complex formula containing numerous variables, the quota has its basis in the belief that the quality of education is at least in part influenced by the physical environment: for every student, there has to be a certain number of books in the library, a certain number of faculty on staff, a certain number of seats in the cafeteria, and so forth. The number of examinees offered admission in any given year (λ) can be derived from the following formula:

$$\lambda = \frac{Q}{\hat{r}}$$

where Q = the MOE quota for the year and $\hat{r}$ = the projected enrollment rate based on historical data. For both regulatory and financial reasons, it is very important universities accurately determine the cut score that will yield the desired number of enrollees: if too few examinees enroll, they will lose money (due to less tuition), but if too many enroll, they face potential penalties, including the loss of government funding or even loss of accreditation if quotas are exceeded routinely.

The Form-H quota for the policy studies department of this particular university was 150 in 2006, and 37% of applicants offered admission to this department historically have accepted. As a result, the target number of admissions offers for this particular year was 405 ($\lambda = \frac{150}{0.37}$). To determine the 405th highest score, summative scores obtained from the actual administration of the test were compiled in Excel and sorted from highest to lowest. The resulting ranking indicates that the 405th highest-ranking examinee achieved a raw score of 45, a score achieved or exceeded by the top 462 examinees. The next higher cut score, 46, was achieved by 385 candidates, 20 fewer than the target number of admission offers. Because the raw number-correct method has no inherent means of breaking ties, a decision had to be made whether to set the cut score at 45 or 46 for the purpose of this analysis. Based on the historic pattern of admissions offers at this university, it is assumed here 462 examinees were offered admission rather than the fewer number of 385 (Ross, 2011). Figure 3 portrays the location of the cut score along the distribution of examinee raw scores.

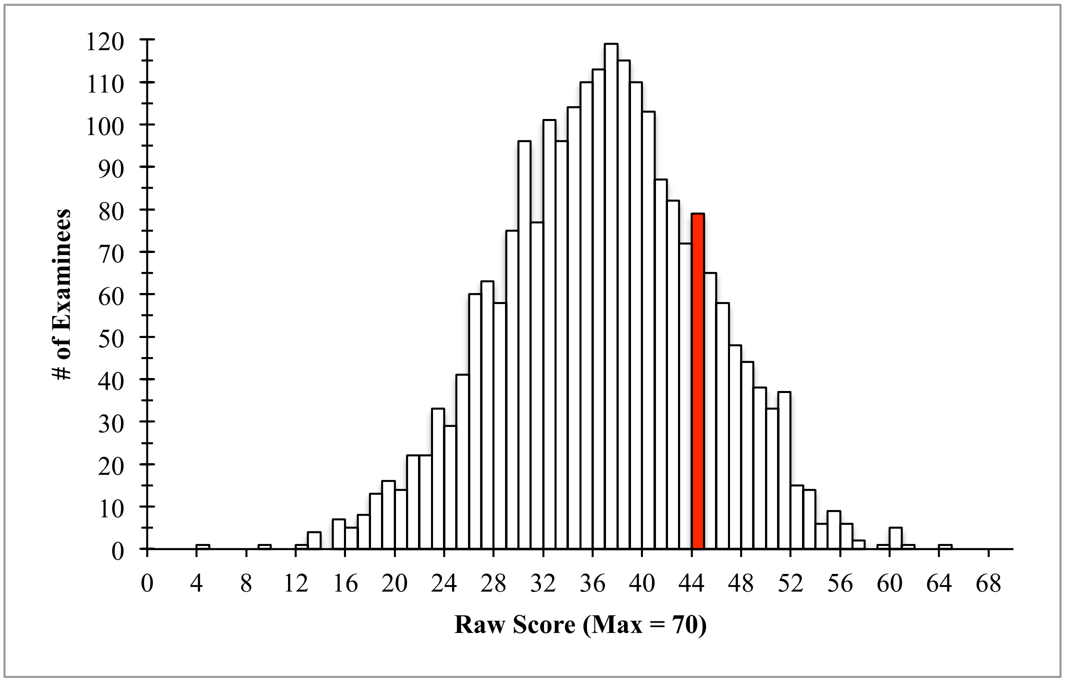


Figure 3. Cut score (45) along the distribution of examinee raw scores

3.3 Problems with the Raw-Score Method

One practical problem with using raw scores for admissions decisions is the lack of sufficient granularity in the scoring scale. As explained, 462 candidates scored at or above 45. Because there is no principled means of deciding who among those that scored 45 should be offered admission so the ideal number of offers can be made, the admissions committee is forced to choose between moving the cut score higher (to 46), which would likely result in a lower-than-target enrollment, or keeping the cut score at 45, which would increase the risk of having a larger-than-desired number of students enroll.

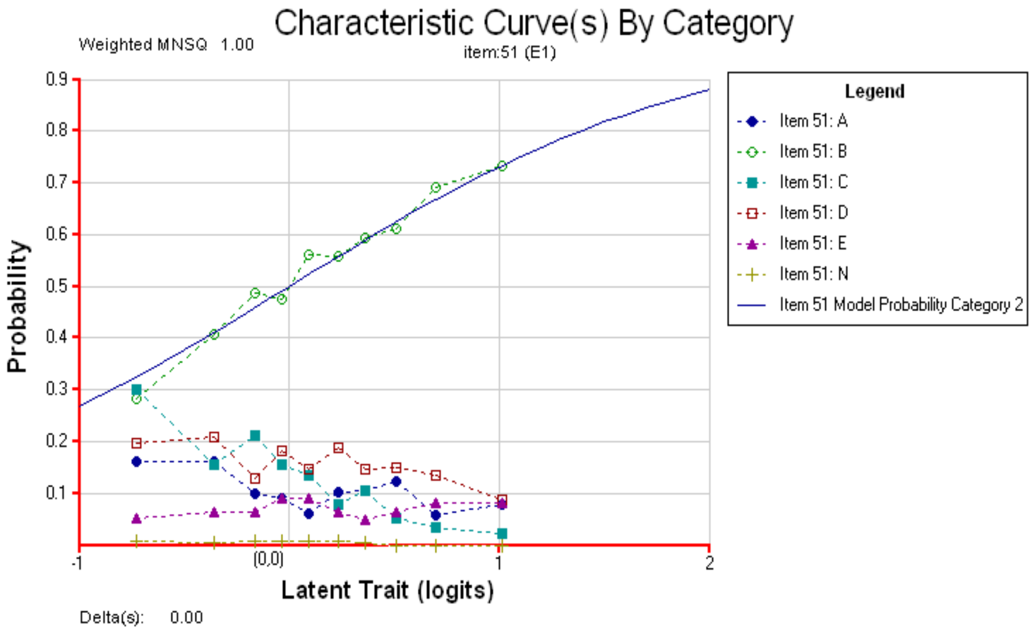
Practical concerns aside, the use of raw scores alone, particularly when compiled from a non-piloted version of the exam, contravenes the ILTA Guidelines for Practice, which mandate either pretesting or post-hoc item analysis be conducted to ensure malfunctioning items are excluded from scoring. In this particular case, there is no confirmation the test-moderation process has been successful despite the time and effort spent creating the exam. Instead, it is taken on faith alone all items are fair estimates of examinee ability. Because this is likely an unreasonable assumption, some form of post-hoc item analysis needs to be conducted to ensure all items are contributing appropriately to the total scores.

Unfortunately, the test moderation panel for Form H involves different combinations of faculty members from year to year, so who could/should conduct this item analysis is unclear. Perhaps more troubling is the fact that many panel members have literature or linguistics backgrounds and often do not consider item analysis to be of value (Ross, 2011). Regardless, conducting such an analysis is critical for ensuring the scores used in making admissions decisions are justifiable, as espoused by ILTA and other international language-testing bodies.

To illustrate the concern of relying solely on the rigor of test development as sufficient for test validation, Figures 4 to 6 illustrate how differently items on Form H are functioning with this set of participants.[10] Item 51, shown in Figure 4, portrays an ideally functioning item. As candidate ability increases (along the horizontal axis), the probability of selecting the keyed option (B) increases (along the vertical axis).

[10] The item characteristic curves (ICCs) portrayed in Figures 4-6 were generated using WINSTEPS as part of the item analysis delineated later in Chapter 3. The ICCs are shown here merely to demonstrate how different items are functioning on the test, thereby demonstrating the need for item analysis. Full item analysis follows in subsequent sections.

Conversely, as ability increases, the probability of selecting one of the distracters decreases, approaching 0 at the highest levels of ability.

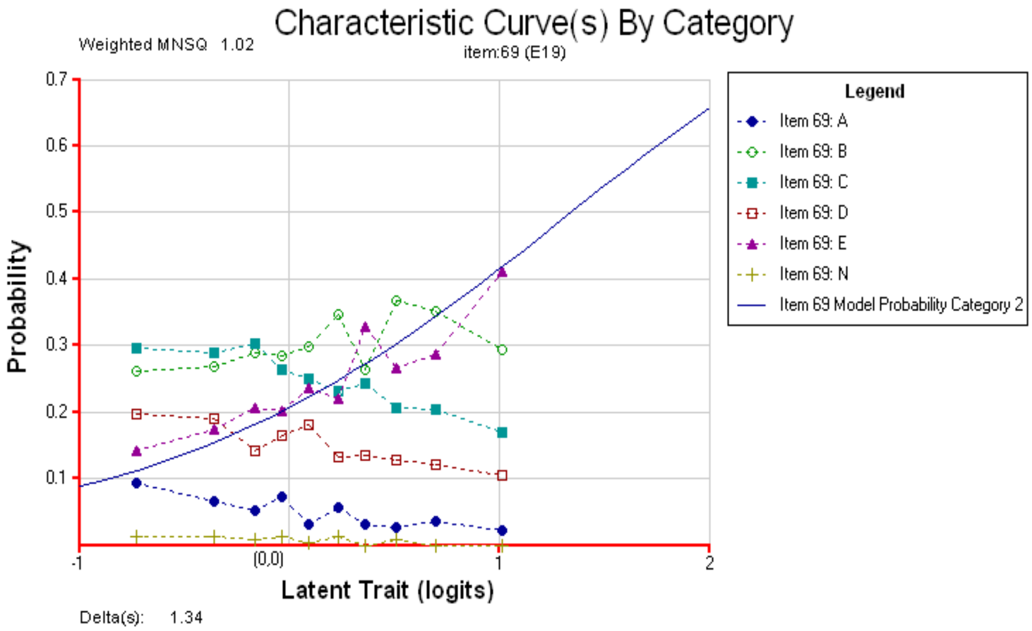


51. The (a) term automobile is commonly (b) applies to a four-wheeled vehicle designed (c) to carry two to six passengers and a limited amount of cargo, as (d) contrasted with a truck. (e) no error

Figure 4. Item that is functioning well

Under the moderation model of test development, all items are expected to function like Item 51. Unfortunately, this is not the case. Despite the extensive moderation of Form H, it appears to contain several malfunctioning items, as suggested by its mediocre internal consistency estimate (KR-20 = 0.78).

Item 69, shown in Figure 5, is illustrative of one of the malfunctioning items that likely contributes to the marginal level of internal consistency of Form H.

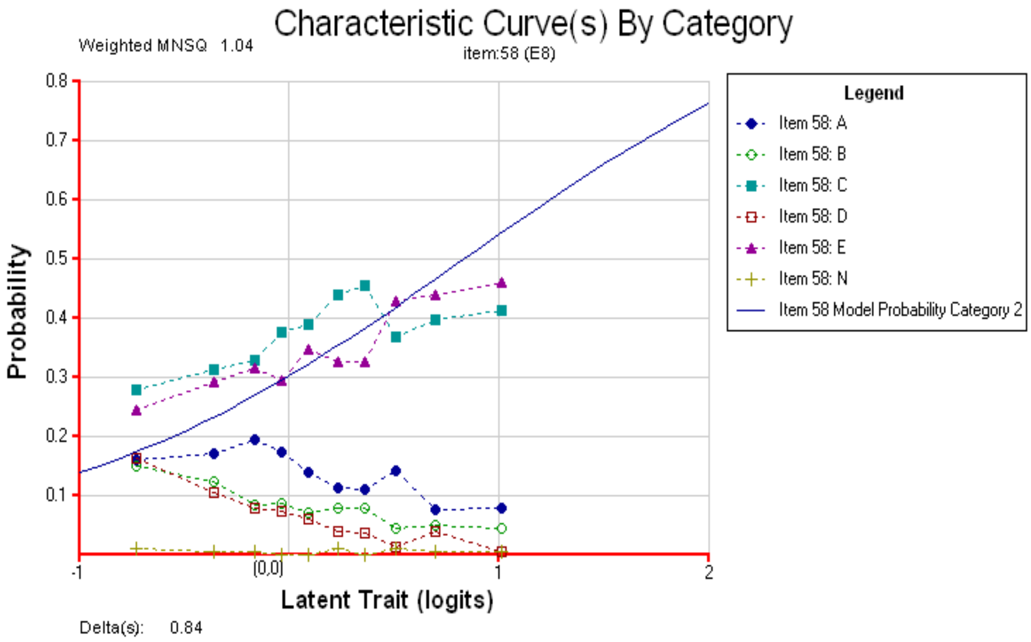


69. The sun is the (a) center of the solar system (b) with nine planets (c) revolving around (d) it. (e) no error

Figure 5. Item not differentiating candidates by ability

As shown, Item 69 does not sufficiently differentiate examinee ability levels. At higher levels of ability (e.g., logit = 1.0), there is only a 40% probability examinees will choose the keyed response (E), while candidates at the lowest level of ability exhibit a 15% probability of choosing the keyed response. Moreover, the probability of choosing one of the distracters approaches 0 only for Option A and in fact even increases in the case of Option B, further indicating this item is not discriminating among ability levels well/properly.

Another type of malfunctioning item is one that appears to have two correct answers (a double-keyed item). In such cases, the test-moderation panel might not detect ambiguity in an option that is designed to be a distracter but actually could be an alternate correct response. Figure 6 illustrates one example of this type of item.



(58) If you are (a) doing the laundry, you (b) should try to wash white things and bright colored things (c) separate, or the colors might (d) ruin the white clothes. (e) no error

Figure 6. Item with two apparent correct responses (a double-keyed item)

As shown, three of the distracters in Item 58 are functioning well, with minimal likelihood of being chosen at higher levels of ability. One of the distracters (E), however, attracted a subset of candidates with a similar ability range as the candidates who selected the keyed option (C). Closer inspection of the item reveals that Option C is indeed the only correct answer from a prescriptive grammar standpoint, given that adverbs – not adjectives – modify verbs (i.e., the correct construction is *"…you should try to wash white things and bright colored things separately."*). However, many examinees who experienced more naturalistic language acquisition (e.g., while living overseas in a predominantly English-speaking environment) may have noticed/learned that native English speakers often permit adjectives to function as adverbs, particularly when they are split from the verb. Conversely, candidates taught English grammar more explicitly and/or those who have had a lot of test-prep training are more likely to be aware of such subtle distinctions and thus better prepared to choose the (prescriptively) correct answer.

The items portrayed in Figures 5 and 6 are only a sample of the items that misfit the models, but they do illustrate the problem of not conducting item analysis as required by the ILTA Guidelines. Following is an introduction to some alternatives to the raw-score method that could be applied to these data to improve the validity of the test scores.

3.4 Alternatives to the Raw-Score Method

Given that item analysis is necessary to validate all test items, there are numerous item-analysis methods in existence from which to choose. For the purposes of comparing such scoring methods and their implications, the raw-score method currently used by the university is used here as the baseline. The objective of the following comparisons is to determine how the ordinal ranking of candidates might vary depending on the method of item analysis applied to the data. In other words, the objective is to determine the magnitude of examinee displacement in rank order with respect to the status quo.

3.4.1 CTT (Kelley's Discrimination Index)

As described in the literature review, one of the earliest forms of item analysis was created by (Kelley, 1939), who developed an item-discrimination index (D) based on the argument that item analysis is maximally effective when comparing the proportion-correct scores between groups comprising the upper and lower 27% of the distribution, but only when the proportion correct (*p*) is approximately 50% (pp. 23-4).[11] Following

[11] Although a lot of researchers use the upper and lower 27% groups by default, Kelley (1939) clearly indicates that 27% is ideal only when items are scored "in graduated amounts" and/or when item reliability

these guidelines, an item analysis was conducted on all 70 items of Form H. Table 5 contains a sample of the items analyzed using Kelley's discrimination index.[12]

Table 5. Excerpt of item analysis based on Kelly's D

Item	Upper Correct (N=616)	Lower Correct (N=646)	Total Correct (N=2320)	p	D	Flag
1	486	398	1592	69%	17%	F
2	391	151	958	41%	40%	
3	352	147	903	39%	34%	
4	355	169	930	40%	31%	
5	320	210	938	40%	19%	F
6	398	156	979	42%	40%	
7	495	289	1478	64%	36%	
8	432	152	1013	44%	47%	
9	443	212	1198	52%	39%	
10	346	190	920	40%	27%	
11	513	262	1457	63%	43%	
12	484	286	1405	61%	34%	
13	347	159	829	36%	32%	
14	273	139	743	32%	23%	
15	444	262	1271	55%	32%	
16	309	217	923	40%	17%	F
17	164	60	352	15%	17%	F
18	319	227	1001	43%	17%	F
19	562	401	1755	76%	29%	
20	487	285	1444	62%	35%	

As shown in Table 5, the second column contains the number of upper-group examinees who answered the item correctly (out of 616: 580 + 36 ties at the lower bound), the third column contains the number of lower-group examinees who answered the item correctly (out of 646: 580 + 66 ties at the upper bound), and the fourth column contains all those who answered the item correctly (out of all 2,320 examinees). The fifth column represents the item difficulty (*p*) and the sixth column represents the discrimination (D) between the upper and lower groups. Items were flagged when they exhibited low

is very low, and that 25% is ideal when items are scored dichotomously and item reliability is fairly high (p. 23). In this case, all items were scored dichotomously and reliability is reasonably high, so the upper and lower quartiles were used for comparison groups rather than the more commonly cited 27%.

[12] Only a subsection of each table is shown in the main text for readability purposes and to illustrate the how each method flags items. Full tables of each item analysis can be found in Appendices B-D.

discrimination values (D ≤ 0.20) and/or when items were excessively easy or difficult ($p \leq 0.20$ or $p \geq 0.80$). Using these criteria, five of the first 20 items (Items 1, 5, 16, 17, and 18) and 19 of the 70 items overall were flagged (see Appendix B.1).

3.4.2 CTT (Point-Biserial Correlation)

Another form of item analysis within CTT involves correlating dichotomous item responses to the total raw score to yield a point-biserial correlation. While there is no firm basis for a point-biserial correlation cutoff, widespread convention within language testing suggests an item with a correlation less than 0.20 is faulty. Using this $r_{pb} < .20$ criterion, 15 of the 70 items were flagged as faulty. Table 6 illustrates a sample of the items with low point-biserial correlations, and Appendix B.2 contains the complete table.

Table 6. Excerpt of item analysis based on point-biserial correlations

Item	r_{pb}	Flag
1	0.18	F
2	0.34	
3	0.30	
4	0.26	
5	0.17	F
6	0.33	
7	0.32	
8	0.38	
9	0.32	
10	0.21	
11	0.38	
12	0.30	
13	0.28	
14	0.21	
15	0.26	
16	0.17	F
17	0.21	
18	0.15	F
19	0.31	
20	0.31	

3.4.3 Rasch/1PL

Item analysis using the IRT 1-parameter logistic (1PL) model – or more specifically, the Rasch equivalent of the IRT-1PL model – was also conducted. Using the software WINSTEPS, information about all 70 items was captured, with a sample of that information displayed in Table 7.

Table 7. Excerpt of item analysis for the Rasch/1PL model

		Infit		Exact Match		
Item	b	MnSq	Zstd	Obs%	Exp%	Flag
1	-0.84	1.04	2.2	68.2	69.6	F
2	0.37	0.95	-3.7	65.8	63.0	
3	0.48	0.98	-1.7	65.0	64.3	
4	0.43	1.00	0.0	64.0	63.6	
5	0.41	1.05	3.8	61.6	63.4	F
6	0.33	0.96	-3.2	65.6	62.6	
7	-0.61	0.97	-2.2	67.2	65.9	
8	0.27	0.93	-5.9	66.9	62.0	
9	-0.08	0.97	-3.1	63.8	60.9	
10	0.45	1.03	2.1	62.8	63.8	F
11	-0.57	0.93	-4.8	68.4	65.3	
12	-0.47	0.98	-1.4	64.4	64.0	
13	0.62	0.98	-1.2	69.6	66.2	
14	0.80	1.02	1.3	69.1	69.0	
15	-0.21	1.00	0.0	62.2	61.5	
16	0.44	1.06	4.2	60.8	63.8	F
17	1.83	0.99	-0.3	84.8	84.8	
18	0.29	1.07	5.5	57.1	62.2	F
19	-1.22	0.96	-1.5	76.6	75.9	
20	-0.54	0.97	-2.0	66.8	65.0	

The second column of Table 7 contains the difficulty (b) estimates of each item as measured in Rasch logits, a unit of measure on an interval-level scale with its mean centered on 0 and a standard deviation of 1. As shown, Item 1 has a difficulty estimate of -0.84 logits, indicating it is relatively easy in that it is nearly 1 standard deviation below the mean level of difficulty of all items. Put another way, it indicates that an examinee with an ability of -0.84 logits has a 50% probability of answering the item correctly. Item 17 on the other hand is considerably more challenging, where an ability of 1.83 logits is required to have a 50% probability of answering that item correctly.

The third column contains the information-weighted fit (infit) mean-square (MnSq) measures of each item, where the mean square represents the randomness of the fit, with an expected value of 1. A value less than 1 indicates the item-response pattern is overly predictable (overfitting), while a value greater than 1 indicates the response pattern contains excessive randomness (underfitting).[13] The Z-std values in the fourth column are the standardized mean-square statistics, with an expected value of 0. It is these values that are the basis of flagging items, where any value greater than 2.0 is deemed faulty, following guidelines suggested by (Linacre, 2002).[14] Using this criterion, 18 of the 70 items were flagged. See Appendix B.3 for a complete list of flagged items.

Other item information included in Table 7 are the observed and expected percentages of examinee performance, where the observed percentage is equal to the proportion of examinees who answered the item correctly and the expected percentage is equal to the proportion of examinees who are expected to answer the item correctly given its level of difficulty (b). Note that the observed percentage of examinees is always lower than the expected percentage when Z-std ≥ 2.0, another indication the item-response patterns of these items underfit the model.

Figure 7 depicts the information the exam is providing at each level of theta. Known as the Test Information Function (TIF), maximum information for the Rasch/1PL model is 14.393, achieved at $\theta = -0.050$. Note that the term *information* is used in IRT to describe the level of precision/reliability at any given level of theta.

[13] Mean-square values are chi-square statistics divided by their degrees of freedom (Linacre, 2002).

[14] Items with z-std values ≤ 2.0 were ignored because they overfit the model, suggesting there may be other dimensions constraining the response pattern, which is of little concern with respect to Rasch's confirmatory model.

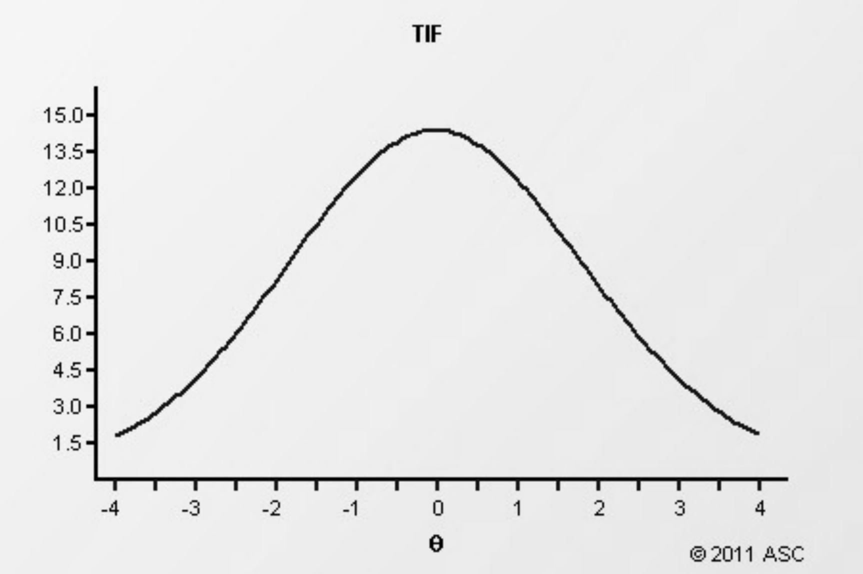


Figure 7. Test Information Function of the Rasch model

As shown, the TIF of the Rasch/1PL model has a normal distribution with its mean centered at/near $\theta = 0$. Note that the amount of information (Y-axis) at each level of theta (X-axis) decreases as the value of theta moves away from 0, with the least amount of information available in the tails of the distribution (e.g., info = 1.5 at ± 4.0 theta).

On the surface, this TIF appears to suggest the Rasch/1PL model fits the data nearly perfectly. However, this symmetry is merely a byproduct of Rasch parameters that force the difficulty parameter to a mean of 0 while holding other parameters (discrimination and guessing) invariant. Recall that the Rasch model forces the data to fit the model, unlike the 2PL and 3PL models, which adjust to fit the data. As such, *all* Rasch TIFs look like the one in Figure 7. Nevertheless, the concept is introduced here to serve as a means of comparison with the TIFs of the 2PL and 3PL models that are introduced later, as well as to orient the reader to the additional information gained over CTT item analysis when employing IRT modeling.

To illustrate the difference in levels of item information available between application of CTT item-analysis methods and IRT modeling, Figure 8 portrays what a CTT TIF would look like were it graphed.

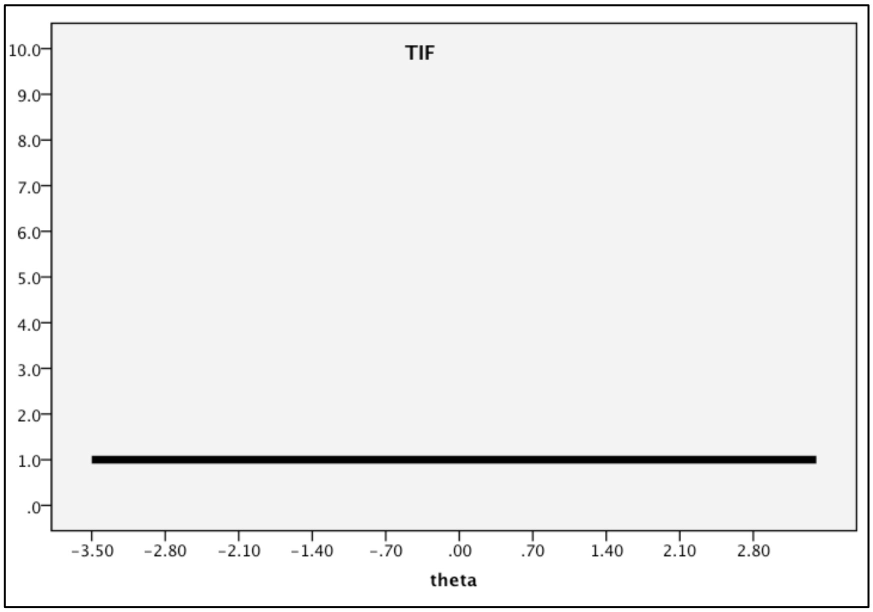


Figure 8. Test Information Function of CTT analyses

What should be immediately evident from Figure 8 is the lack of item information available when CTT analyses are employed. Rather, there is information, but it is the same at every level of theta (denoted by the solid horizontal line because CTT analysis occurs at the test level rather than at the item level). In other words, information is averaged across all levels of theta, thereby obviating the need to visually depict the item information generated by CTT item analysis. It is exhibited here only as a means of comparison with IRT information curves/functions.

Returning to the information available with the Rasch/1PL model, Figure 9 is a graph of the Conditional Standard Error of Measurement (CSEM) Function, which is the inverse of the Test Information Function and estimates the amount of error in theta estimation at each level of theta. Because the CSEM is the inverse of the TIF, the minimum CSEM is always located at the same point on the distribution as is maximum information, which in this case is at θ = -0.050.

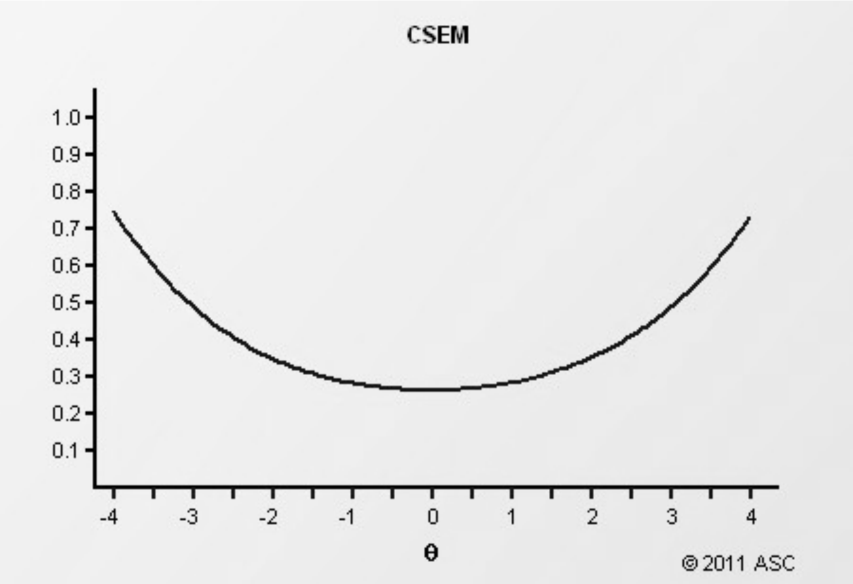


Figure 9. CSEM of the Rasch model

Figure 10 is a visual depiction of the Test Response Function (TRF) for the Rasch/1PL model. The TRF is the sum of all individual item response functions,[15] and its corresponding Test Response Curve (TRC) illustrates the proportion (left Y-axis) or number (right Y-axis) of items that examinees are predicted to answer correctly at each level of theta.

[15] Assuming all items are locally independent, meaning the response to any one item does not influence the response of any other item.

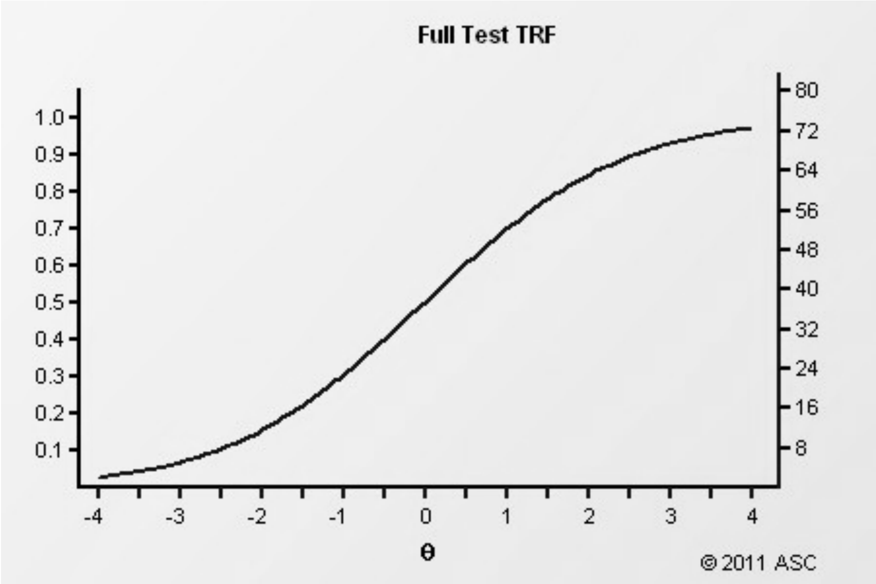


Figure 10. Test Response Function of the Rasch model

As shown, low-ability examinees ($\theta = -4.0$) are predicted to answer almost no items correctly while examinees of the highest ability ($\theta = 4.0$) are predicted to answer nearly 100% of the items correctly.

3.4.4 IRT 2PL

The next item analysis conducted was application of the IRT 2-parameter logistic (2PL) model to the data set. Table 8 contains an excerpt of the resulting item analysis when applying the 2PL model to the data using Xcalibre (Version 4.1).[16]

[16] The options for input and output specifications within Xcalibre are described in Appendix E.

Table 8. Excerpt of item analysis for the 2PL model

Item	a	b	Flag
1	0.242	-1.947	F
2	0.472	0.475	
3	0.39	0.725	
4	0.313	0.778	
5	0.208	1.091	F
6	0.439	0.449	
7	0.437	-0.845	
8	0.534	0.305	
9	0.423	-0.121	
10	0.254	0.979	F
11	0.564	-0.651	
12	0.394	-0.714	
13	0.371	0.978	
14	0.269	1.675	
15	0.328	-0.384	
16	0.192	1.264	F
17	0.373	2.852	
18	0.176	0.899	F
19	0.482	-1.550	
20	0.413	-0.790	

Columns 2 and 3 of Table 8 contain the discrimination (a) and difficulty (b) parameter estimates of each item resulting from application of the IRT-2PL model to the data. The a-parameter for this model has a mean of 0.365 and standard deviation of 0.137, while the model forced the b-parameter to center its mean at 0 with a standard deviation of 1. Items were flagged if either value exceeded the acceptable parameter ranges suggested by Xcalibre ($a \geq 0.30;\ -3.0 \leq$ b ≤ 3.0). Overall, 23 of the 70 items were identified as misfitting, all for unacceptably low discrimination parameter estimates ($a < .30$). See Appendix B.4 for the complete item-analysis table.

Figure 11 displays a graph of the Test Information Function for the 2PL model. As shown, maximum information obtained is 6.275 at θ = -0.90.

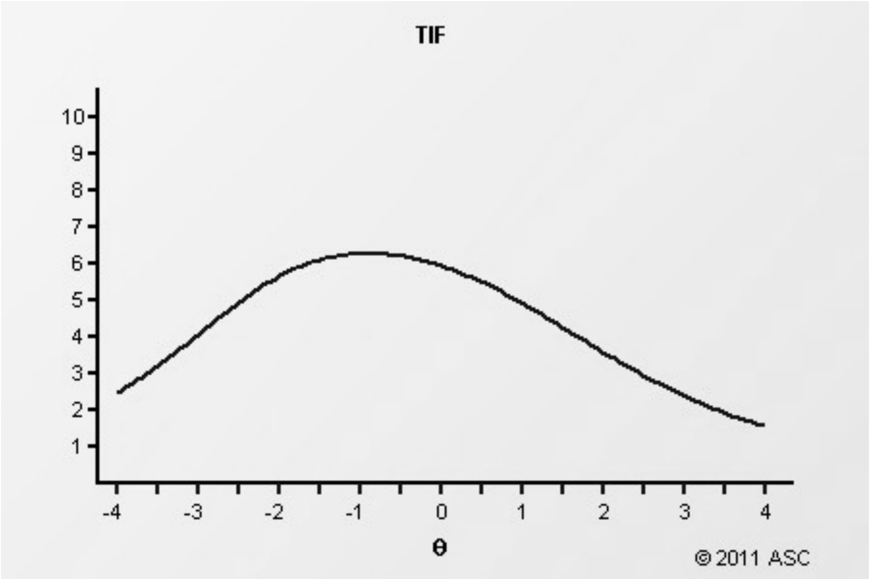


Figure 11. Test Information Function of the 2PL model

In comparison to the Rasch-1PL model, the 2PL Test Information Function (TIF) is considerably flatter, indicating items discriminate less near the mean but overall better across the distribution. Note too that maximum information was not obtained near $\theta = 0$ as it was in the Rasch model, but instead at θ = -0.90. This is visually apparent, with the peak of the curve on the left half of the distribution, well to the left of $\theta = 0$.

Figure 12 displays a graph of the Conditional Standard Error of Measurement (CSEM) Function for the 2PL model. As a reminder, the CSEM function is the inverse of the TIF and estimates the amount of error in theta estimation at each level of theta. For the 2PL model, the minimum CSEM modeled is 0.399 at θ = -0.90, which again is the same location as the point of maximum information.

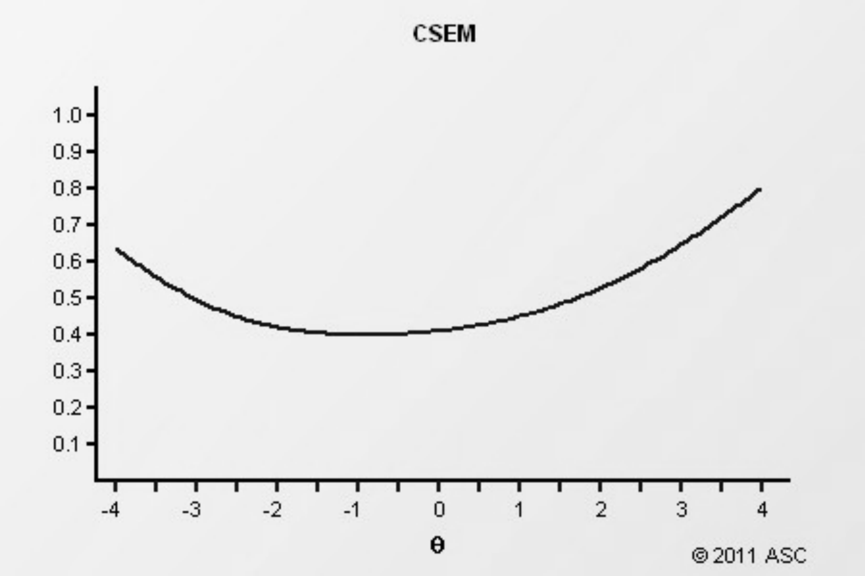


Figure 12. CSEM of the 2PL model

Figure 13 is a graph of the Test Response Function (TRF) for the 2PL model.

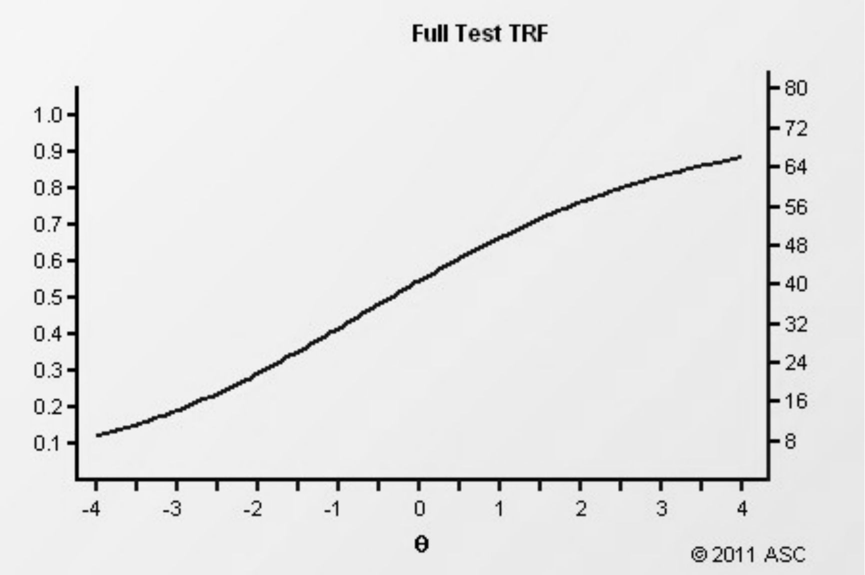


Figure 13. Test Response Function of the 2PL model

As illustrated, the 2PL model does not fit the data particularly well. The flat TRF across all levels of theta indicates items are not differentiating among examinees very well. To further investigate why this might be the case, Figure 14 portrays the scatterplot of the b-parameter (difficulty) by the a-parameter (discrimination) in the 2PL model.

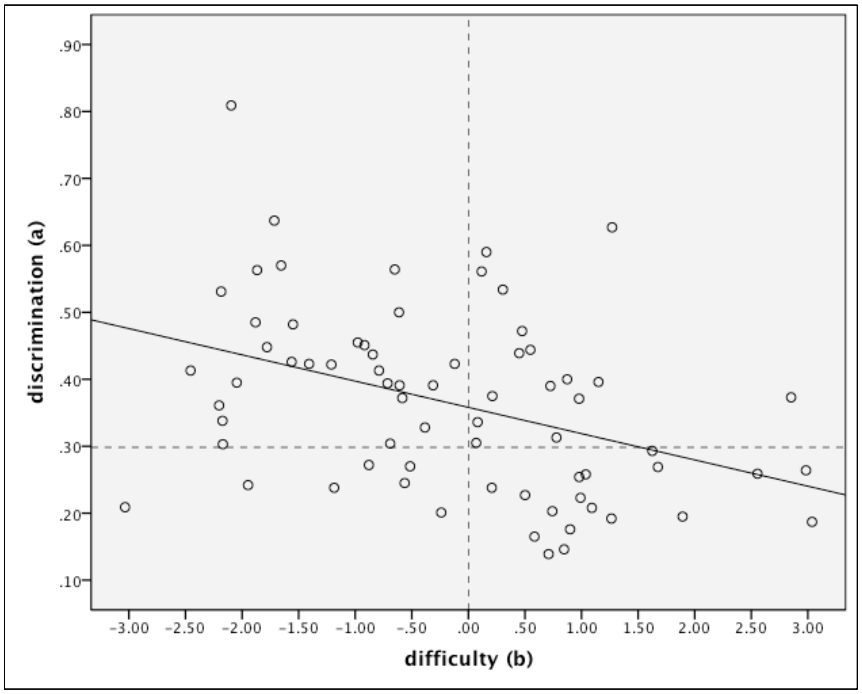


Figure 14. Scatterplot of difficulty (b) by discrimination (a)

The broken horizontal line inserted into the scatterplot crosses the Y-axis at the minimum level of discrimination considered acceptable for the model (0.30). As mentioned, 23 of the 70 items have unacceptably low discriminatory power when the 2PL model is fit to the data. The broken vertical line that crosses the X-axis at $\theta = 0$ was added to divide the graph into quadrants to further illustrate how items are functioning. Of particular note is item 36 (the most extreme outlier in the upper left quadrant), which exhibits the greatest discriminatory power of all items (a = .809) despite being one of the easiest (b = -2.096). In fact, visual inspection of the graph suggests easier items in general exhibit greater discriminatory power than more difficult items, which is confirmed by the solid fit line modeling the negative correlation between the two parameters (r = - .401, $p < .01$). It also concurs with the theta where maximum information is obtained (θ = -0.90). Given that the cut score for the 2PL model occurs near $\theta = 1.0$, this bias toward more information being available at the easier end of the difficulty contiuum is problematic for model fit.

To further examine the distribution of items at each level of theta, Figure 15 portrays a histogram of the difficulty (b) parameter estimates for the 2PL model.

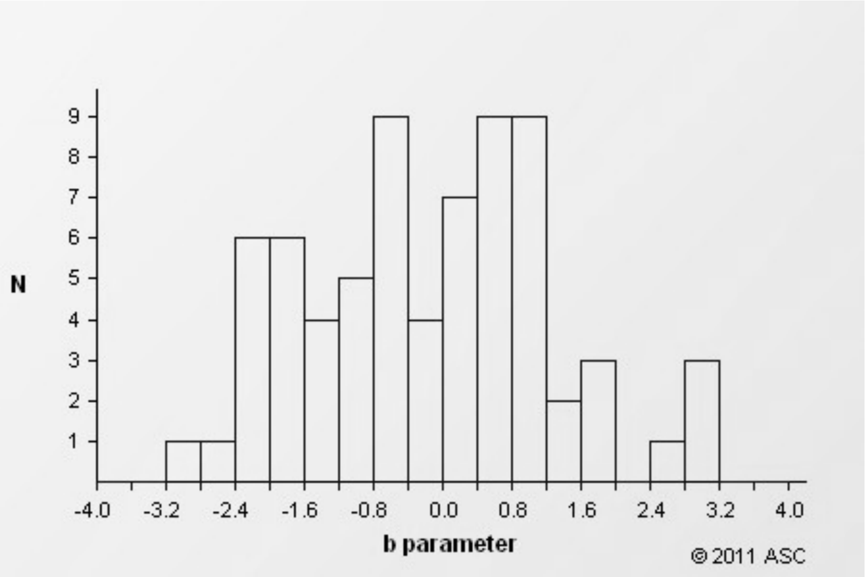


Figure 15. Histogram of the b-parameters of the 2PL model

As shown, there is confirmation the graph skews positive, meaning there are more items in the left half of the distribution than the right. That said, nine of the 70 items have difficulty estimates near the cut score ($\theta = 1.0$) that can help differentiate among candidates near the cut line.

3.4.5 IRT 3PL

The final item-analysis method employed is the IRT 3-paramter logistic (3PL) model. Table 9 contains an excerpt of the resulting item analysis when the 3PL model was employed using Xcalibre (Version 4.1).

Table 9. Excerpt of item analysis for the 3PL model

Item	a	b	c	Flag
1	0.278	-0.895	0.214	F
2	0.704	0.888	0.173	
3	0.583	1.189	0.174	
4	0.58	1.333	0.210	
5	0.405	1.893	0.227	
6	0.748	0.913	0.197	
7	0.539	-0.242	0.202	
8	0.872	0.715	0.188	
9	0.584	0.470	0.198	
10	0.553	1.575	0.232	
11	0.687	-0.196	0.190	
12	0.507	-0.030	0.208	
13	1.052	1.265	0.227	
14	0.535	2.023	0.192	
15	0.457	0.400	0.210	
16	0.355	2.162	0.223	
17	0.938	2.345	0.118	
18	0.278	2.030	0.209	F
19	0.534	-1.045	0.204	
20	0.488	-0.189	0.196	

Columns 2 through 4 of Table 9 contain the discrimination (a), difficulty (b), and guessing (c) parameter estimates of each item resulting from application of the 3PL model. The a-parameter for this model has a mean of 0.525 and standard deviation of 0.192, while the model forced the b-parameter to center its mean at 0 with a standard deviation of 1; the c-parameter for this model has a mean of 0.20 and a standard deviation of 0.02. Items were flagged if any of the three-parameter estimates exceeded the acceptable parameter ranges suggested by Xcalibre ($a \geq 0.30;\ -3.0 \leq b \leq 3.0;\ c < .40$) As shown, only two of the first 20 items were flagged for not fitting the model. Overall, only 10 of the 70 items were identified as misfitting the model, all for unacceptably low discrimination parameter estimates ($a < 0.30$). See Appendix B.5 for a complete list of the items flagged by the 3PL model.

Figure 16 is a graph of the Test Information Function (TIF) or the 3PL model. Maximum information is 7.822, obtained at θ = 1.250.

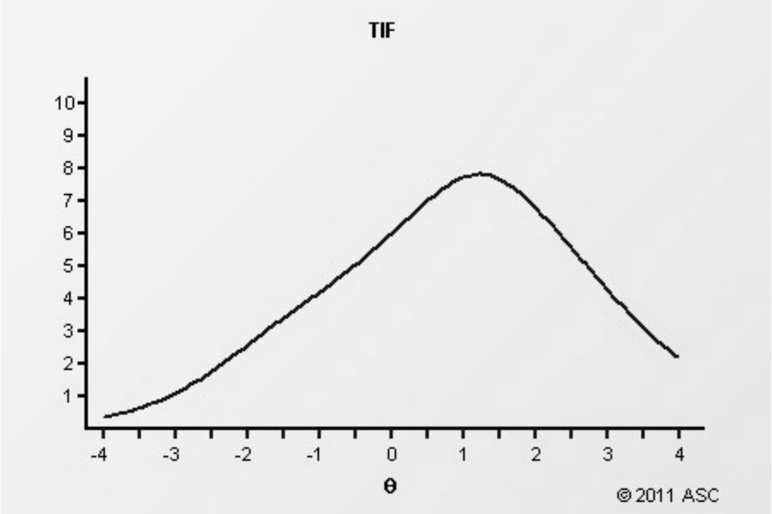


Figure 16. Test Information Function of the 3PL model

Note that the TIF of the 3PL model is similar in shape to the Rasch/1PL model in that it has a steeper peak and more symmetric distribution than did the 2PL TIF (Figure 11). Because the Y-axis of this graph is on a different scale than that of the Rasch model (see Figure 7), the maximum item information obtained in the 3PL model is considerably lower than that of the 1PL model. Nevertheless, the 3PL model discriminates very well near its peak. Note too that the 3PL peak is on the right half of the graph, which differs from both the 1PL and 2PL model. Because the cut score for the 3PL model is at $\theta = 1.1129$, there is much more information available around the cut score for the 3PL model than the other two IRT models, which is very favorable with respect to model fit.

Figure 17 is a graph of the Conditional Standard Error of Measurement (CSEM) Function for the 3PL model, which again is the inverse of the TIF. The minimum CSEM obtained is 0.358 at $\theta = 1.25$.

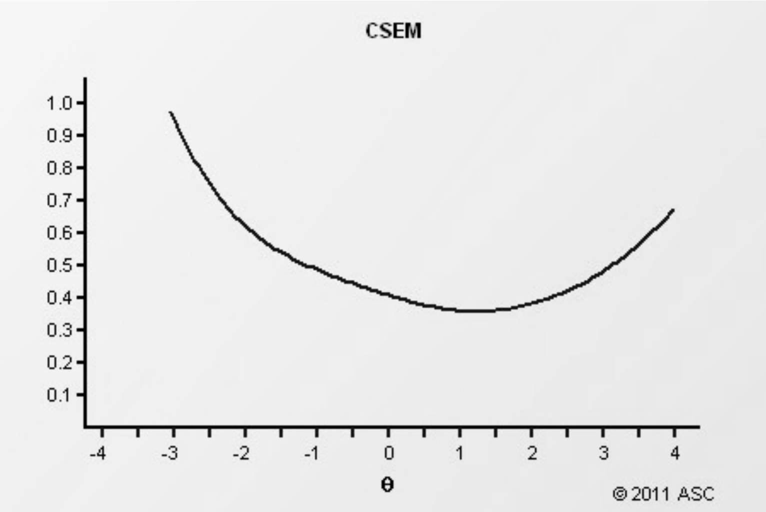


Figure 17. CSEM of the 3PL model

As with the other two IRT models, the 3PL model exhibits its worst fit in the tails of the distribution. As shown above, the CSEM nearly reaches 1.0 at the far left of the distribution, where examinee ability and item difficulty is lowest. This finding makes sense because there are fewer items and examinees at this end of the distribution in which to find reliable item-pattern responses. Generally speaking, the reliability of a test is proportional to its number of items, so with so few items and examinees at this end of the distribution, error in parameter estimation is expected to be quite high, and it is.

To view model fit from a different perspective, Figure 18 depicts the scatterplot of the b-parameter (difficulty) by the a-parameter (discrimination) for the 3PL model.

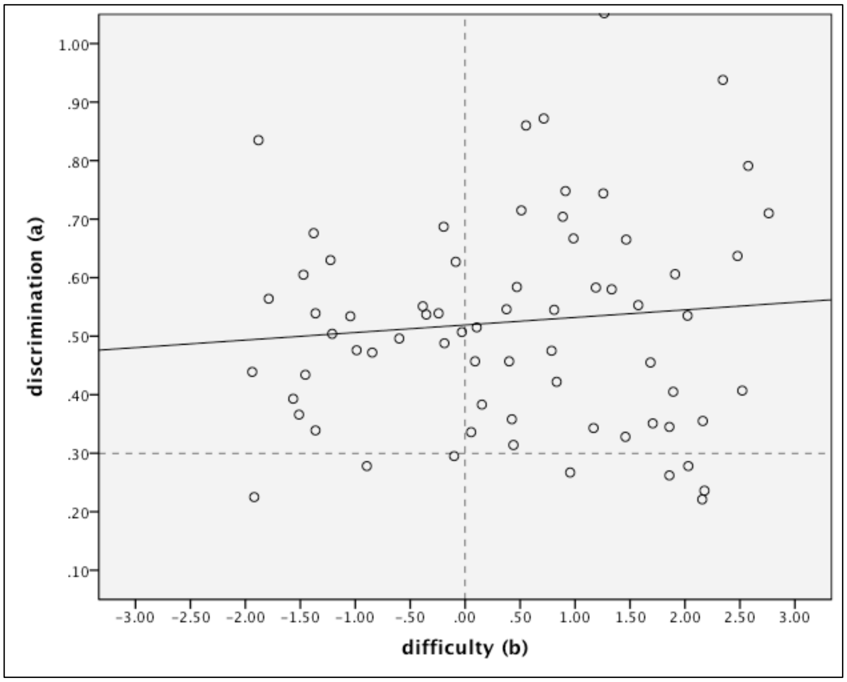


Figure 18. Scatterplot of difficulty (b) by discrimination (a)

Recall that the broken horizontal line inserted into the scatterplot crosses the Y-axis at the minimum level of discrimination considered acceptable (a = 0.30). The broken vertical line that crosses the X-axis at $\theta = 0$ was added to break the graph into quadrants to further illustrate how items are functioning in the 3PL model. As shown, the majority of items discriminate sufficiently. Item 36 is of particular note again, as it is the most extreme outlier in the upper left quadrant. However, it is not the most discriminating item overall as it is when the 2PL model is employed. Four other items discriminate better with the 3PL model, as illustrated by the items in the upper right quadrant that cross the Y-axis at values higher than Item 36. Moreover, a greater number of well-discriminating items are relatively difficult with the 3PL model, as opposed to the 2PL model where a greater number of items are relatively easy. This is visual confirmation the 3PL model seems to fit the data better, which is in part confirmed by the fact that the cut score for the 3PL is at $\theta = 1.1129$ and maximum information was obtained at $\theta = 1.25$, values considerably closer than those obtained with the 2PL model (where maximum information is at $\theta = -.90$ and the cut score is at $\theta = 1.0$).

Another difference between the scatterplots of the 2PL and 3PL models is the correlation between their respective difficulty and discrimination parameters. Whereas the 2PL model exhibits a statistically significant negative correlation between the two (r = - .401, $p < .01$), the 3PL model exhibits a statistically non-significant correlation (r = .089, n.s.), and the little correlation that does exist is positive rather than negative, as indicated by the slightly positive slope of the solid fit line that is portrayed in Figure 18.

To further examine the distribution of items when the 3PL model is applied, Figure 19 portrays a histogram of the difficulty (b) estimates of each item on the test.

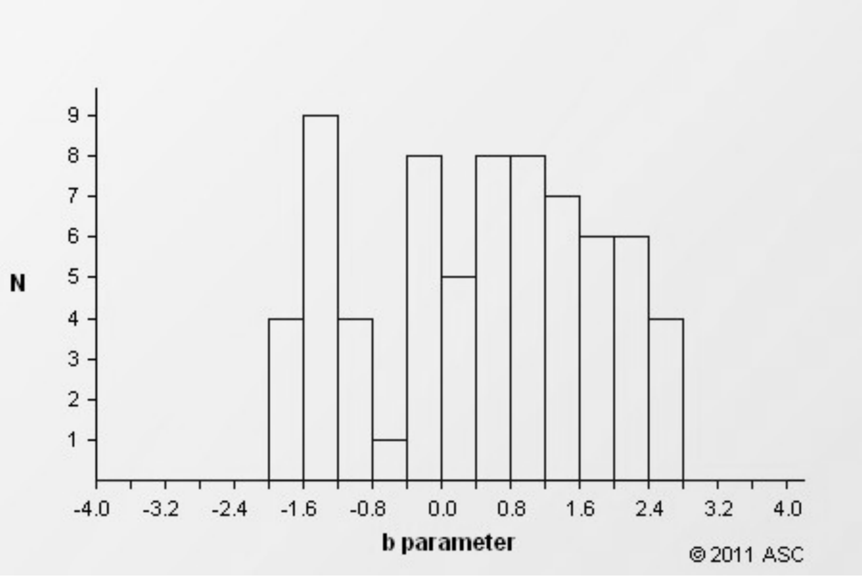


Figure 19. Histogram of the b-parameters of the 3PL model

As shown, the graph skews negative, with a greater number of items in the right half of the distribution. This is a desirable quality with respect to model fit because the cut score is at $\theta = 1.25$, indicating numerous items are at/near the cut score, thereby providing maximum information where it is most important for decision making.

3.5 Summary of Analyses

Given the number of analyses conducted to this point, this next section was created to synthesize the findings. To begin, Table 10 summarizes all of the items flagged by each model as misfitting.

Table 10. Summary of flagged items across models

Model	CTT (Kelley)	CTT (Pt Bis)	Rasch/1PL	IRT 2PL	IRT 3PL
Flagged Item #	1	1	1	1	1
	5	5	5	5	
			10	10	
	16	16	16	16	
	17				
	18	18	18	18	18
	21				
		22	22	22	
	23				
	28				
				31	
		32	32	32	
	33	33	33	33	33
	34	34	34	34	34
	35	35	35	35	35
	36				
	37	37	37	37	37
				38	
	39				
	41	41	41	41	41
	42	42		42	
			47	47	
				48	
	49		49	49	49
		57	57	57	57
	58	58	58	58	58
			65	65	
			66	66	
	69	69		69	
Total # of Items	19	15	18	23	10

As shown, there is considerable variation in the items flagged as faulty, both in terms of which items were flagged and the total number flagged. For example, only eight of the items were flagged by all five models (Items 1, 18, 33, 34, 35, 37, 41, and 58), while some items were flagged by only one of the models (e.g., Items 17, 23, 31). Moreover, the IRT-3PL model flagged only 10 of the 70 items overall, whereas the IRT-2PL model flagged more than twice as many (23).

Another means of comparison across models is the cross-tabulation values of each model pair. Cross-tabulations indicate the level of agreement (classification consistency) between model-pair rank orders. Table 11 is a summary of the cross-tabulations for each of the item-analysis methods employed relative to the baseline raw scores.

Table 11. Classification consistency across models

		Raw Score		Classification
		Admit	Reject	Consistency
CTT (Kelley)	Admit	383	50	94.4%
	Reject	79	1808	
CTT (PtBis)	Admit	392	54	94.7%
	Reject	70	1804	
Rasch/1PL	Admit	376	41	94.5%
	Reject	86	1817	
IRT 2PL	Admit	369	37	94.4%
	Reject	93	1821	
IRT 3PL	Admit	381	24	95.5%
	Reject	81	1834	

Recall that the purpose of using the raw-score data as the baseline is to determine the magnitude of displacement across models when compared to actual placement. As shown, each of the methods classifies about 95% – or about 2200 of the 2,320 – examinees the same as the raw-score method. However, anywhere from 105 (in the Raw-3PL comparison) to 130 (in the Raw-2PL comparison) of the examinees are classified differently. More specifically, the alternate-model outcomes would lead to a rejection of 70 to 93 more examinees than would the raw-score method, while they would lead to the admittance of 24 to 54 more examinees than would the raw-score method.

Another indication of the level of agreement between model pairs is the Kappa statistic. Table 12 illustrates the Kappa values for each pairwise comparison.

Table 12. Kappa statistics of each pairwise model comparison

Pairwise Comparison	Statistic	Std Dev	Approx T	Approx sig
Raw-CTT (Kelley)	0.821	0.015	39.599	0.000
Raw-CTT (Pt Bis)	0.830	0.015	39.998	0.000
Raw-Rasch/1PL	0.822	0.015	39.665	0.000
Raw-IRT 2PL	0.816	0.016	39.425	0.000
Raw-IRT 3PL	0.851	0.014	41.134	0.000

As shown, all pairwise comparisons exhibit statistically significant Kappa values ($\kappa \geq 0.8, p \ < .001$), with the 3PL model exhibiting the strongest agreement ($\kappa = 0.851$). While these Kappa values are indeed statistically significant, they indicate a fair amount of classification inconsistency, a matter of importance explored in subsequent chapters.

3.6 Test of IRT Assumptions

As illustrated, all of the models employed in this study achieve comparable levels of agreement with the original rankings derived from application of the raw-score method. However, as explained in the literature review, the IRT family of models requires two primary assumptions to be satisfied in order for their use to be considered valid. One is the assumption of unidimensionality, meaning all of the items on the test are measuring only one dimension of ability, in this case English proficiency.

3.6.1 Test of Unidimensionality

To test whether the assumption of unidimensionality is reasonably satisfied, a confirmatory factor analysis was conducted using structural equation modeling (SEM), where 1-factor and 2-factor models were applied to the data to test goodness of fit. Figures 20 and 21 illustrate the standardized factor loadings of the relationships between each latent trait and the observed measures of each trait, generated using AMOS 6.0.

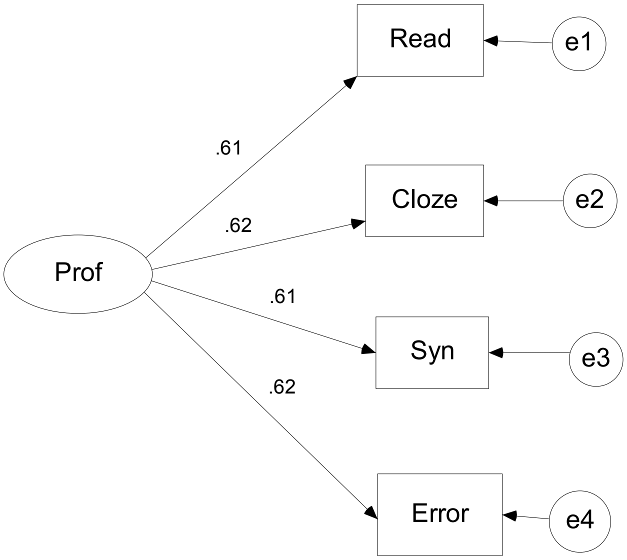


Figure 20. 1-factor SEM of the full data set

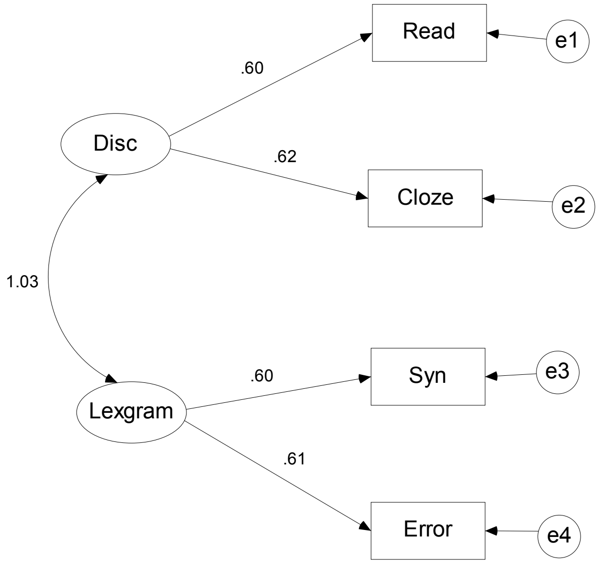


Figure 21. 2-factor SEM of the full data set

As shown, the latent trait of the 1-factor model (Prof: English proficiency) was split into two hypothesized traits for the 2-factor model (Disc: discourse ability, and LexGram: lexico-grammatical ability). However, as the standardized factor loadings between each latent trait and the observed measures of those traits show, the 1-factor model fits nearly identically to the 2-factor model, suggesting there is no meaningful advantage to splitting the latent trait of English proficiency into two traits. In short, the assumption of unidimensionality seems satisfied by these data. This conclusion is confirmed by a comparison of the Chi-square (χ^2) statistics of the two models.

Table 13. Chi-square comparison of CFA models

Model	NPAR	CMIN	DF	P	CMIN/DF
1-Factor	12	3.392	2	.183	1.696
2-Factor	13	2.356	1	.125	2.356

As shown in Table 13, both Chi-square statistics are statistically non-significant (CMIN/DF = 1.696, *p* = .183 and CMIN/DF = 2.356, *p* = .125, respectively), indicating each model fits the data well.[17] In fact, the 1-factor model fits even better than the 2-factor model as evidenced by the lower Chi-square (CMIN/DF) value, thereby providing further evidence the assumption of unidimensionality has been satisfied for these data.

[17] Note that the null hypothesis of the Chi-square goodness-of-fit test is that the model fits the data. As a result, these statistically non-significant findings indicate the null hypothesis cannot be rejected, which is a good thing because it means the models fit the data well.

3.6.2 Test of Local Item Independence

The second assumption required of IRT is local item independence, which again means the response to each item is independent of all other item responses. To determine whether this assumption is satisfied by the data, a testlet response theory (TRT) model was applied to the data using SCORIGHT (Version 3.0) because of the presence of three large testlets on the test (Sections I, II, and III – see Appendix A).

According to the SCORIGHT User's Manual (X. Wang, Bradlow, & Wainer, 2005), any testlet variance (λ_{ij}) – a measure of the examinee-by-testlet interaction – in excess of 0.30 is problematic as it indicates that 30% or more of the variance in examinee scores on the testlet is attributable to the item dependence found among the testlet items.

Table 14. Estimated gamma variances

Testlet	Variance	S.E.
1	0.5100	0.2502
2	0.4971	0.2523
3	0.4152	0.1564

As shown in Table 14, all three testlets exhibit a large testlet effect, with the first and second testlets exhibiting especially strong testlet effects, with about half of the variance in examinee scores in these two testlets attributable to examinee-by-testlet interaction. Because of this clear violation of the assumption of local item independence, there is considerable question as to whether it is appropriate to apply IRT models to these data, an issue explored in detail later.

4 Data Analysis II (Random 80%)

4.1 Participants

Recall that with IRT modeling, the 2PL and 3PL models are adjusted to maximize fit to the data. As a result, they are exploratory in nature and require replication of their findings before they can be considered valid (unlike CTT and Rasch/1PL models, which are confirmatory). Because the particular exam under investigation is both single use and single administration, there is no opportunity to validate the two models using data collected during another administration or from a parallel form of the exam. As a result, two similar sets of item analyses to those conducted in Chapter 3 were performed using resampled data to determine whether the nature and magnitude of the differences between models would hold across “different” examinee populations.

The first resampling analysis involved removal of a random 20% of the original candidate pool, leaving 1,856 of the original 2,320 examinee scores in tact. The motivation for choosing this subset of the population is to determine how classification consistency might change were there a smaller number of candidates in the pool but with a similar distribution of ability levels.

4.2 Descriptive Statistics

Descriptive statistics of the resampled data were calculated to determine whether they had similar characteristics to the original data. Table 15 displays the descriptive statistics for both the original data set and the resampled data set.

Table 15. Descriptive statistics of the original data and first resampling

N	Min	Max	Mean	Std Dev	Skewness		Kurtosis	
					Statistic	SE	Statistic	SE
1856	10	61	37.340	8.181	-.035	.057	-.171	.114
2320	5	65	37.303	8.254	-.064	.051	-.080	.102

As shown in Table 15, the range of scores for this candidate pool is somewhat narrower than the original range, with a minimum score of 10 and a maximum of 61. However, the means are very similar (37.340 vs. 37.303), as are the standard deviations (8.181 vs. 8.254), which suggest the two data sets have similar distributions. Figure 22 illustrates this is indeed the case.

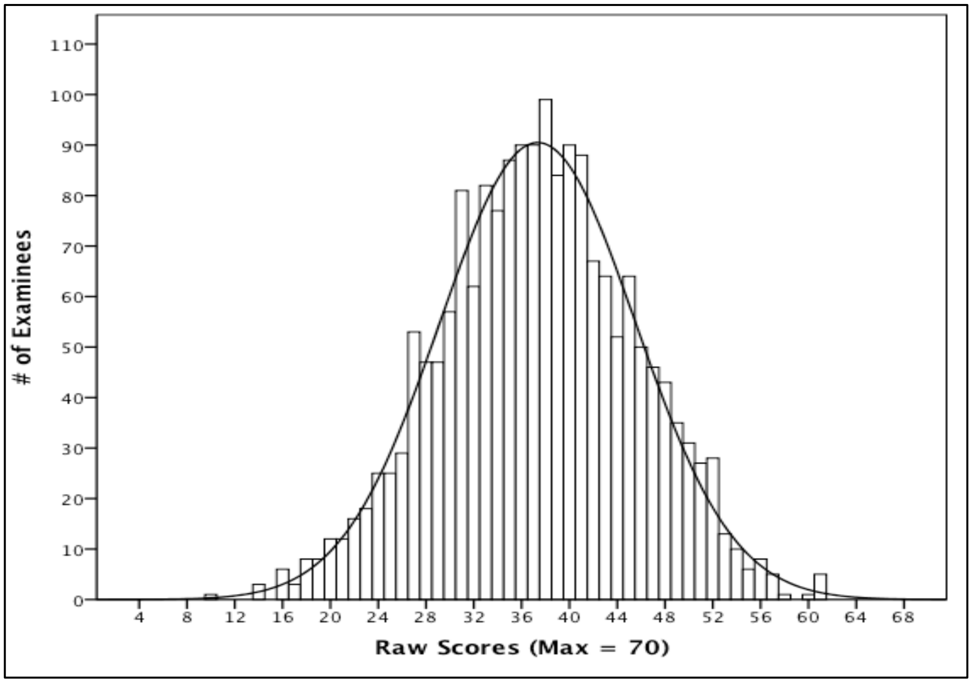


Figure 22. Distribution of examinee scores for the first resampling

As shown in Table 15 and Figure 22, this first resampling of the data exhibits slightly less skewness than the original distribution (-.035 vs. -.064), but it is more platykurtic than the original distribution (Kurtosis = -.171 vs. -.080). Nevertheless, the distribution of these data approximate a normal distribution like the original data.

4.3 Cut-Score Determination

The same protocol for cut-score determination was used with this resampling of the data as with the original data. Recall from Chapter 3 that the cut score is a function of the Japanese Ministry of Education annual quota and the historical acceptance rate for the policy studies department of this particular university, where:

$$\lambda = \frac{\mathrm{Q}}{\hat{\mathrm{r}}}$$

and sigma (λ) represents the number of admissions offers to be made. In 2006, the admissions year for which this test was administered, $\mathrm{Q} = 150$ and $\hat{\mathrm{r}} = 37\%$, so the target number of admissions offers was 405 ($\lambda = \frac{150}{0.37} = 405$). To determine the 405th highest score for this examinee pool, raw scores obtained from the actual administration of the test were compiled in Excel and sorted from highest to lowest. The resulting ranking indicates the 405th highest-ranking examinee achieved a raw score of 44, a score achieved or exceeded by the top 425 examinees. The next higher cut score, 45, was achieved by 373 candidates, 32 fewer than the target number of admission offers. Based on the historic pattern of admissions offers at this university, it is assumed here 425 examinees would have been offered admission rather than the fewer number of 373. Figure 23 on the following page depicts the cut score location along the distribution of examinee scores.

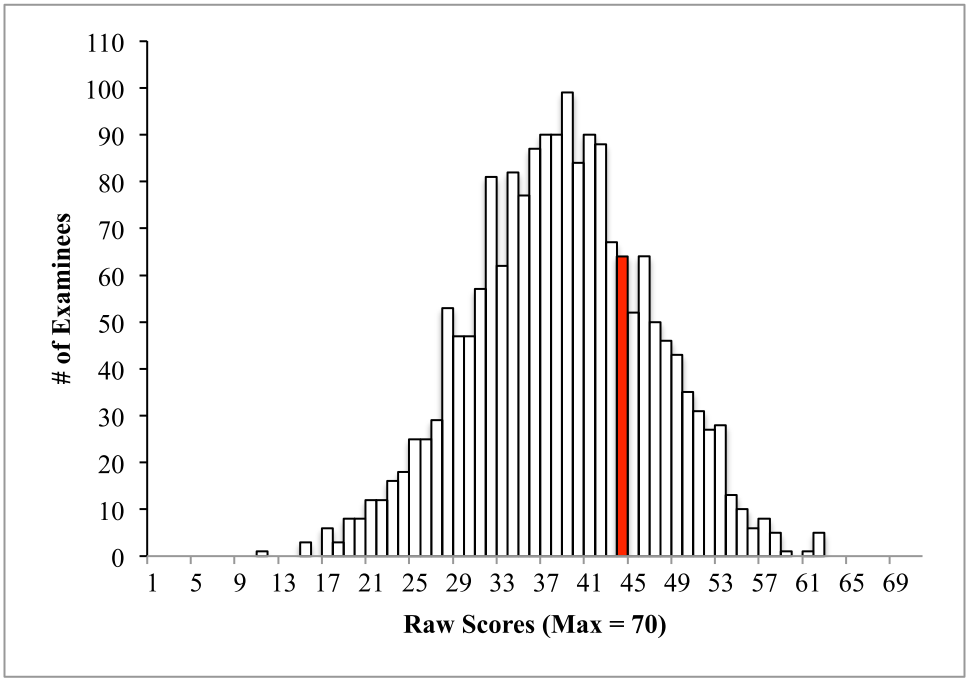


Figure 23. Cut score (44) along the distribution

4.4 Item Analyses

To help identify similarities and differences in the data sets (across examinee populations), the same set of item analyses performed on the original data set was performed on the resampled data. Following is a description of the results generated from each item analysis that was applied to the resampled data.

4.4.1 CTT (Kelley)

Once again employing the item-analysis method introduced by Kelley (1939), the performance of the upper 25% and lower 25% of the distribution was conducted. Table 16 contains an excerpt of this item analysis.

Table 16. Excerpt of item analysis based on Kelley's D

Item	UG	LG	Total	*p*	D	Flag
1	380	300	1273	69%	19%	F
2	306	119	767	41%	39%	
3	276	114	730	39%	34%	
4	272	130	750	40%	30%	
5	248	160	748	40%	20%	
6	317	124	797	43%	41%	
7	385	220	1181	64%	36%	
8	343	121	817	44%	47%	
9	339	164	968	52%	37%	
10	261	149	750	40%	24%	
11	394	199	1173	63%	42%	
12	371	212	1124	61%	35%	
13	280	124	676	36%	33%	
14	201	104	589	32%	21%	
15	345	200	1016	55%	32%	
16	234	163	741	40%	16%	F
17	122	45	281	15%	16%	F
18	249	180	808	44%	16%	F
19	436	305	1403	76%	30%	
20	374	218	1148	62%	34%	

As shown, four of the first 20 items were flagged as faulty, all for unreasonably low discrimination values (D < 0.20). In addition, Item 17 was flagged for being so difficult (p = 0.15) that use of Kelley's index has less validity because of where the item falls along the distribution. Overall, 21 of the 70 items were flagged as faulty, all for having low discrimination indexes and/or levels of difficulty outside the range considered acceptable for this form of item analysis (p < 0.20 or p > 0.80). See Appendix C.1 for the complete item analysis of this data set.

4.4.2 CTT (Point Biserial)

The second type of item analysis performed on the resampled data is examination of the point-biserial correlations, which again is a measure of the reliability between individual test items and the overall test. Recall that while there is no firm benchmark for acceptable levels of point-biserial correlation, convention within the language-testing field suggests a correlation of at least 0.20 is acceptable.

Table 17. Excerpt of item analysis based on point-biserial correlations

Item	rpb	Flag
1	0.20	
2	0.34	
3	0.30	
4	0.26	
5	0.18	F
6	0.33	
7	0.31	
8	0.38	
9	0.32	
10	0.20	
11	0.38	
12	0.30	
13	0.28	
14	0.20	
15	0.27	
16	0.16	F
17	0.21	
18	0.14	F
19	0.31	
20	0.30	

As illustrated, three of the first 20 items were identified as having unacceptably low correlations (Items 5, 16 and 18). Overall, 15 of the 70 items were flagged for having unacceptably low correlations (see Appendix C.2 for a complete list of flagged items).

4.4.3 Rasch/1PL

Item analysis within the Item Response Theory (IRT) framework was also conducted. The first of these analyses was application of the 1-parameter logistic (1PL) model, or more specifically, the Rasch equivalent of the 1PL (see Section 2.1.3.3.1 of the literature review for additional detail). For the Rasch model, the decision criterion for item acceptability is a Z-standard value equal to or less than 2.0 (Linacre, 2002).

Table 18. Excerpt of item analysis for the Rasch/1PL model

		Infit		Exact Match		
Item	b	MnSq	Z-std	Obs%	Exp%	Flag
1	-0.84	1.03	1.33	68.4	69.5	
2	0.37	0.95	-3.35	65.9	62.9	
3	0.46	0.98	-1.56	64.7	64.0	
4	0.41	1.00	-0.09	63.5	63.4	
5	0.42	1.05	3.19	61.6	63.5	F
6	0.30	0.96	-3.24	66.2	62.2	
7	-0.60	0.97	-1.90	67.2	65.7	
8	0.25	0.93	-5.57	67.2	61.8	
9	-0.10	0.97	-2.76	64.1	60.8	
10	0.41	1.03	2.01	62.2	63.4	F
11	-0.59	0.93	-4.46	69.1	65.5	
12	-0.47	0.98	-1.57	64.7	63.9	
13	0.59	0.98	-1.20	69.3	65.7	
14	0.81	1.03	1.32	69.1	69.2	
15	-0.21	1.00	-0.08	61.4	61.4	
16	0.44	1.06	3.83	59.5	63.7	F
17	1.83	0.99	-0.17	84.7	84.9	
18	0.27	1.07	5.35	56.9	62.0	F
19	-1.21	0.96	-1.53	76.7	75.8	
20	-0.52	0.97	-1.67	65.8	64.6	

As shown in Table 18, four of the first 20 items were flagged as faulty when applying the Rasch/1PL model (Items 5, 10, 16, and 18), with Item 18 particularly misfitting the model (Z-std = 5.35). Overall, 17 of the 70 items misfit the model (see Appendix C.3 for a complete list of the flagged items).

4.4.4 IRT 2PL

The second IRT-model application was the 2-parameter logistic (2PL) model. Recalling that the 2PL model accounts for both difficulty (b) and discrimination (a) among items (vs. only difficulty in the case of the Rasch/1PL model), the criteria for flagging items as faulty are unacceptably low discrimination and/or unacceptably low or high difficulty parameter estimations.

Table 19. Excerpt of item analysis for the 2PL model

Item	a	b	Flag
1	0.273	-1.736	F
2	0.470	0.472	
3	0.382	0.705	
4	0.317	0.741	
5	0.217	1.056	F
6	0.450	0.391	
7	0.431	-0.849	
8	0.547	0.277	
9	0.420	-0.157	
10	0.253	0.899	F
11	0.576	-0.662	
12	0.407	-0.697	
13	0.375	0.911	
14	0.265	1.717	F
15	0.331	-0.380	
16	0.195	1.219	F
17	0.381	2.796	
18	0.172	0.854	F
19	0.497	-1.504	
20	0.402	-0.780	

As shown in Table 19, six the of the first 20 items were flagged as faulty when the 2PL model was applied to the data, all for unacceptably low discrimination values ($a < .30$). Overall, 27 of the 70 items were flagged when the 2PL model was applied (see Appendix C.4 for the complete item-analysis table).

Because of the greater complexity of the IRT-2PL model compared to the Rasch/1PL model or either of the CTT item-analysis methods, much more information is generated from the analysis. Recall that one example of this is the Test Information Function (TIF), which is an indication of how much information about examinee performance the test is providing at each level of theta. For this resampling of the data, maximum information is 6.233 at $\theta = -0.850$, as shown in the graphical representation of the Test Information Function (TIF) portrayed in Figure 24.

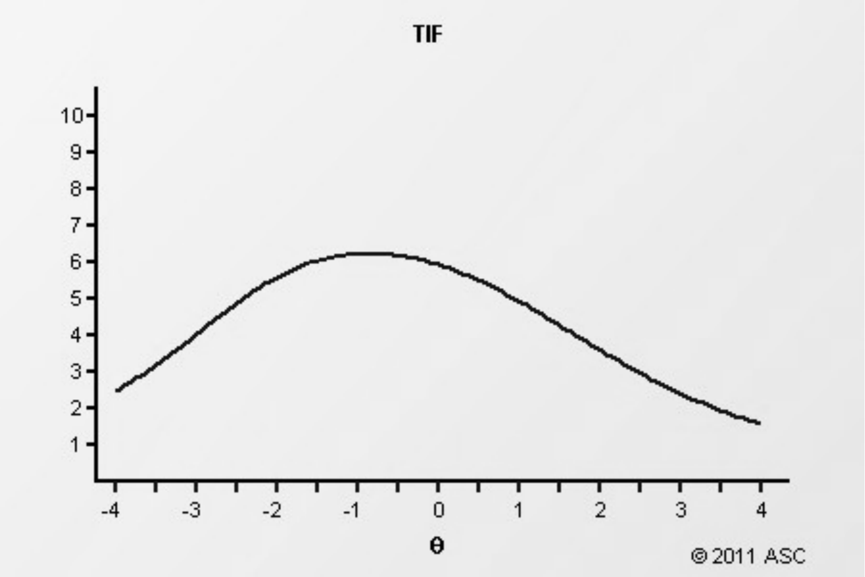


Figure 24. Test Information Function of the 2PL model

Like the TIF for the 2PL model of the original data set, this TIF is somewhat flat, with its peak well left of the mean. As a result, it seems to indicate discrimination is highest at relatively low levels of theta, which is troubling considering the fact that the cut score is higher than average for these data ($\theta = 1.0026$).

Figure 25 displays a graph of the Conditional Standard Error of Measurement (CSEM) Function for the 2PL model, which again is the inverted function of the TIF and estimates the amount of error in theta-estimation for each level of theta. The minimum CSEM for this application of the 2PL model is 0.401, achieved at θ = -0.850.

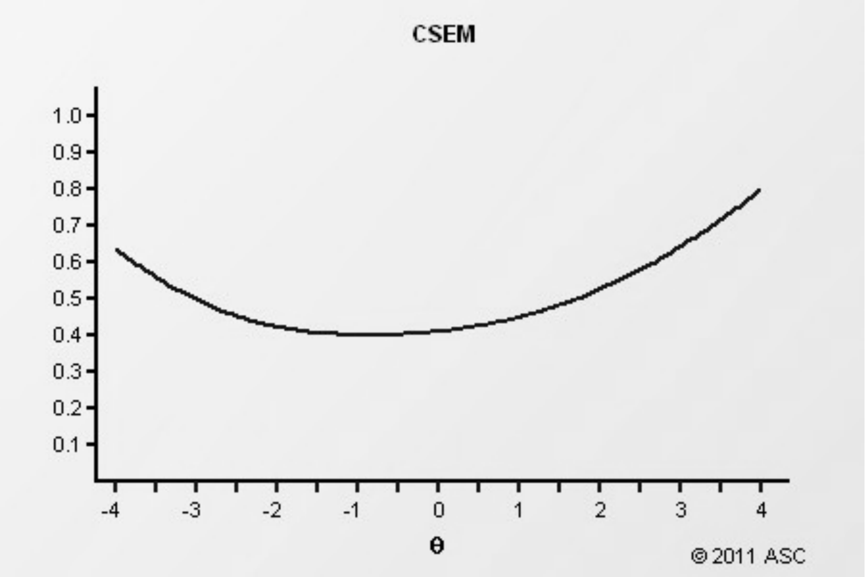


Figure 25. CSEM of the 2PL model

As shown, the CSEM function is generally high across the distribution, which indicates the 2PL model does not seem to fit the data particularly well. This finding mirrors the finding of the first analysis (see Figure 12).

Figure 26 displays a graph of the Test Response Function (TRF) for all items on the exam. Recall that the TRF predicts the proportion/number of items that examinees are predicted to answer correctly at any given level of theta. The left Y-axis portrays the predicted proportion-correct, while the right Y-axis represents the predicted number-correct for all levels of theta.

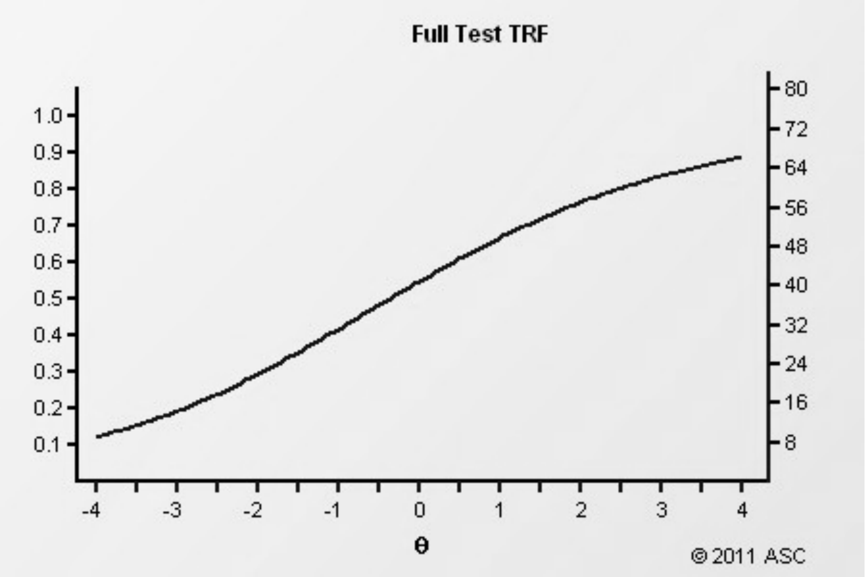


Figure 26. Test Response Function of the 2PL model

Synthesizing the information portrayed in Figures 24-26, it is clear the 2PL model does not fit the data very well, much like in the case of the original candidate pool (see Chapter 3, Figures 10 through 12). Once again, maximum information is obtained at a level of theta (-0.850) well below that of the cut score (1.0026). In other words, there is much greater error in parameter estimates around the cut score than desirable, as was the case in the original data analysis.

To view the fit of the 2PL model to the data from another perspective, Figure 27 illustrates the scatterplot of the b-parameter (difficulty) estimates by the a-parameter (discrimination) estimates.

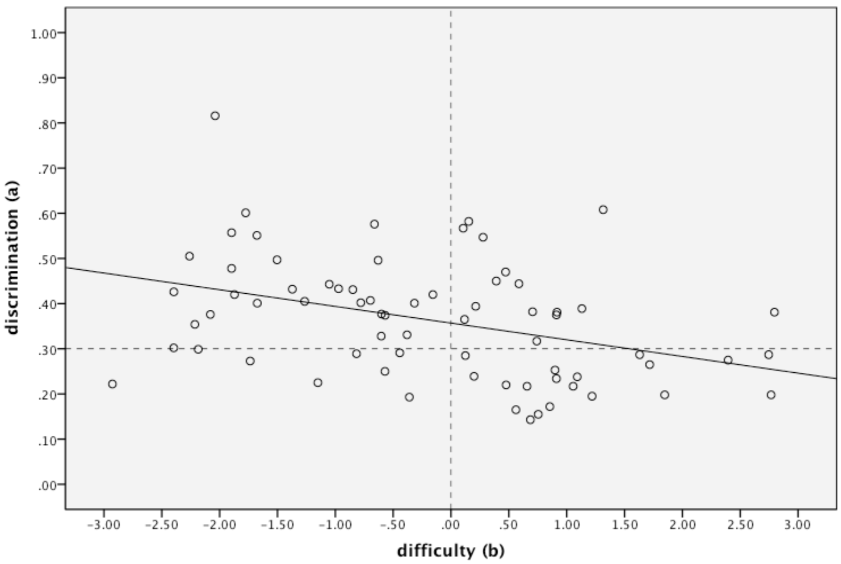


Figure 27. Scatterplot of difficulty (b) by discrimination (a)

The broken horizontal line inserted into the scatterplot crosses the Y-axis at the minimum level of discrimination considered acceptable for the model (a = 0.30). As mentioned, 23 of the 70 test items have unacceptably low discriminatory power when the 2PL model is fit to the data. The broken vertical line that crosses the X-axis at $\theta = 0$ was added to break the graph into quadrants to further illustrate how items are functioning in the 2PL model. Of particular note is item 36 (the most extreme outlier in the upper left quadrant). As was the case in the analysis of the full data set, Item 36 exhibits the greatest discriminatory power of all the items (a = .816) despite being one of the easiest (b = -2.039). Once again, visual inspection of the graph suggests easier items in general exhibit greater discriminatory power than more difficult items, which is confirmed by the slight negative correlation between the two parameters (r = - .1.208). It also concurs with the theta where maximum information is obtained (θ = -0.850). Given that the cut score for the 2PL model occurs near θ = 1.0026, this bias toward more information being available at the easier end of the difficulty contiuum is problematic for model fit.

Finally, Figure 28 portrays a histogram of the difficulty (b) estimates of each item on the test, further illustrating the relative facility of the items as a whole.

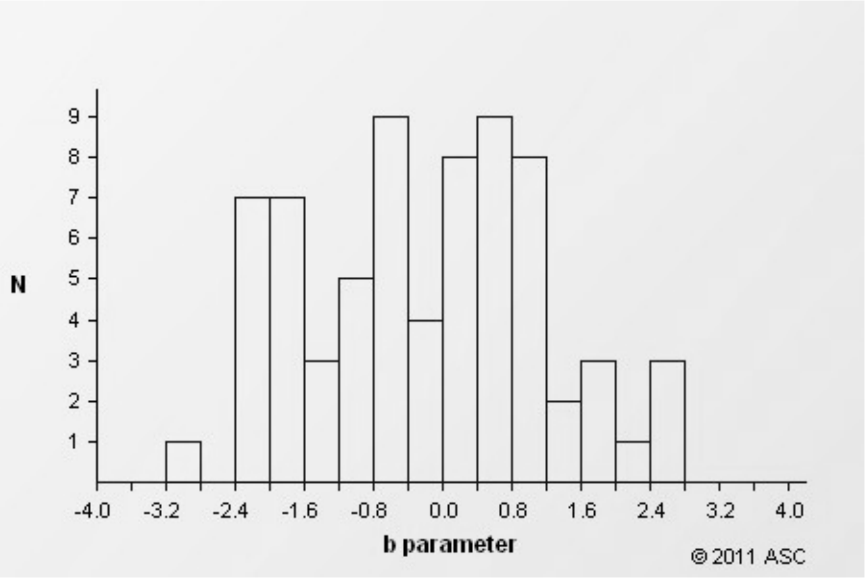


Figure 28. Histogram of the b-parameters of the 2PL model

As shown, the majority of items are on the left half of the distribution (below $\theta = 0$), which corresponds to the shape of the TIF curve in Figure 24. Nevertheless, eight of the 70 items at least approximate the level of difficulty at the cut score of $\theta = 0.750$, which helps contribute to the classification accuracy of the model where it is most important.

4.4.5 IRT 3PL

One final analysis conducted was application of the IRT-3PL model. Building on the foundation of the 2PL model, recall that the 3PL model estimates not only the difficulty (b) and discrimination (a) parameters, but also a guessing parameter (c) based on the belief that examinees will at least make an educated guess at any dichotomously scored item when there are no penalties for guessing.

Table 20 depicts an excerpt of the item analysis that results from application of the 3PL model on the resampled data.

Table 20. Excerpt of item analysis for the 3PL model

Item	a	b	c	Flag
1	0.32	-0.75	0.22	
2	0.73	0.91	0.19	
3	0.59	1.21	0.19	
4	0.58	1.32	0.21	
5	0.41	1.88	0.23	
6	0.78	0.88	0.20	
7	0.54	-0.20	0.21	
8	0.91	0.70	0.19	
9	0.59	0.47	0.21	
10	0.54	1.57	0.23	
11	0.70	-0.20	0.20	
12	0.53	-0.01	0.21	
13	1.05	1.24	0.23	
14	0.54	2.10	0.20	
15	0.46	0.42	0.22	
16	0.36	2.16	0.23	
17	0.93	2.40	0.12	
18	0.29	2.05	0.22	F
19	0.56	-0.98	0.21	
20	0.49	-0.12	0.21	

As shown, only one of the first 20 items was flagged as faulty (Item 18, for an unacceptably low correlation: a < 0.30). Overall, only nine of the 70 items misfit the 3PL model, the fewest number of items among all models employed (see Appendix C.5 for the complete item-analysis table).

Figure 29 further illustrates the fit of the 3PL model to the data. As shown, the graphical representation of the Test Information Function (TIF), a display of how much information about examinee scores the test is providing at each level of theta, reaches its maximum level of 8.042 at $\theta = 1.250$.

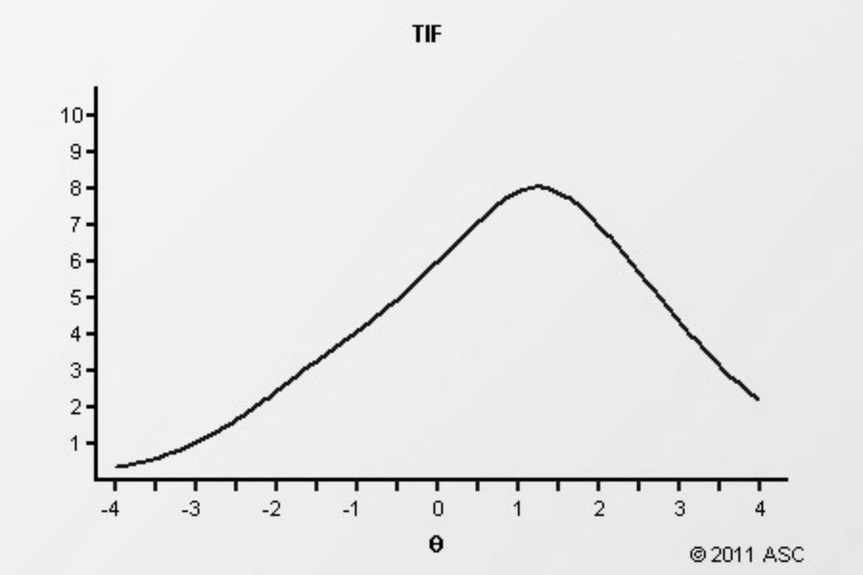


Figure 29. Test Information Function of the 3PL model

As was the case with the original data set, application of the 3PL model to this data set yields a fairly peaked curve, with maximum information near the cut score, which again is a desirable outcome for these data.

Figure 30 displays the Conditional Standard Error of Measurement (CSEM) of the 3PL model, which again is the inverse of the TIF function and portrays the amount of error in each theta estimate. In this case, minimum CSEM is 0.353, achieved at θ = 1.250.

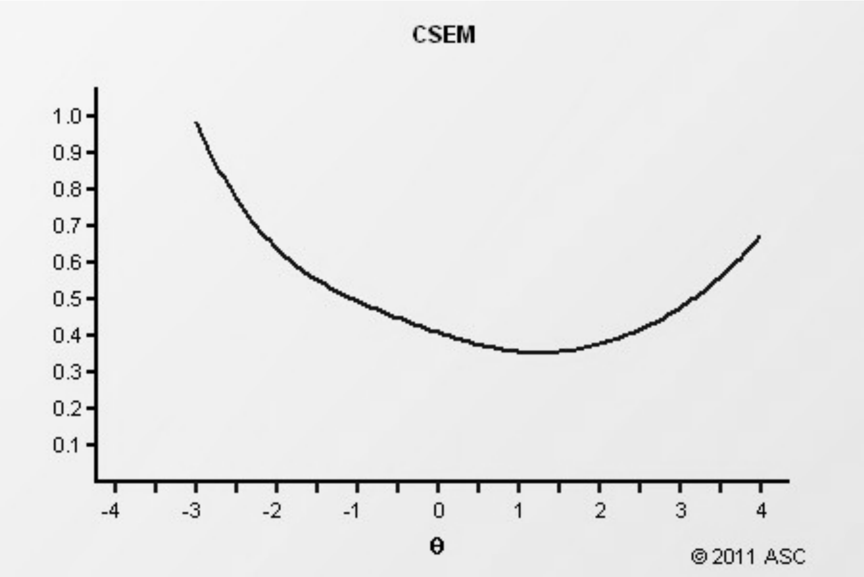


Figure 30. CSEM of the 3PL model

One other graphical representation of how the model fits these data is the Test Response Function (TRF), shown in Figure 31. Recall that the TRF predicts the proportion/number of items that examinees are expected to answer correctly at each level of theta. The left Y-axis represents the predicted proportion correct while the right Y-axis depicts the predicted number of correct responses.

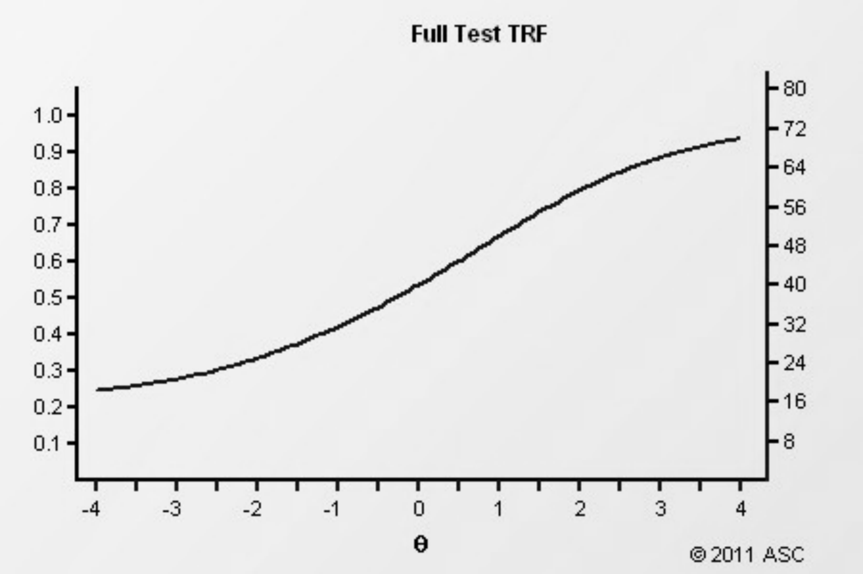


Figure 31. Test Response Function of the 3PL model

Taken together, Figures 29-31 illustrate that the 3PL model fits the resampled data pretty well, which replicates the findings of the original data analysis. This conclusion is confirmed in Figure 32 below.

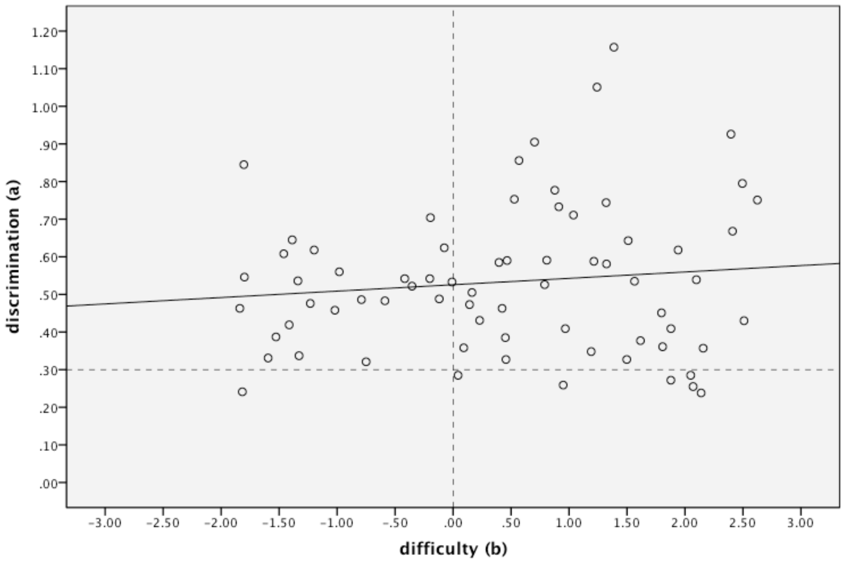


Figure 32. Scatterplot of difficulty (b) by discrimination (a)

The broken horizontal line inserted into the scatterplot crosses the Y-axis at the minimum level of discrimination considered acceptable (a = 0.30). The broken vertical line that crosses the X-axis at $\theta = 0$ was added to break the graph into quadrants to further illustrate how items are functioning in the 3PL model. As shown, the majority of items discriminate sufficiently when the 3PL model is applied to the data, with a majority of the items above the broken horizontal line. Item 36 is of particular note again, as it is the most extreme outlier in the upper left quadrant. However, it is not the most discriminating item overall as it is when the 2PL model is applied to the data. Five other items actually discriminate better within the 3PL model (Items 8, 13, 17, 53, and 60), illustrated by those items in the upper right quadrant that cross the Y-axis at values higher than Item 36. Moreover, a greater number of well-discriminating items are relatively difficult (see

Figure 33), as opposed to the 2PL model, where a greater number of items are relatively easy (see Figure 28).

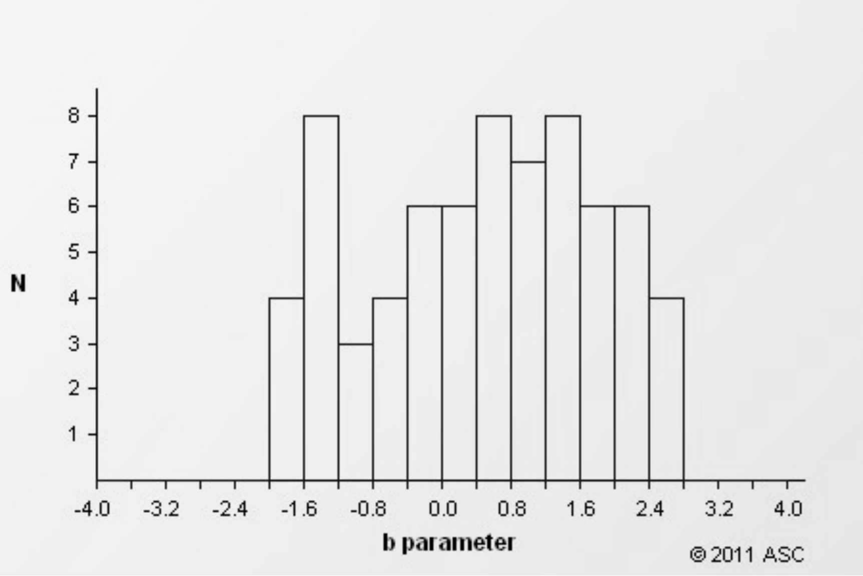


Figure 33. Histogram of the b-parameters of the 3PL model

Overall, Figures 32 and 33 confirm that the 3PL model fits the data better than the 2PL model, which is further confirmed by the fact that the cut score for the 3PL is at $\theta = 1.1129$ and maximum information was obtained at $\theta = 1.250$, values considerably closer on the distribution than the corresponding values obtained with the 2PL model (where maximum information is at $\theta = -.850$ and the cut score is at $\theta = 1.0026$).

4.5 Summary of Analyses

In an effort to synthesize the respective outcomes of each item analysis, Table 21 is a summary of the items flagged as faulty according to each model's criteria. As shown, the 2PL model identified the most number of misfit items (27), while the 3PL model identified the fewest, exactly one-third the number of the 2PL model (9).

Table 21. Summary of the items flagged across models

Model	CTT (Kelley)	CTT (Pt Bis)	Rasch/1PL	2PL	3PL
Flagged Items	1			1	
		5	5	5	
			10	10	
				14	
	16	16	16	16	
	17				
	18	18	18	18	18
	21				
			22	22	
	23				
	24				
	26			26	
	28				
				30	
				31	
		32	32	32	
	33	33		33	33
	34	34	34	34	34
	35	35	35	35	35
	36				
	37	37	37	37	37
			38	38	
	39				
	41	41	41	41	41
	42	42		42	
			47	47	
				48	
	49	49	49	49	49
	50				
		57	57	57	57
	58	58	58	58	58
				63	
			65	65	
		66	66	66	
	69	69		69	
# Items Flagged	21	15	17	27	9

As is the case with the analysis of the full data set, there is considerable variation in the items flagged across the models. For example, only seven of the items were flagged across all models (Items 18, 34, 35, 37, 41, 49, and 58), which is one fewer than in the full-data analysis. Interestingly, all seven of these items are in the group identified in the analysis of the full data set. As with the original analysis, several items were identified as misfitting only one model, although there are many more in this second analysis (Items 14, 17, 21, 23, 24, 28, 30, 31, 36, 39, 48, 50, and 63). Interestingly, all of the items identified as misfitting only one model were identified by either the CTT-Kelley method or the IRT-2PL model.

Another means of comparison across models is the cross-tabulation values of each model pair. As a reminder, cross-tabulations indicate the level of agreement (classification consistency) between model rankings. Table 22 is a summary of the cross-tabulations for each of the models relative to the baseline raw scores.

Table 22. Classification consistency across models

		Raw Score		Classification Consistency
		Admit	Reject	
CTT (Kelley)	Admit	379	51	94.8%
	Reject	46	1380	
CTT (PtBis)	Admit	377	48	94.8%
	Reject	48	1383	
Rasch/1PL	Admit	386	80	93.6%
	Reject	39	1351	
IRT 2PL	Admit	381	24	96.3%
	Reject	44	1407	
IRT 3PL	Admit	378	27	96.0%
	Reject	47	1404	

As shown, classification consistency ranges from about 93.6% (for the Rasch model) to a maximum of 96% (for the 3PL model). In other words, a similar level of classification consistency was found with this resampling as with the full data set, albeit with a slightly wider range than with the original data. Nevertheless, as stated earlier, a 4-6% difference in classification is non-trivial: anywhere from 74 (in the Raw-3PL comparison) to 119 (in the Raw-Rasch/1PL comparison) of the 1,856 examinees are classified differently across models when compared to the baseline raw-score method. More specifically, the alternate-model outcomes would lead to a rejection of 39 to 48 more examinees than would the raw-score method, while they would lead to the admittance of 24 to 80 more examinees than would the raw-score method.

One other indication of the level of agreement between model pairs is the Kappa statistic. Table 23 illustrates the Kappa values for each pairwise comparison.

Table 23. Kappa statistics of each pairwise model comparison

Pairwise Comparison	Statistic	Std Dev	Approx T	Approx sig
Raw-CTT (Kelley)	0.853	0.015	36.732	0.000
Raw-CTT (Pt Bis)	0.854	0.015	36.771	0.000
Raw-Rasch/1PL	0.824	0.015	35.580	0.000
Raw-IRT 2PL	0.894	0.013	38.555	0.000
Raw-IRT 3PL	0.885	0.013	38.153	0.000

As shown, all pairwise comparisons exhibit statistically significant Kappa values ($\kappa \geq 0.8, p \ < .001$), with the 2PL model exhibiting the strongest agreement ($\kappa = 0.894$) and the 3PL model second strongest agreement ($\kappa = 0.885$). Overall, these values indicate respectable levels of agreement, but they are far from ideal, particularly given the high-stakes nature of the exam.

4.6 Test of IRT Assumptions

Because the 3PL model was designed specifically to model dichotomously scored multiple-choice items, it is of little surprise it seems to fit both the original data and the resampled data well. Recall, however, that all IRT models require satisfaction of two strong assumptions: unidimensionality and local item independence. Failure to satisfy either of these assumptions often leads to overestimations of parameter precision and underestimations of error, which if severe enough, can undermine the validity of the model's application to the data. As a result, and in a fashion similar to the original analysis in Chapter 3, both of these assumptions were tested using the resampled data.

4.6.1 Test of Unidimensionality

To test whether the assumption of unidimensionality is reasonable for the resampled data, a confirmatory factor analysis was conducted using structural equation modeling (SEM), where 1-factor and 2-factor models were applied to the data to test their respective goodness of fit. Figures 34 and 35 illustrate the standardized factor loadings of the relationships between each latent trait and the observed measures of each trait.

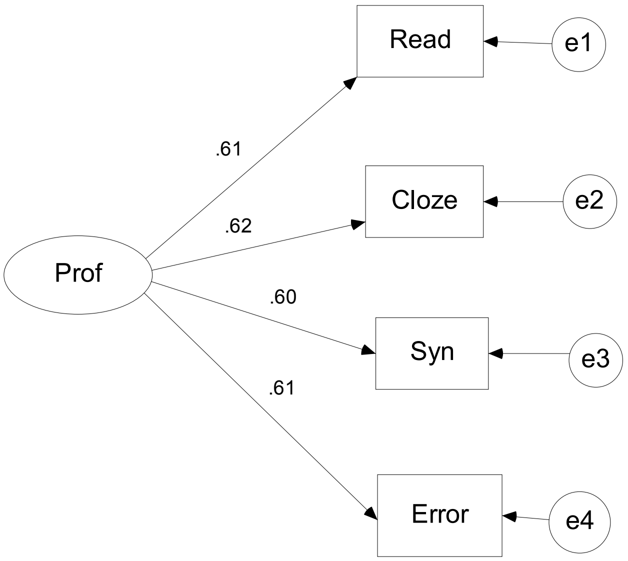


Figure 34. 1-factor SEM of the first resampling

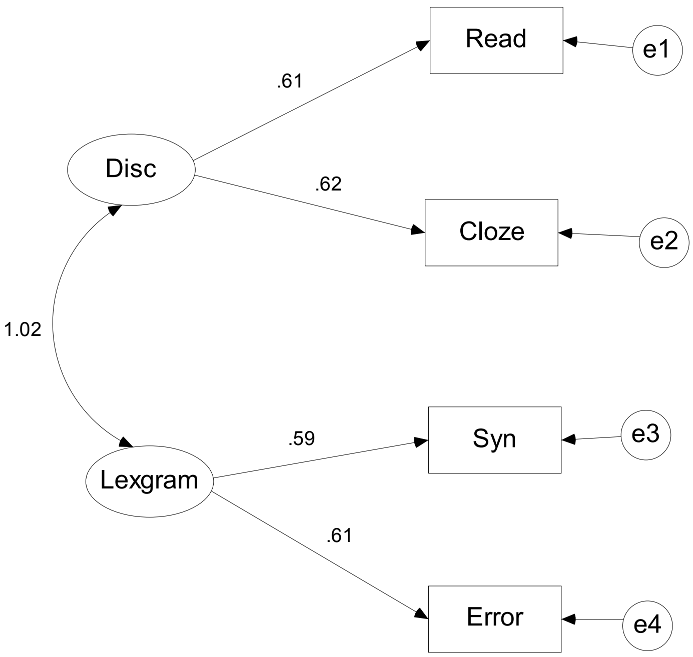


Figure 35. 2-factor SEM of the first resampling

As shown, the latent trait of the 1-factor model (English proficiency) was split into two hypothesized traits for the 2-factor model (discourse ability and lexico-grammatical ability). However, as the standardized factor loadings show, the 1-factor model fits nearly identically to the 2-factor model (e.g., General Proficiency → Reading Comprehension = .61; Discourse → Reading Comprehension = .61), suggesting there is no advantage to splitting the latent trait of English proficiency into two traits. This conclusion is confirmed by a comparison of the Chi-square statistics of the two models.

Table 24. Chi-square comparison of CFA models

Model	NPAR	CMIN	DF	P	CMIN/DF
1-Factor	12	3.160	2	.206	1.580
2-Factor	13	2.776	1	.096	2.776

As shown, both Chi-square statistics are non-significant (CMIN = 3.160, *p* = .206 and CMIN = 2.776, *p* = .096, respectively), indicating each model fits the data well. In fact, the 1-factor model fits the data better, recalling that the null hypothesis is the model fits the data. In other words, the further from statistical significance the Chi-square value, the better the fit. In this case, p = .206 for the 1-factor model and p = .096 for the 2-factor model, so the 1-factor model is a better fit overall. In short, the assumption of unidimensionality appears satisfied, a finding similar to that of the original analysis.

4.6.2 Test of Local Item Independence

To test for local item independence, the three testlets on the test (Sections I, II, and III – see Appendix A) were modeled using SCORIGHT 3.0, software created by Wang et al (2005) specifically to model testlet effects. Table 25 shows the variance in item scores accounted for by examinee-testlet interactions ($\gamma_{id(j)}$).

Table 25. Estimated gamma variances

Testlet	Variance	S.E.
1	0.7583	0.5172
2	0.4069	0.1614
3	0.5899	0.4810

As shown, all three testlets exhibit a large testlet effect, with the first and third testlets exhibiting especially strong effects. Over 75% of the variance in examinee scores on Testlet 1 is due to examinee-testlet interaction, meaning less than 25% of the variance in Testlet-1 scores can be attributed to examinee ability. Though not quite as extreme for the other testlets, nearly 60% of the variance in Testlet-3 scores and over 40% of the variance in Testlet-2 scores is attributable to this interaction. For this reason, it would seem inappropriate to apply the 2PL and 3PL IRT models to these data,[18] the same conclusion reached with the original data set.

[18] Because the Rasch-1PL model is confirmatory, it could be applied to these data despite the large testlet effects, although application of the Rasch model presents other problems as discussed later.

5 Data Analysis III (Middle 80%)

5.1 Participants

A second resampling of the original data was conducted to further validate the generalizability of the findings presented in Chapter 3. In this case, resampling involved removal of the upper 10% and lower 10% of the candidate pool (based on raw scores). Removing the top 10% ensures a new cut score will be established (since the top 20% of the original examinee pool would have been offered admission based on the 2006 quota – 462/2,320), while removing the bottom 10% helps restore symmetry to the distribution (to better approximate a normal distribution). The motivation for choosing this subset of the examinee population is to determine how classification consistency would change if there were a more restricted range of ability among the examinees, as well as forcing the cut score more toward the center of the distribution to determine what impact, if any, doing so would have on the classification accuracy of different item-analysis methods.

5.2 Descriptive Statistics

After removal of the upper 10% and lower 10% of examinees from the data, the new sampling includes 1,856 of the original 2,320 examinees. Table 26 displays the descriptive statistics of the original data set and this second resampled data set.

Table 26. Descriptive statistics of the original data and second resampling

N	Min	Max	Mean	Std Dev	Skewness		Kurtosis	
					Statistic	SE	Statistic	SE
1856	27	48	37.350	5.520	.003	.057	-.933	.114
2320	5	65	37.303	8.254	-.064	.051	-.080	.102

As shown, the range of scores for this candidate pool is considerably narrower than that of the original candidate pool, with the minimum score 27 and a maximum of 48. The means are very similar (37.350 vs. 37.303), but the standard deviation of this restricted-range data is much narrower (5.520 vs. 8.254), suggesting about two-thirds of all scores fall within a very tight range (between 32 and 43) and 95% of the scores fall between 26 and 48.[19] Interestingly, the distribution exhibits virtually no skew (Skewness = .003), but it is considerably more platykurtic than the original distribution (Kurtosis = -.933), with fewer scores near the peak and in the tails than would be found in a normal distribution of scores. Figure 36 depicts the histogram of this data set.

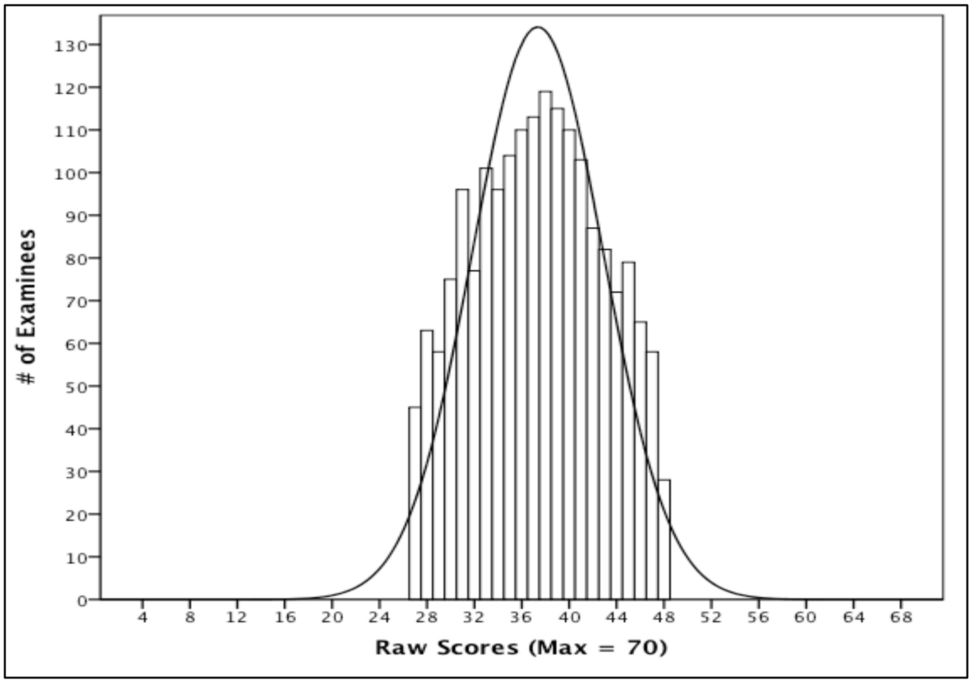


Figure 36. Distribution of examinee scores for the second resampling

[19] In reality 100% of the scores fall between 27 and 48, as they are the minimum and maximum scores of the distribution.

5.3 Cut-Score Determination

Once again, the same protocol for cut-score determination was used with these data as with the previous two sets of data. As a reminder, the annual quota (Q) for $2006 = 150$ and the historical acceptance rate $(\widehat{\mathrm{r}}) = 37\%$, so the target number of admissions offers (λ) was 405 $(\lambda = \frac{150}{0.37})$. To determine the 405th highest score, raw scores obtained from the actual administration of the test were compiled in Excel and sorted from highest to lowest. The resulting ranking indicates that the 405th highest-ranking examinee achieved a raw score of 42, a score achieved or exceeded by the top 471 examinees. The next higher cut score, 43, was achieved by 384 candidates, 21 fewer than the target number of admission offers. Based on the historic pattern of admissions offers at this university, it is assumed 471 examinees would be offered admission rather than the fewer number of 384. Figure 37 portrays the location of the cut score along the distribution.

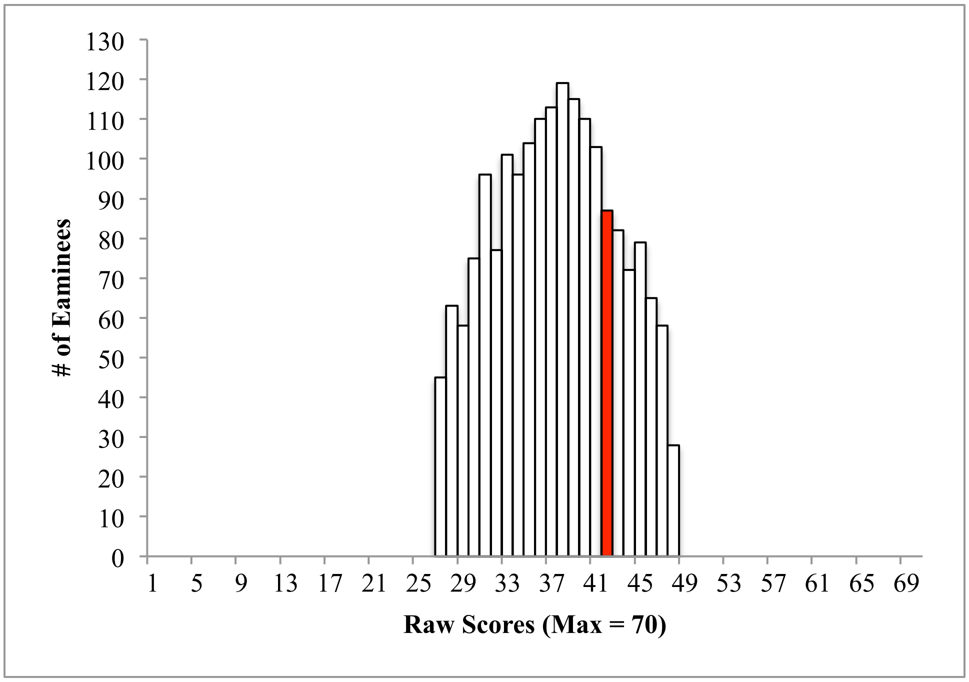


Figure 37. Cut score (42) along the distribution

5.4 Item Analyses

To help identify similarities and differences in the data across examinee populations, the same set of item analyses performed on the original data set was performed on this set of resampled data. Following are the findings of these analyses.

5.4.1 CTT (Kelley)

Once again employing the test validation method introduced by Kelley (1939), an item analysis was conducted comparing the performance of the upper 25% of the distribution against that of the lower 25%. Table 27 contains a sample of the items analyzed using Kelley's discrimination index.

Table 27. Excerpt of item analysis based on Kelley's D

Item	UG	LG	Total	*p*	D	Flag
1	352	321	1271	68%	12%	F
2	277	133	750	40%	33%	
3	250	136	720	39%	27%	
4	242	134	715	39%	25%	
5	226	175	737	40%	14%	F
6	263	126	746	40%	31%	
7	358	253	1190	64%	27%	
8	281	130	773	42%	34%	
9	304	191	953	51%	27%	
10	236	150	708	38%	21%	
11	381	232	1186	64%	36%	
12	343	245	1118	60%	25%	
13	201	134	610	33%	17%	F
14	175	126	577	31%	13%	F
15	316	211	1016	55%	26%	
16	232	180	742	40%	14%	F
17	93	46	250	13%	11%	F
18	231	197	813	44%	11%	F
19	421	350	1425	77%	21%	
20	356	251	1179	64%	27%	

The second column of Table 27 contains the number of upper-group examinees who answered the item correctly (out of 471: 464 + 7 ties at the lower bound), the third column contains the number of lower-group examinees who answered the item correctly (out of 515: 464 + 51 ties at the upper bound), and the fourth column contains all those who answered the item correctly (out of all 1,856 examinees). The fifth column represents the item difficulty (p) and the sixth column represents the discrimination (D) between the upper and lower groups. Items were flagged when they exhibited low discrimination values ($\mathrm{D} < 0.20$) and/or when items were excessively easy or difficult ($p \leq 0.20$ or $p \geq 0.80$). Using these criteria, 35 of the 70 items were flagged, a full half of the test (see Appendix D.1 for the full item-analysis table).

5.4.2 CTT (Point Biserial)

The second analysis performed on this set of resampled data is examination of the point-biserial correlations between individual test items and the overall test. Recall that while there is no firm benchmark for acceptable levels of point-biserial correlation, convention within the language-testing field suggests a correlation of at least 0.20 is acceptable.

Table 28. Excerpt of item analysis based on point-biserial correlations

Item	rpb	Flag
1	0.10	F
2	0.27	
3	0.22	
4	0.19	F
5	0.10	F
6	0.23	
7	0.21	
8	0.28	
9	0.21	
10	0.16	F
11	0.29	
12	0.20	
13	0.14	F
14	0.11	F
15	0.21	
16	0.11	F
17	0.13	F
18	0.09	F
19	0.19	F
20	0.23	

As shown, 10 of the first 20 items were flagged for having item-test correlations below 0.20. Overall, a staggering 42 of the 70 items were flagged for having insufficient point-biserial correlations (see Appendix D.2).

5.4.3 Rasch/1PL

Item analysis within the Item Response Theory (IRT) framework was also conducted. The first of these analyses was application of the 1-parameter logistic (1PL) model, or more specifically, the Rasch equivalent of the 1PL (see Section 2.1.3.3.1 of the literature review for additional detail). For the Rasch model, the decision criterion for item acceptability is a Z-standard value at or below 2.0, as suggested by Linacre (2002).

Table 29. Excerpt of item analysis for the Rasch/1PL model

		Infit		Exact Match		
Item	b	MnSq	Zstd	Obs%	Exp%	Flag
1	-0.80	1.02	1.21	68.5	68.5	
2	0.40	0.97	-2.83	62.8	60.9	
3	0.47	0.98	-1.26	62.6	62.0	
4	0.48	1.00	-0.36	61.9	62.2	
5	0.43	1.03	2.20	60.4	61.4	F
6	0.40	0.98	-1.60	62.2	61.0	
7	-0.61	0.99	-0.94	64.8	64.4	
8	0.34	0.96	-3.48	63.4	60.1	
9	-0.07	0.99	-1.11	59.6	57.8	
10	0.49	1.01	0.54	61.7	62.4	
11	-0.60	0.96	-2.89	65.6	64.2	
12	-0.44	0.99	-0.47	61.5	61.4	
13	0.73	1.01	0.69	67.0	67.0	
14	0.82	1.02	1.04	68.8	68.8	
15	-0.20	0.99	-1.06	59.9	58.6	
16	0.42	1.02	1.99	59.4	61.2	
17	1.91	1.00	-0.01	86.5	86.5	
18	0.25	1.03	3.31	55.4	59.1	F
19	-1.24	0.99	-0.38	76.9	76.8	
20	-0.58	0.98	-1.30	64.9	63.9	

As shown, only two of the first 20 items of the test were flagged for having excessive Z-standard values. Moreover, only seven of the 70 total items were flagged as faulty by the Rasch/1PL model (see Appendix D.3), an outcome quite surprising given the large number of items flagged by the other models for this data set.

5.4.4 IRT 2PL

The IRT 2-parameter logistic (2PL) model was also applied to the data. As with the other two sets of analyses, the difficulty (b) and discrimination (a) parameter estimates were used as the basis for examining item fit. As a reminder, unacceptably low discrimination (a < 0.30) and/or unacceptably low or high difficulty parameter estimations (b < -3.0 or b > 3.0) were the benchmarks used in the analysis. Table 30 portrays an excerpt of the item analysis findings for the 2PL model.

Table 30. Excerpt of item analysis for the 2PL model

Item	a	b	Flag
1	0.177	-2.559	F
2	0.340	0.680	
3	0.261	1.028	F
4	0.206	1.306	F
5	0.149	1.560	F
6	0.280	0.826	F
7	0.264	-1.351	F
8	0.348	0.571	
9	0.228	-0.185	F
10	0.194	1.417	F
11	0.393	-0.942	
12	0.238	-1.083	F
13	0.206	1.994	F
14	0.185	2.448	F
15	0.235	-0.518	F
16	0.147	1.545	F
17	0.303	3.644	F
18	0.130	1.024	F
19	0.313	-2.342	
20	0.285	-1.210	F

As shown, virtually every item is identified as misfitting when the 2PL model is applied. Sixteen of the first 20 items and 54 of the 70 total items were flagged, an obvious problem with respect to model fit. See Appendix D.4 for the complete item-analysis table.

The Test Information Function (TIF) for the 2PL model is depicted in Figure 38, which as a reminder is a graphical representation of how much information about examinee performance the test is providing at each level of theta. For this resampling of the data, maximum information is 2.899 at θ = -1.350.

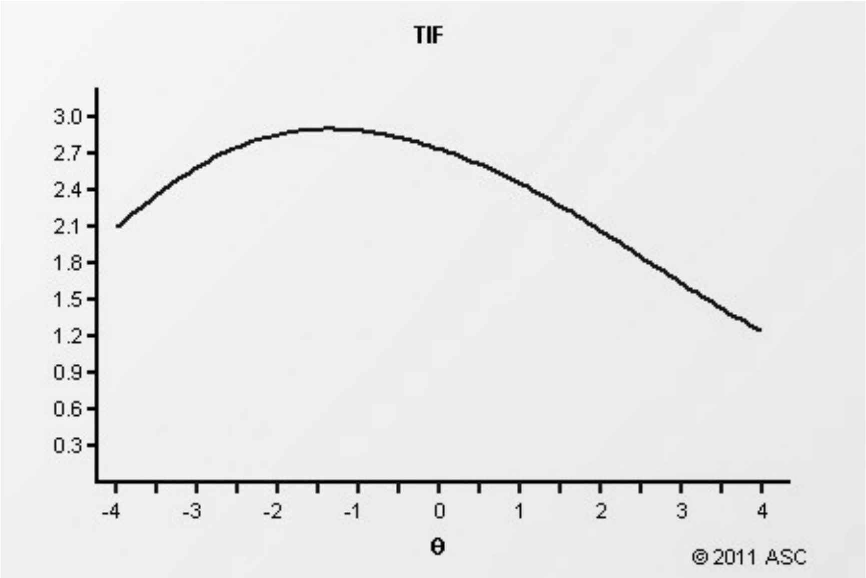


Figure 38. Test Information Function of the 2PL model

Figure 39 portrays the Conditional Standard Error of Measurement (CSEM), which again is the inverse of the TIF. For the 2PL model, the minimum CSEM is 0.587, achieved at θ = -1.350. As shown, there is an enormous amount of error at each level of theta, which of course is not surprising given how poorly the model fits the data.

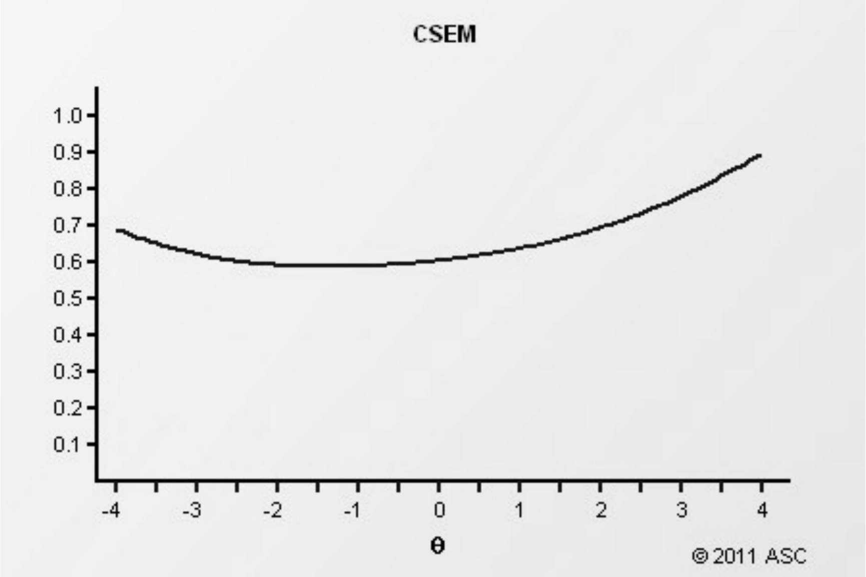


Figure 39. CSEM of the 2PL model

Figure 40 displays a graph of the Test Response Function (TRF) for the 2PL model. As explained in Chapter 3, the TRF portrays the proportion or number of items examinees are predicted to answer correctly at each level of theta. The left Y-axis represents the predicted proportion correct and the right Y-axis represents the predicted number correct at any given level of theta.

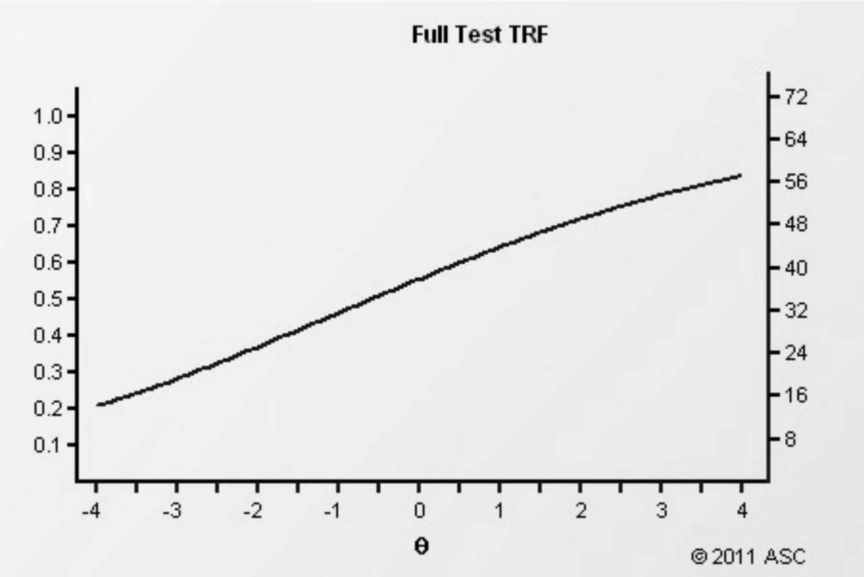


Figure 40. Test Response Function for the 2PL model

As shown, the TRF is further confirmation the 2PL fits the data poorly. There is very little discrimination among examinees across the entire distribution, as evidenced by the very flat slope of the function at all levels of theta.

Figure 41 is the scatterplot of the difficulty (b) by discrimination (a) parameters for the 2PL model.

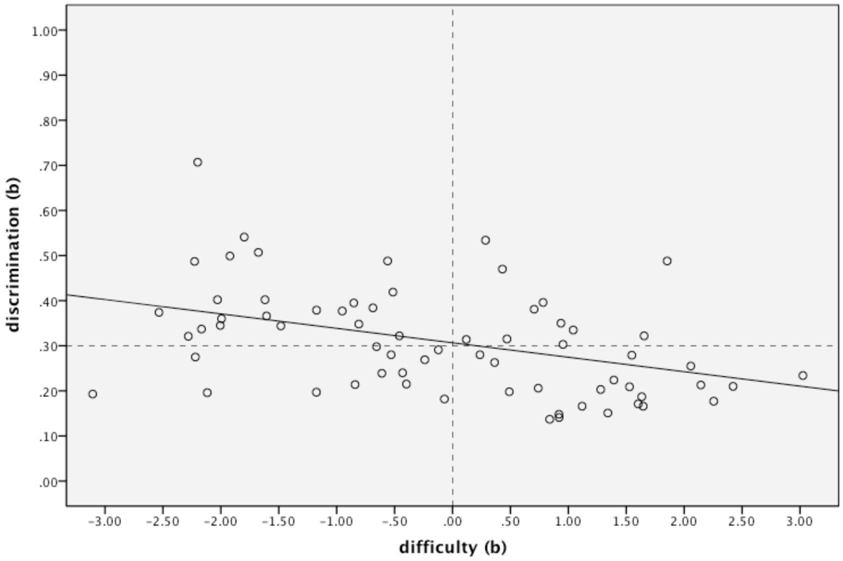


Figure 41. Scatterplot of difficulty (b) by discrimination (a)

Once again, the broken horizontal line inserted into the scatterplot crosses the Y-axis at the minimum acceptable level of discrimination for the model (a = 0.30). As mentioned, 54 of the 70 test items have unacceptably low discriminatory power and/or are unacceptably easy or difficult when the 2PL model is fit to the data.

The broken vertical line that crosses the X-axis at $\theta = 0$ was added to break the graph into quadrants to further illustrate how items are functioning in the 2PL model. Of particular note once again is item 36 (the most extreme outlier in the upper left quadrant). As was the case in the analysis of the full data set and this restricted-range data set, Item 36 exhibits the greatest discriminatory power of all the items (a = .707) despite being one of the easiest (b = -2.200). Moreover, easier items in general exhibit greater discriminatory power than more difficult items, which is evidenced by the slight negative correlation between the two parameters (r = - 1.93). It also concurs with the theta where maximum information is obtained (θ = -1.350). Given that the cut score for the 2PL model for these data occurs near θ = 0.604, this bias toward more information being available at the easier end of the difficulty contiuum continues to be problematic. This finding is confirmed by the distribution of b-parameter estimates as shown in Figure 43.

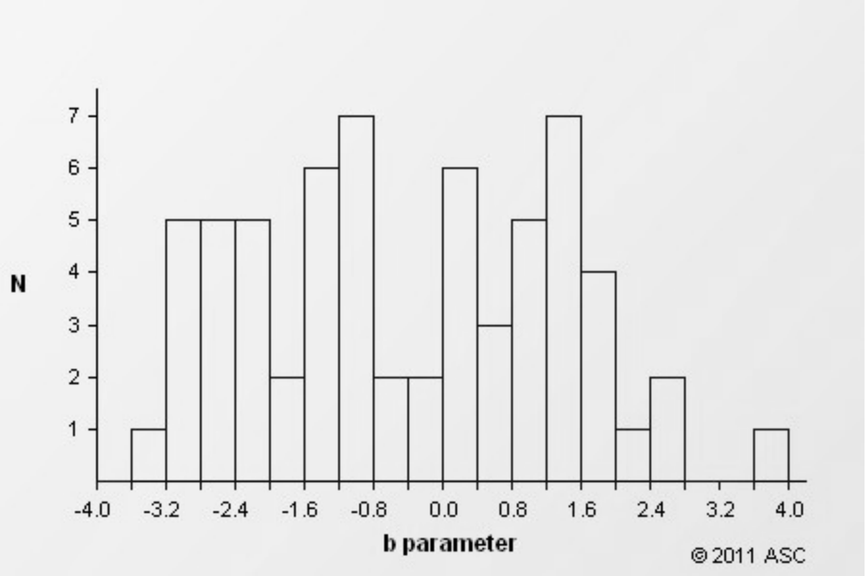


Figure 42. Histogram of the b-parameters of the 2PL model

As shown, the majority of items are in the left half of the distribution, confirming the fact that maximum item information can be found at lower levels of theta (i.e., among the easier items on the exam).

5.4.5 IRT 3PL

The final model applied to the restricted-range data set is the IRT-3PL model. Table 31 portrays an excerpt of the item analysis findings for this model application.

Table 31. Excerpt of item analysis for the 3PL model

Item	a	b	c	Flag
1	0.207	-0.696	0.275	F
2	0.604	1.312	0.225	
3	0.483	1.814	0.235	
4	0.440	2.162	0.250	
5	0.405	2.636	0.287	K
6	0.505	1.600	0.234	
7	0.357	-0.110	0.259	
8	0.581	1.250	0.225	
9	0.356	1.071	0.254	
10	0.444	2.255	0.257	
11	0.515	-0.115	0.248	
12	0.331	0.269	0.261	
13	0.544	2.515	0.250	
14	0.510	3.060	0.254	F
15	0.369	0.737	0.257	
16	0.377	2.745	0.285	K
17	0.838	3.359	0.138	F
18	0.279	2.624	0.271	F
19	0.374	-1.272	0.266	
20	0.370	-0.054	0.259	

As a reminder, the 3PL model estimates three parameters: difficulty (b), discrimination (a), and pseudo-guessing (c). In this case, low discrimination estimates (a < .30), excessively easy or difficult items (b < -3.0 or b > 3.0), and/or excessively high guessing-parameter estimates (c > .40) all trigger flags. In addition, some items that have highly unexpected response patterns can trigger a flag, denoted by K, for a possible keying error.

As shown in Table 31, six of the first 20 items were flagged when the 3PL model was applied, two of which for possible keying errors (Items 5 and 16). This is actually not the case, thereby indicating badly misfitting items. It appears the correlation for each is very low, presumably because the items are quite difficult and just could not correlate well within ability parameters because of such a restricted range of abilities. See Appendix D.5 for the complete item-analysis table.

As with all IRT analyses, additional information is available to describe/portray model fit. Figure 43 is a graph of the Test Information Function (TIF) for 3PL model. Recall the TIF is a graphical representation of how much information the test is providing at each level of theta. Maximum information is 4.178, achieved at θ= 2.050.

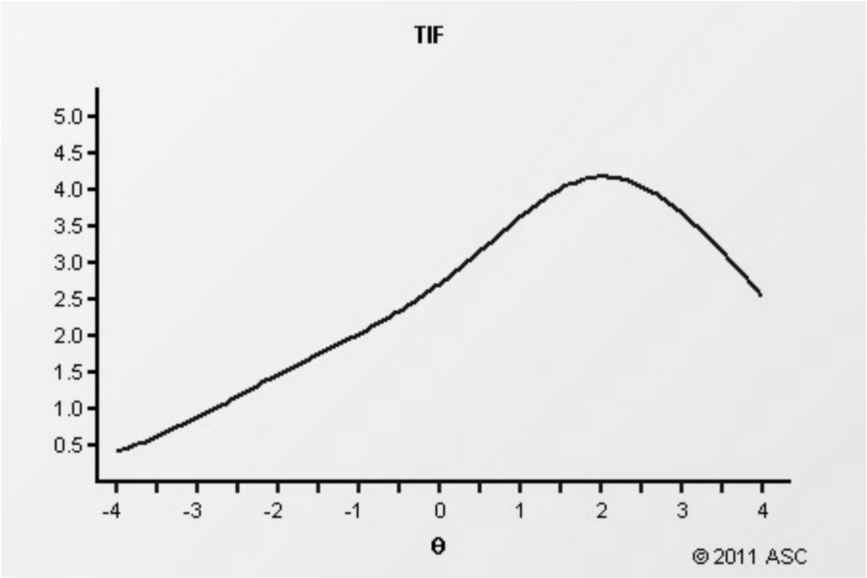


Figure 43. Test Information Function of the 3PL model

The inverse of the TIF is the Conditional Standard Error of Measurement (CSEM) Function, and it illustrates the amount of error in theta estimation for each level of theta. In this case, the minimum CSEM is 0.489, achieved at θ= 2.050, as shown in Figure 44.

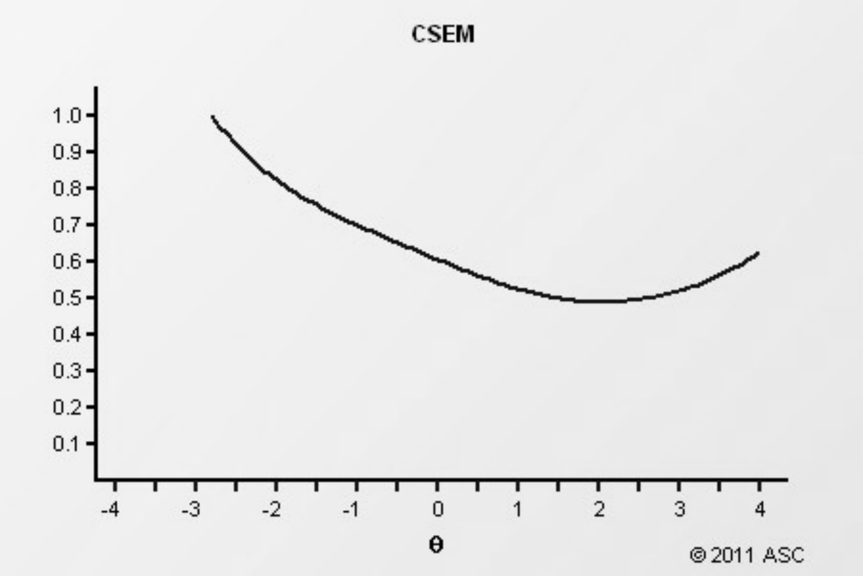


Figure 44. CSEM of the 3PL model

Figure 45 is a graph of the Test Response Function (TRF) for 3PL model. Recall that the TRF predicts the proportion/number of items examinees will answer correctly as a function of theta. The left Y-axis represents the anticipated proportion correct while the right Y-axis represents the anticipated number correct.

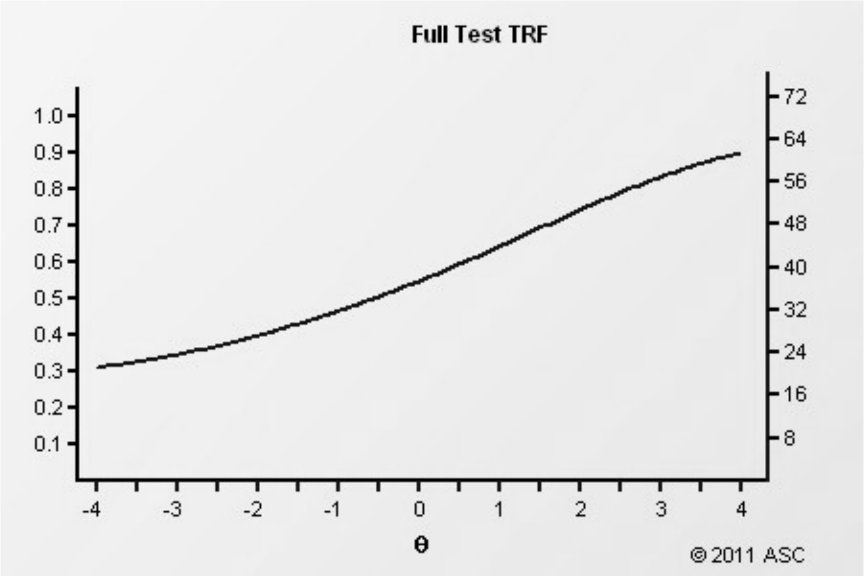


Figure 45. Test Response Function of the 3PL model

As is the case with the 2PL model, the TRF illustrates that the 3PL model does not fit the restricted-range data as well as it did the other two data sets. For example, the slope of the TRF is comparatively flat (c.f., Figure 25). Nevertheless, the model does appear to discriminate best toward the upper end of the theta continuum, a finding confirmed by the theta at which maximum information was achieved (θ = 2.050). This is generally positive since the cut scores across analyses tend to be at high levels of theta.

To investigate model fit from a different perspective, Figure 46 depicts the scatterplot of the difficulty (b) by discrimination (a) parameter estimates.

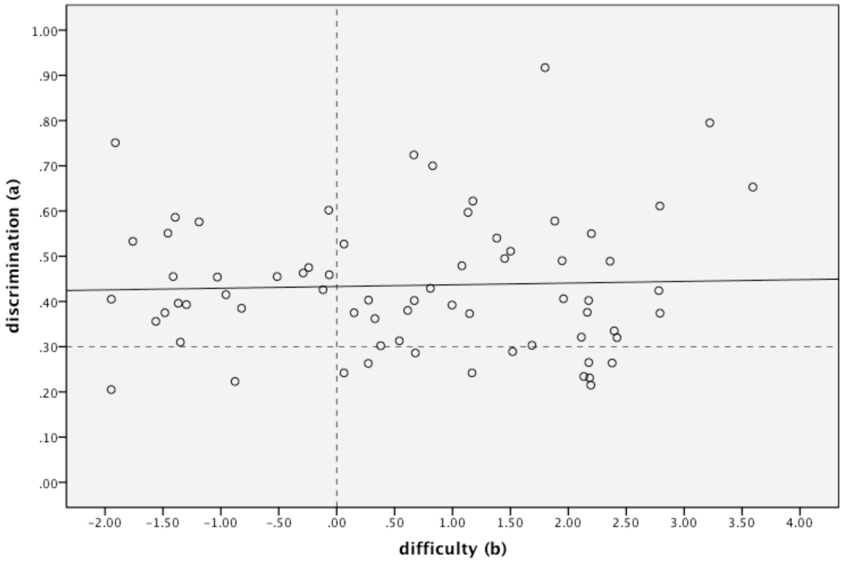


Figure 46. Scatterplot of difficulty (b) by discrimination (a)

To maintain continuity with the other scatterplots contained in this dissertation, a broken horizontal line was inserted into the scatterplot that crosses the Y-axis at the minimum level of discrimination considered acceptable (a = 0.30). The broken vertical line that crosses the X-axis at $\theta = 0$ was added to break the graph into quadrants to further illustrate how items are functioning in the 3PL model. As shown, many more items fall below the level of acceptable discrimination in this application of the 3PL than in the previous two. Nevertheless, a greater number of well-discriminating items are relatively difficult, which again is desirable given the location of the cut score (θ = .715). The slightly positive slope of the regression line (the solid line, r = 0.01) further confirms the finding that more difficult items are discriminating better than easier items. This differs from the 2PL model, where a greater number of easy items discriminate better. A comparison of Figures 42 and 47 demonstrate this trend between the two models.

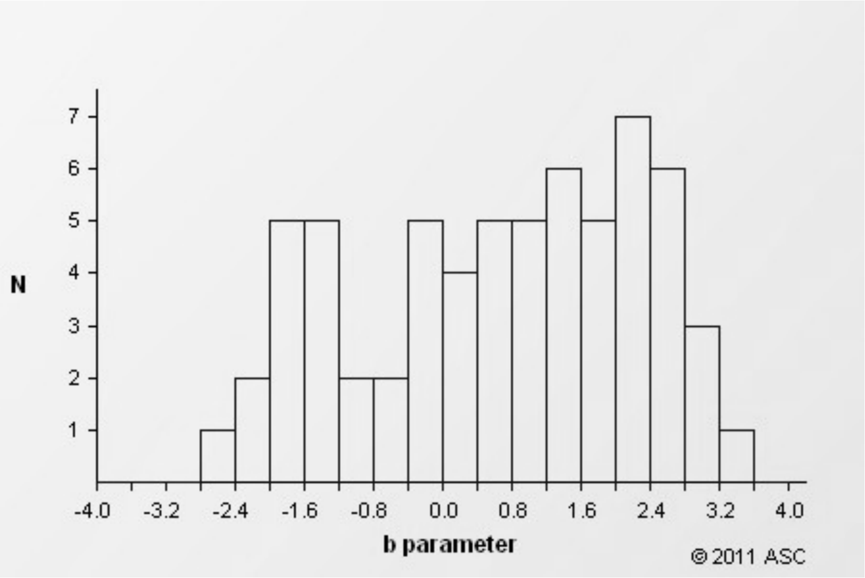


Figure 47. Histogram of the b-parameters of the 3PL model

5.5 Summary of Analyses

In an effort to synthesize the respective outcomes of each item analysis, Table 32 is a summary of the items flagged as faulty according to each model's criteria.

Table 32. Summary of the items flagged across models

Model	CTT (Kelley)	CTT (Pt Bis)	Rasch/1PL	2PL	3PL
Flagged Items	1	1		1	1
				3	
		4		4	
	5	5	5	5	5
				6	
				7	
				9	
		10		10	

				12	
	13	13		13	
	14	14		14	14
				15	
	16			16	16
	17			17	17
	18		18	18	18
				20	
	21	21		21	
	22	22		22	22
	23	23		23	
	24	24		24	
		25		25	25
	26	26		26	26
		27		27	
	28				
		29		29	
		30		30	
	31	31		31	31
	32	32	32	32	
	33	33		33	
	34	34	34	34	
	35	35	35	35	
	36	36		36	
	37	37	37	37	
		38		38	
	39	39		39	
	40	40		40	40
	41	41		41	
	42	42		42	
				45	
	46	46		46	
		47		47	47
	47	48		48	48
	49	49		49	49
	50				
		51		51	
		52			
		54			
	55	55		55	
	57	57	57	57	57
	58	58		58	58
	61				
				62	
		63		63	
				64	
	65	65		65	65
	66	66		66	66
		68		68	
	69	69		69	69
				70	
# Flagged	35	42	7	54	19

As shown, the two Classical Test Theory (CTT) item-analysis methods identified half or more of the items as faulty. Application of the 2PL model led to even greater misfit, with a staggering 54 of 70 items flagged, almost all of which were due to insufficient discrimination parameter estimates. Interestingly, the Rasch model only identified seven items as misfitting, which is the fewest not only among the models applied to these data, but among all model applications across all data sets. While there could be other reasons this is the case, it is likely due to the fit-decision criterion (Z-std > 2.0). As shown in Table 30, Items 2, 8, and 11 also "misfit" the model in that they fit too well (are overly predictable). It appears that if items were flagged for both underfitting and overfitting the model (Z < 2.0 and Z > 2.0) as is the case with other models, the number of flagged items would greatly increase for the Rasch/1PL model.

Regardless, the trend of much large numbers of items flagged with these data indicates model fit quickly broke down when the range of scores became narrower. This is particularly troubling considering admit-reject decisions for this entrance exam revolve around a very narrow slice of the distribution, far narrower in range than the range of data included in this second resampling analysis. Recall that classification accuracy really is only a function of a narrow band of scores along the distribution, given that the majority of examinees will either "pass" (be admitted) or "fail" (be rejected) regardless of the item-analysis method employed because of how far their scores are from the cut line. This substantial degradation of model fit as range narrows will be discussed in Chapter 6 as it relates to the classification-accuracy decisions to be made for these data.

As delineated in Chapters 3 and 4, another means of model comparison is examination of the cross-tabulation values of each model pair. As a reminder, cross-tabulation values indicate the level of agreement (classification consistency) between model rankings of examinee scores. Table 33 is a summary of the cross-tabulations for each of the models relative to the baseline raw scores for this restricted-range data set.

Table 33. Classification consistency across models

		Raw Score		Classification Consistency
		Admit	Reject	
CTT (Kelley)	Admit	375	58	91.7%
	Reject	96	1327	
CTT (PtBis)	Admit	347	58	90.2%
	Reject	124	1327	
Rasch/1PL	Admit	384	75	91.3%
	Reject	87	1310	
IRT 2PL	Admit	365	40	92.1%
	Reject	106	1345	
IRT 3PL	Admit	383	22	94.1%
	Reject	88	1363	

As shown, each alternate model classifies 90-94% of the examinees consistently, which is about 3-5% lower than was the case with the other two data sets (see Tables 11 and 22). In other words, anywhere from 110 (in the Raw-3PL comparison) to 182 (in the Raw-Point-Biserial comparison) of the 1,856 examinees are classified differently. More specifically, the alternate-model outcomes would lead to a rejection of 87 to 124 more examinees than would the raw-score method, while they would lead to the admittance of 22 to 75 more examinees than would the raw-score method.

One other comparison of classification consistency between model pairs considered here is the Kappa statistic. Table 34 illustrates the Kappa values for each pairwise comparison.

Table 34. Kappa statistics of each pairwise model comparison

Pairwise Comparison	Statistic	Std Dev	Approx T	Approx sig
Raw-CTT (Kelley)	0.775	0.017	33.437	0.000
Raw-CTT (Pt Bis)	0.751	0.018	32.297	0.000
Raw-Rasch/1PL	0.768	0.017	33.074	0.000
Raw-IRT 2PL	0.782	0.017	33.864	0.000
Raw-IRT 3PL	0.836	0.015	36.189	0.000

As shown, all pairwise comparisons exhibit statistically significant Kappa values ($p < .001$), but unlike the other two sets of analyses, all but the IRT-3PL model has a Kappa value below what is considered acceptable ($\kappa \geq 0.8$). Note too that the Rasch/1PL model exhibits poorer agreement than the IRT-2PL model despite having identified far fewer items as misfitting. Taken together, these Kappa values illustrate the number of problems this restricted-range data set creates for this set of test items.

5.6 Test of IRT Assumptions

Given that the data for this third data analysis is contrived and fails to approximate a normal distribution, the application of some of the item-analysis methods used in this section is probably inappropriate. Nevertheless, to maintain consistency across analyses, the IRT assumptions were tested once again. Below is a summary of the findings of each test.

5.6.1 Test of Unidimensionality

As with the other data analyses, confirmatory factor analysis was run using structural equation modeling (SEM) to determine whether one factor alone could account for the majority of variance among examinee scores. Figures 48 and 49 illustrate the 1- and 2-factor models created, as well as their standardized factor loadings.

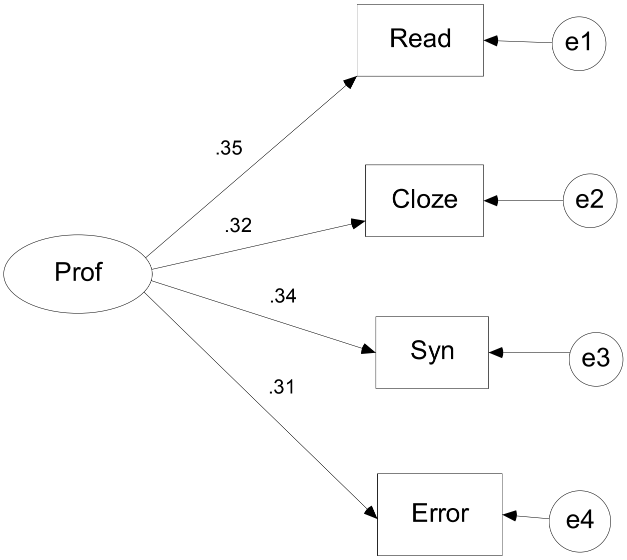


Figure 48. 1-factor SEM of the restricted-range data

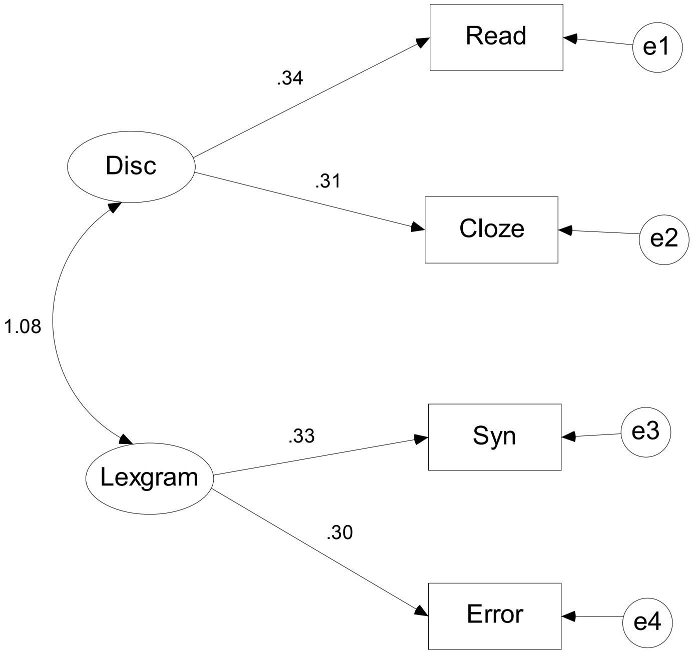


Figure 49. 2-factor SEM of the restricted-range data

Once again, the models fit similarly, indicating there is no advantage to modeling two latent traits (factors) rather than only one. English proficiency (Prof) as a single factor loads equally well (even slightly better) on the four observed measures than do the two hypothesized factors (.35 > .34 on the reading-comprehension section, .32 > .31 on the cloze section, .34 . > .33 on the synonyms section, and .31 > .30 on the error-identification section, respectively). As a result, it seems reasonable to conclude the underlying factor being measured is sufficiently unidimensional to justify application of IRT to the data, at least with respect to this one assumption. Of course, this finding is relative and says nothing about the overall fit of either model. As it happens, both models failed to converge, indicating they fit the data extremely poorly. This is confirmed by their respective Chi-square statistics (CMIN/DF), shown in Table 35.

Table 35. Chi-square comparison of CFA models

Model	NPAR	CMIN	DF	*p*	CMIN/DF
1-Factor	12	9.525	2	.009	4.762
2-Factor	13	9.312	1	.002	9.312

As shown, both Chi-square statistics are statistically significant (p = .009 and p = .002), which means the null hypothesis that the models fit the data must be rejected. This finding is not surprising given how many items the item-analysis methods identified as misfitting, but it remains disconcerting, as will be explained in the next chapter.

5.6.1 Test of Local Item Independence

To complete the analysis of IRT assumptions, SCORIGHT (Version 3.0) was once again employed to examine the magnitude of any testlet effects on the test (see Sections I, II, and III of the test in Appendix A). Table 36 illustrates the testlet effects as reflected by the variance in gamma (a parameter measuring examinee-testlet interaction).

Table 36. Estimated gamma variances

Testlet	Variance	SE
1	0.3647	0.1718
2	0.3248	0.1310
3	0.4274	0.3512

As shown, the gamma variance of each testlet is considerably lower for this restricted-range data than for the other two data sets. While a positive finding on the surface, this is probably due more to the level of noise among the data than a truly smaller testlet effect. Testlet 3 illustrates this point, where the standard error of the variance (SE = .3512) is nearly as large as the statistic itself (Variance = .4274). As a result, it is questionable whether one could claim the assumption of local item independence is satisfied by these data. Regardless, it is clear from all three data sets that the size of the testlets included in the exam (15 items, 20 items, and 15 items, respectively, and 50 of the 70 items overall) is a structural problem that should be addressed by the examination-creation committee, a recommendation explored in greater detail in the next chapter.

6 Discussion

As described in Chapters 3 through 5, a set of five item-analysis methods was applied to three sets of data to determine how test outcomes might change across methods. Following is a discussion of the cumulative findings with respect to the six research questions identified in Section 1.3 of this dissertation.

6.1 Research Question 1

> *Do the item-analysis methods prescribed by Classical Test Theory (CTT) and Item Response Theory (IRT) identify any test items as faulty/misfitting, and if so, do they identify the same items?*

The answer to the first half of this question is yes. Analysis of all three data sets yielded similar results, meaning each item-analysis method flagged several items. However, the answer to the second half of the question is no. As illustrated throughout the analyses, both the number and the nature of items flagged varied quite considerably among the methods. Recall that the IRT-3PL model flagged the least number of items during analysis of the full data set (only 10 of the 70), whereas the IRT-2PL model flagged more than twice as many (23). Moreover, only eight items were flagged by all five models, while several items were flagged by only one of the models. The same is true for the random-80% data set, where the 3PL model flagged nine items overall but the 2PL model flagged three times as many (27). Additionally, only six items were flagged across all of the models, but many more were flagged by only one of the models. Finally, with the restricted-range data set, differences in item flagging varied the most, where only two of the 70 items were flagged by every model despite the fact a staggering 54 items were flagged by the 2PL model. Clearly, model choice greatly influences which items are flagged as faulty.

It is important to note, however, that both the number and nature of items flagged by each model is a function of the decision criteria applied to the output of that model, criteria that are somewhat arbitrary in nature. As explained elsewhere, the language-testing field commonly considers a point-biserial correlation of 0.20 or greater as acceptable, Linacre (2002) suggested a Z-standard infit value of less than 2.0 as acceptably fitting, and Xcalibre has its default flag specifications set to a minimum a-parameter of 0.30, b-parameters between -3.0 and 3.0, and a maximum c-parameter of 0.40. All of these decision criteria lack theoretical motivation and are subject to change. For example, if items that "overfit" the Rasch model (i.e., are too predictable, with Z-standard infit values < -2.0) are also excluded from scoring – a decision criterion as sound theoretically as removing only underfitting items – then an additional 13 items would have been found faulty with the full data set analyzed in this study. As a result, these findings must be tempered by the fact that they are a function of the chosen decision criteria and that model outcomes would change were different criteria applied.

6.2 Research Question 2

Does the identification of misfitting items influence the classification consistency of examinee scores across methods?

While the answer to the first research question is important in itself, if ultimately there is no impact on the ordinal ranking of examinees, there really is no consequence (nor benefit) of applying one particular scoring method instead of another for these particular data. In fact, given that the various methods flagged such different item sets, it could be argued that the raw-score method is as valid as any of the other scoring methods. However, as shown in each of the data-analysis chapters, each item-analysis method identified anywhere from 4-9% of examinees that would have been classified differently had it been employed instead of the raw-score method. This finding is troubling. Despite the widely varying nature and number of items flagged across models, the number of students who would be displaced by each model is similarly large: 105-130 of the 2,320 examinees would be classified differently in the full data set, 74-119 (of 1856) in the random-80% data set, and 110-182 (of 1856) in the restricted-range data set. These displacements result in Kappa statistics ranging from good (e.g., 3PL/Full Data Set: $\kappa = .851$) to relatively poor (e.g., Point-Biserial/Restricted-Range Data Set: $\kappa = .751$), suggesting displacement is indeed a concern with these data sets.

Similar variation in displacement is also evident when the alternate item-analysis methods are compared to each other. Table 37 illustrates the amount of variation present between/among the methods themselves.

Table 37. Cross-tabulation results between/among methods

		CTT (Kelley)		CTT (Pt Bis)		Rasch/1PL		IRT 2PL		IRT 3PL	
		Reject	Admit	Reject	Admit	Reject	Admit	Reject	Admit	Reject	Admit
CTT (Kelley)	Reject	1,874	0	1,829	58	1,840	47	1,815	72	1,827	60
	Admit	0	446	45	388	63	370	99	334	88	345
CTT (Pt Bis)	Reject			1,886	0	1,853	21	1,809	65	1,822	52
	Admit			0	434	50	396	105	341	93	353
Rasch/1PL	Reject					1,903	0	1,823	80	1,835	68
	Admit					0	417	91	326	80	337
IRT 2PL	Reject							1,914	0	1,888	26
	Admit							0	406	27	379
IRT 3PL	Reject									1,915	0
	Admit									0	405

As shown, the IRT 2PL-3PL model comparison exhibits the most consistency of all pairwise comparisons, where “only” 53 examinees would have been displaced (26 would have been admitted with the 3PL method but would have been rejected by the 2PL method, while 27 would have been admitted with the 2PL but not the 3PL). On the opposite end of the spectrum, there is a displacement of 170-171 examinees between the 2PL model and the CTT and Rasch methods, respectively. These findings seem to indicate that the differences in items flagged by each method do indeed impact the classification consistency of examinee scores. This conclusion is reinforced by the Kappa values of the pairwise comparisons of all methods.

Table 38. Levels of agreement between methods

Pairwise Comparison	Kappa
Kelley-Point Biserial	0.855
Kelley-Rasch	0.842
Kelley-2PL	0.751
Kelley-3PL	0.785
Point Biserial - Rasch	0.899
Point Biserial-2PL	0.756
Point Biserial-3PL	0.791
Rasch-2PL	0.747
Rasch-3PL	0.781
2PL-3PL	0.921

As shown in Table 38, the 2PL-3PL pairwise comparison exhibits the greatest level of agreement (κ = .921), while the lowest levels of agreement are between the 2PL model and the other methods (Kelley-2PL: κ = .751, Point Biserial-2PL: κ = .756, Rasch-2PL: κ = .747). The 3PL model also exhibits relatively weak levels of agreement with the other methods (Kelley-3PL κ = .785, Point Biserial-3PL κ = .791, and Rasch-3PL κ = .781), demonstrating that the two IRT models generate substantially different outcomes than the CTT and Rasch methods when applied to these data.

To summarize, the identification of misfitting items not only influences the classification consistency of examinee scores when comparing item-analysis methods to the baseline raw-score method, but also when comparing the methods to each other. This is an important finding, one that reinforces the importance of the decision regarding which method to apply to a testing context. Stated differently, this finding is of concern because of its implications with respect to wrongly admitting or rejecting examinees due either to the absence of item analysis or to a mismatch of item-analysis method and data.

6.3 Research Question 3

Which item-analysis method is likely to lead to the highest level of classification accuracy given the nature of the test data?

As explained, there is clear evidence each item-analysis method yields results substantially different results from the raw-score method when it comes to the ordinal ranking of examinees. However, because each of the methods also yield results different from each other, it is necessary to determine which of these methods is likely to yield the greatest classification accuracy for these data. Note, however, that only a narrow slice of the examinee pool is affected by the application of different scoring methods, so classification accuracy really is concerned only with this portion of the candidate pool rather than the pool at large. In other words, the majority of examinees will be admitted or rejected consistently regardless of the scoring method employed, so it is only those examinees whose scores falls close enough to the cut score that they may be admitted in some cases and rejected in others. To get a sense of the number of examinees whose scores might be affected by the choice of item-analysis method, Table 39 portrays two of the most extreme classification discrepancies within the random-80% data set.

Table 39. Examples of examinee misclassification

		Examinee A		Examinee B		Placement Variance
Method	Cut Score	Score	Rank	Score	Rank	
Raw Score	44.0000	43.0000	T-426	45.0000	T-310	116
CTT-Kelley	33.0000	34.0000	T-305	31.0000	T-513	-208
CTT-PB	0.5200	0.8200	T-188	0.2500	T-580	-392
Rasch-1PL	0.4700	0.7800	T-222	0.1800	T-650	-428
IRT-2PL	0.7465	1.1436	243	0.4530	567	-323
IRT-3PL	0.7565	1.1407	210	0.4934	577	-367

As shown, Examinee B would have been admitted using the raw-score method (Raw Score = 45, Cut Score = 44), but Examinee A would have been rejected (Raw Score = 43, Cut Score = 44). Under every other scoring method, however, Examinee B would have been rejected and Examinee A admitted. The displacement in ordinal ranking is non-trivial too, with up to a 428-place difference between the two examinees under the Rasch method, compared to only a 116-place difference under the raw-score method. While these two examinees are among the most extreme cases, 300-400 candidates attained scores close enough to the cut score where their fates rest solely on the choice of scoring method. Table 40 illustrates an example of this method-choice variance.

Table 40. Example of examinee classification variance

		Examinee C	
Method	Cut Score	Score	Rank
Raw Score	44.0000	43.0000	T-426
CTT-Kelley	33.0000	32.0000	T-431
CTT-PB	0.5200	0.4300	T-429
Rasch-1PL	0.4700	0.5700	T-342
IRT-2PL	0.7465	0.7150	421
IRT-3PL	0.7565	0.7733	399

As shown, Examinee C would be rejected if the raw-score method, either of the CTT methods, or the IRT-2PL model were employed but admitted if either the Rasch-1PL model or IRT-3PL model were employed. This is very disconcerting, particularly because several hundred examinees would be classified differently depending on the scoring method employed. As a result, it is critical administrators identify and employ the scoring method most likely to lead to the highest classification accuracy for this testing context. Unfortunately, no language-testing body provides any guidance on how administrators can make this decision, and because many of these administrators are not language-testing experts themselves, they are ill equipped to make the determination on their own.

Therefore, while it is important to note there is no definitively correct answer with respect to which method would achieve the highest classification accuracy for this particular context, there are defensible criteria against which each method can be evaluated. Below is a description of the decision-making process applied here to illustrate which of the item-analysis methods previously described would likely to lead to the highest classification accuracy for this testing context. In brief, pairwise comparisons are made in sequence from the least-sophisticated scoring method to the most-sophisticated scoring method, with a decision being made at each stage until only one method remains.

6.3.1 Raw-Score vs. CTT (Kelley)

As illustrated throughout this study, the raw-score method lacks validity despite the fact it is arguably the most widely used scoring method in the language-testing field. Validity aside, the lack of granularity in the raw-score scale is problematic for these data in terms of admissions overflow. Recall that strict admissions criteria are in place to ensure students have adequate access to faculty, libraries, cafeterias, and other on-campus facilities. Because of the crude scale of the raw-score method, the university must offer admission to more examinees than desired to avoid a shortage of enrollees, but doing so could result in accreditation sanctions. Equally importantly, offering admission based on raw scores will likely lead to a reduction in the overall quality of instruction, not only because larger classes/higher student-to-teacher ratios reduce the amount of instruction available to each student, but also because faculty will likely have to accommodate less-qualified students during class, thereby reducing the amount of level-appropriate guidance and instruction available to examinees who entered the university fully qualified. In short, the application of a more fine-grained scoring method would not only be more defensible theoretically, it would if nothing else yield an important practical benefit. For this reason, Kelley's D would be preferred to the raw-score method.

6.3.2 CTT (Kelley) vs. CTT (Point-Biserial)

As application of the CTT-Kelley method is a vast improvement over the raw-score method, so too is the application of point-biserial correlations over Kelley's D. The primary reason for this improvement is because the entire population sample is included in the analysis when calculating point-biserial correlations rather than only 50-54% of the sample when Kelley's D is calculated (the upper and lower 25-27% groups of the distribution). At the very least, including the entire examinee population in the calculation of item effectiveness has greater face validity, but it also seems mathematically more valid because more data are included in the analysis. The only potential downside to choosing point-biserial calculations over Kelley's D is that they are more complex and time-consuming to calculate when calculated manually. However, there are numerous relatively inexpensive user-friendly software programs (e.g., SPSS, SAS) that can calculate point-biserial correlations quickly. As a result, using point-biserial correlations as the basis for item removal would be preferable to Kelley's D for these data (and really any data where there is one continuous variable and one dichotomous variable).

6.3.3 CTT (Point-Biserial) vs. Rasch/1PL

Because IRT models can extract a greater amount of item information from data than simpler item-analysis methods like those within CTT, it would seem they would always be superior to CTT methods. However, the Rasch/1PL model in particular poses a challenge to this general rule because of its confirmatory perspective. As a reminder, Rasch models require data to fit the model, so any misfitting data should be discarded from the analysis. The problem for this particular testing context is that the discarding of data would include not only misfitting items but also misfitting examinees, of which there are over 100 in the full data set. This poses two problems. The first is how these examinee scores should be evaluated with respect to all of the other examinee scores given that it would be unethical to remove them from admissions consideration merely because they misfit the Rasch model. The second problem is the philosophy of the Rasch model itself. As Yen (2006) explained, many test developers believe it is inappropriate for a scaling model to be the driving force in test construction, as is the case with the Rasch model (p. 124). In this case, the Rasch philosophy certainly seems to go against the collective belief of the test developers given how much time university faculty spend creating the exam year after year. As a result, it would seem unlikely these stakeholders would accept the Rasch model as a legitimate scoring method, thereby leaving point-biserial correlations as a better alternative despite the additional item information gleaned from application of the Rasch model.

6.3.4 CTT (Point-Biserial) vs. IRT 2PL

Because the two- and three-parameter logistic IRT models are exploratory rather than confirmatory, they lack the same ethical concern regarding the treatment of misfitting persons. Instead, it is merely a question of which, if either, fits the data well enough to justify their added complexity over relatively straightforward point-biserial correlations. In the case of the 2PL model, this potential gain in information does not appear to justify its application over point-biserial calculations. As shown in Figures 11 and 12 (pp. 88-89), for instance, maximum information is available at a low level of theta for the 2PL model (Figure 11), but the cut score is at a much higher level of theta, where the conditional standard error of measurement is quite high (Figure 12). As a result, even though the 2PL model seems to fit the data reasonably well overall, it fits relatively poorly near the cut score, thereby diminishing its classification accuracy for the 300-400 candidates whose scores are impacted most by the choice of scoring method. For this reason, point-biserial correlations remain the best item-analysis choice for these data.

6.3.5 CTT (Point-Biserial) vs. IRT 3PL

Unlike the 2PL model, which has a lot of error around the cut score, the 3PL model fits the data particularly well near the cut (see, e.g., Figures 16 and 17 on pp. 93 and 94, respectively). As a result, the 3PL model seems superior to point-biserial correlations as the basis for item removal given all of the additional information gleaned from its application. This makes sense theoretically because the 3PL model was designed specifically to model dichotomously scored items that do not penalize examinees for guessing. More specifically, the 3PL models difficulty, discrimination, and guessing, all parameters likely to contribute to the item-scoring patterns of examinees taking a multiple-choice test that does not penalize guessing.[20] The only potential drawbacks of applying the 3PL model here would be insufficient data, which is a non-issue, and the inability to compute and interpret the results. As explained elsewhere, however, modern software programs like Xcalibre 4.1 make barriers to use minimal, even for the uninitiated. In short, then, the IRT-3PL model seems to be the item-analysis method that will result in the highest classification accuracy for this particular test. The only question that remains is whether the application of *any* IRT model is appropriate for these data, which is the focus of the research questions addressed in the following two sections.

6.4 Research Question 4

Can a test designed to measure English language proficiency satisfy the IRT assumption of unidimensionality?

Given the classification-accuracy analysis in the preceding sections, it would appear the IRT-3PL model is the most appropriate method to employ among the five evaluated. Recall, however, that IRT modeling requires two very strong assumptions to be satisfied for their application to be considered valid. The first assumption is unidimensionality, meaning only a single latent trait is being measured. As shown in all three analyses (see, e.g., Figures 20 and 21, p. 100), it seems reasonable to assume unidimensionality is satisfied, so the answer to this research question is yes. That is, it is reasonable to assume a latent trait loosely termed *English Proficiency* is the construct being measured by the test. As a result, application of the 3PL model remains valid with respect to this assumption.

[20] On some exams, examinees are penalized for guessing, where, e.g., an additional .25 points is subtracted from incorrect answers as opposed to receiving only a score of 0 for unanswered items.

6.5 Research Question 5

Can a test that contains testlets satisfy the IRT assumption of local item independence?

Unlike unidimensionality, the second IRT assumption – local item independence (LII) – clearly was not satisfied by these data. Across all of the analyses, testlet effects account for 40-75% of the total variance in examinee scores. It is for this reason employing the 3PL model is unjustified with these data, despite all of its advantages with respect to modeling long dichotomously scored tests taken by large examinee pools. Fortunately, there is a comparable item-analysis method available, that being Testlet Response Theory (TRT) analysis. Recall from Section 2.1.3.4.2.1 that TRT analysis was created to address violations of local item independence by modeling the examinee-testlet interaction among items within a testlet. Mathematically, the equation is written as:

$$t_{ij} = a_j\left(\theta_i - b_j - \gamma_{id(j)}\right)$$

where t_{ij} represents the TRT-based ability of examinee i on item j, a_j is the discriminating power of item j, θ_i is the theta-ability of examinee i, b_j is the difficulty of item j, and $\gamma_{id(j)}$ is the *testlet effect* (the interaction between examinee i and item j of testlet d). By definition, $\gamma_{id(j)} = 0$ for all independent items. Therefore, $\gamma_{id(j)}$ is the set of parameters that differentiates the TRT model from the IRT-3PL model. Otherwise, the two models are mathematically equivalent.

While TRT analysis does have the disadvantage of resulting in the loss of information (see Figure 1, p. 36), this is of less concern for this testing context because the exam is single use and single administration, and because the ordinal ranking of examinees is the only consideration. In other words, the resulting information loss is a reasonable sacrifice given that the examinee-testlet interaction of each testlet is modeled, which is of vital importance given how dramatically these interactions influence testlet scores. Table 41 contains an excerpt of the TRT item analysis performed on the full data set to illustrate how its outcome differs from the other item-analysis methods.

Table 41. Excerpt of item analysis for TRT

Item	a	b	c	Flag
1	0.343	-1.209	0.206	
2	0.669	0.899	0.059	
3	0.524	1.251	0.049	
4	0.550	1.657	0.127	
5	0.826	2.976	0.304	
6	0.954	1.167	0.163	
7	0.665	-0.824	0.066	
8	1.007	0.746	0.109	
9	0.642	0.107	0.060	
10	0.756	2.294	0.240	
11	0.899	-0.620	0.060	
12	0.723	-0.309	0.133	
13	1.723	1.776	0.238	
14	0.475	2.572	0.090	
15	0.445	-0.180	0.060	
16	0.491	2.883	0.202	
17	1.164	2.895	0.085	
18	0.300	1.896	0.099	
19	0.759	-1.627	0.051	
20	0.617	-0.769	0.046	

Of immediate note is the fact that not a single item of the first 20 was flagged for misfitting the model. All of the a-parameter estimates are acceptable ($a \geq .30$), as are all of the b- and c-parameters ($-3.0 < b < 3.0$ and $c < .40$). Overall, only seven of the items were flagged as misfitting the TRT model, two of which had unacceptably low discrimination parameter estimates (Items 34 and 35), while the other five (Items 37, 39, 41, 42 and 58) had b-parameter estimates that lay outside the acceptable range of difficulty. See Appendix B.6 for the complete TRT item-analysis table.

By way of comparison, recall that the 3PL model identified as faulty 10 of the 70 items when it was applied to the full data set. Interestingly, some but not all of the items flagged as faulty by the TRT model are the same as those identified by the 3PL model. For example, both models flagged Items 34, 35, 37, 41, and 58, but only the TRT model flagged items 39 and 42. Meanwhile, the 3PL model flagged five other items that the TRT model did not (Items 1, 18, 33, 49, and 57).

Despite the relative similarities in terms of item flagging, their classification consistency with respect to raw scores varied, as shown in Table 42.

Table 42. Classification consistency across models

		Raw Score		Classification Consistency
		Admit	Reject	
3PL	Admit	381	24	95.5%
	Reject	81	1834	
TRT	Admit	364	41	94.0%
	Reject	98	1817	

Note there is a slight drop in classification consistency when the TRT model is compared to the 3PL model. Whereas over 95% of the examinees were classified the same between the 3PL and raw-score rankings, 94% were classified the same between the TRT and raw-score rankings. Put another way, 129 examinees were classified differently under the TRT model, compared to only 105 for the 3PL model. Interestingly, the 3PL and TRT models classified examinees quite differently between themselves, as shown in Table 43.

Table 43. Classification consistency between 3PL and TRT

		3PL		Classification Consistency
		Admit	Reject	
TRT	Admit	348	57	95.0%
	Reject	57	1858	

This finding provides evidence that item-analysis methods not only classify examinees differently when compared to the baseline raw-score ranking, but also among themselves, even when they are similar in design and identify similar items as misfitting. In this particular case, the large testlet effects found in the three exam testlets undoubtedly explain the variance in classification consistency between the TRT and 3PL models. In fact, its is because of the presence of this variance that the TRT model is the one that seems most appropriate to employ with these data rather than the 3PL model.

6.6 Research Question 6

If both IRT assumptions can be satisfied, does classification accuracy improve sufficiently to justify the added computational complexity IRT modeling requires?

As demonstrated, the two IRT assumptions could not be satisfied with these data due to severe local item dependence among the items contained in each testlet. However, the TRT model, which is an extension of the IRT-3PL model, can be employed, so the question is whether the level of complexity inherent in the model increases the validity of test scores enough to justify its use over a simpler item-analysis method like CTT.

While some of the limitations of CTT are of no concern here because the exam is single use and single administration (e.g., population dependence and item dependence), other CTT assumptions are problematic. First is the assumption of item equivalence, where it is assumed each item contributes equally to the measure of examinee ability, given that each correct response earns a score of 1 and the total score is the sum of correct responses. As shown, however, this clearly is not the case (see, e.g., Figure 16, p. 93). The amount of information varies quite markedly across the distribution of ability levels, so it is unreasonable to score an exam where each item is given equal weight in the final score. This is especially true for these data because only a small subsection of the population is ultimately effected by the scoring method. Put another way, extracting the most information possible from the items around the cut score is of utmost importance, so treating all of the items as if they contribute equally is indefensible.

The second problem with resorting to a simpler form of analysis like CTT is its treatment of measurement error, where it is assumed the standard error of measurement (SEM) is constant across all ability levels. As shown throughout the analyses, this is decidedly not the case (see, e.g., Figure 17, p. 94). The error at each level of theta changes dramatically and is especially prevalent in the tails of the distribution, so it is important to minimize error as much as possible around the levels of theta where admissions decisions are made. As illustrated, this is accomplished through the application of more sophisticated modeling procedures like IRT 3PL and TRT.

One final argument that could be made against application of more sophisticated modeling is the lack of familiarity among the university's faculty, both in terms of computation and interpretation. However, this argument holds little credibility because modern software packages (e.g., Xcalibre for IRT, SCORIGHT for TRT) are all user-friendly enough that users with little or no experience can learn how to run them and interpret the data simply by reading the respective user manual. Nevertheless, rescoring remains a challenge in this particular context if for no other reason than because Japanese universities traditionally release the full exam (with answer key) to the media soon after the exam is administered so examinees can calculate their scores to get a sense of how well they performed. Depending on the timing of the identification of malfunctioning items and recalculation of scores, implementation of the ILTA Guidelines could result in some candidates being moved off the admissions roster after they have formed an opinion about their chances of being accepted. Put another way, examinees would be understandably disturbed if their admissions fate were changed post hoc. Nevertheless, scoring tests that has several faulty items included is not justifiable. As a result, TRT analysis should be employed over CTT analysis because of the substantial gains in classification accuracy, particularly around the cut score. Assuming this takes place, ample forewarning to all stakeholders, with a full explanation of the changes to scoring and the rationale for such, should help mitigate the consternation that would result from any post-hoc changes to the ordinal ranking of examinees.

7 Conclusion

The impetus for this dissertation was to determine whether choice of item-analysis method would influence the ordinal ranking of examinees who took a large-scale university entrance exam. As the analysis of each data set shows, it clearly does. In all three analyses, anywhere from 4% to 9% of the examinees were classified differently from their original raw-score-based classification. As a result, it seems clear that scoring exams using the raw-score method, especially without any item analysis, is inappropriate.

This finding reinforces the guidance espoused by international testing-bodies like the International Language Testing Association (ILTA) regarding the necessity of item pretesting and/or post-hoc item analysis. Although one can argue about which item-analysis method is most appropriate for this particular testing context, there is no question some form of item analysis is necessary. Regardless of the data set analyzed or the item-analysis method employed, several items were flagged as faulty without fail. Moreover, it is argued that TRT modeling will provide the highest classification accuracy for these data because testlet effects are present, which prevents use of the 3PL model given the clear violation of local item independence.

The only question that remains is whether administrators of this exam are willing to use a scoring method that is not only unfamiliar to the majority of stakeholders, but one that would require the current long-standing scoring system be discarded. Given that these exams are single use and single administration, there would be no repercussions with respect to past exam classifications, and future scores could be linearly transformed onto a scale that is more readily identifiable by the public if that would increase the likelihood of stakeholder endorsement.

Whatever the outcome of this particular exam, the findings of this study make it clear that ILTA and other language-testing bodies need to expand their guidelines to include information about the item-analysis methods available to exam scorers, as well as under what conditions they should be employed. Recall that current ILTA guidelines call for the following procedures with respect to item testing and scoring:

> The work of the task and item writers needs to be edited before pretesting. If pretesting is not possible, the tasks and items should be analysed after the test has been administered but before the results are reported. Malfunctioning or misfitting tasks and items should not be included in the calculation of individual test takers' reported scores. (p. 2)

While these guidelines clearly call for item analysis prior to the final scoring of tests, they remain devoid of guidance with respect to which item-analysis method to employ. As the findings of this study demonstrate, different item-analysis methods generate different sets of items that should be removed before scoring. The ILTA committee therefore needs to expand its guidelines to explain which item-analysis method should be applied to which contexts. Criteria for making such a decision include the number of examinees taking the test, the number of items on the test, the nature of the items (e.g., multiple choice, constructed response), the medium of delivery (e.g., computer adaptive, paper and pencil), the scoring method (dichotomous, polytomous, or a combination of both), the number of dimensions being measured, whether local item dependence is likely due to the presence of numerous/large testlets, and whether the items will be administered only once or on multiple occasions.

Figure 50 is a decision tree that represents an example of the type of delineation the ILTA committee could provide to language testers.

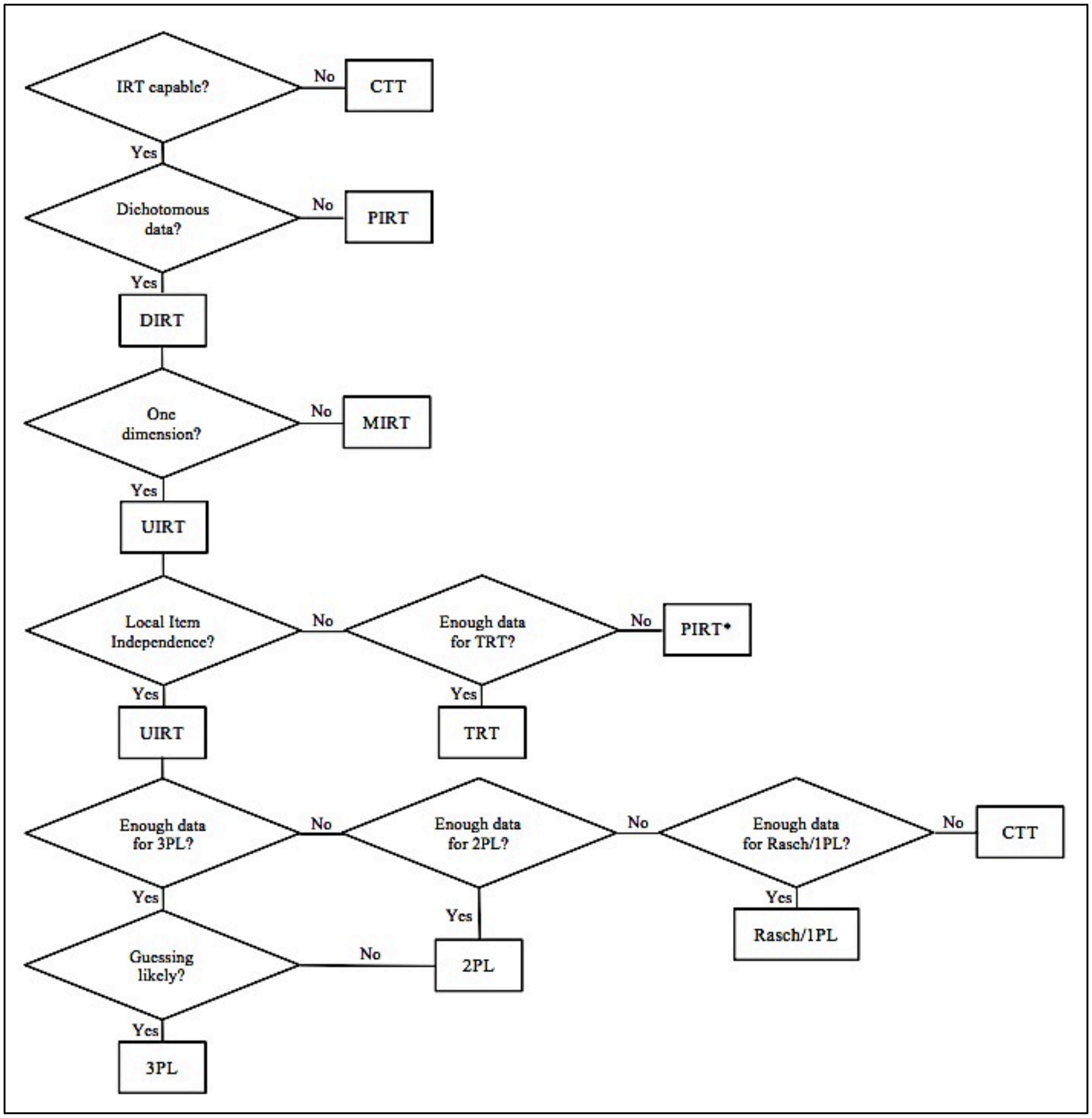


Figure 50. Proposed model-choice decision tree

As shown, the most fundamental decision is whether exam scorers have the ability to run IRT analysis. As explained earlier, the barriers to implementing IRT analysis are considerably lower these days due to more user-friendly IRT software and well-written supporting documentation. However, even if IRT analysis remains beyond the reach of exam scorers, CTT analysis at a minimum should be conducted because it is highly unlikely every item on a test will be functioning as intended. CTT item-analysis methods like Kelley's D and point-biserial correlation are easy to compute and interpret, thereby ensuring tests undergo at least a rudimentary form of item analysis given the questionable validity of the raw-score method.

In contexts where IRT analysis is a possibility, a whole range of item-analysis methods emerges. For polytomous items (e.g., items that employ partial-credit scoring or items whose answers are plotted along a continuum, like a Likert scale), polytomous IRT (PIRT) is the best choice.[21] For dichotomous data, like those analyzed in this study, dichotomous IRT (DIRT) would be more appropriate.

The next two decision points relate to the assumptions of IRT modeling, unidimensionality and local item independence. If the data are multidimensional, then multidimensional IRT (MIRT) analysis should be employed. If, however, the data are sufficiently unidimensional, as they are in this study, then unidimensionality IRT (UIRT) analysis should be employed. Furthermore, if local item independence is violated, then testlet response theory (TRT) analysis should be conducted if there are sufficient data, as was the case in this study. If there are insufficient data, then PIRT should be employed, where the items of each testlet would be scored as a super-item using partial-credit scoring rather than treating them as individual items scored dichotomously.[22]

[21] Elaboration of PIRT modeling is beyond the scope of this paper. See Yen & Fitzpatrick (2006), pp. 115-8, for an introduction to PIRT, as well as for addition references.

[22] The decision tree is annotated with PIRT* at this point to indicate that PIRT modeling would be the only means of resolving local item dependence post hoc should insufficient data exist to conduct TRT analysis. However, a better solution, at least long-term, would be to redesign the test to eliminate the likelihood of LID in the first place. The best way to achieve this is through the reduction or elimination of testlets on the exam, or at least a reduction in the number of items in each testlet.

If item responses are shown to be locally independent, then the choice of UIRT model must be made, which often is dictated by the number of data available. Although there are no absolute rules for determining the number of data required by each model, some minimum thresholds have been suggested, as cited in Yen & Fitzpatrick (2006):

1PL – 20 items and 200 examinees (Wright & Stone, 1979)

2PL – 30 items and 500 examinees (Hulin, Lissak, & Drasgow, 1982)

3PL – 60 items and 1000 examinees (Hulin et al., 1982)

As shown, there should be no fewer than 20 items and 200 examinees for which data are available if IRT analysis at any level is to be employed. Otherwise, CTT analysis should be employed. When the number of data does not limit model selection, however, the 3PL model often will be the best choice unless there is reason to believe guessing will play no (systematic) role in the item response patterns of examinees (e.g., if guessing is penalized or distracter analysis indicates no distracter is more attractive to one subset of the examinee population than another). If this is the case, the 2PL model will likely fit the data as well or better than the 3PL model. Finally, the confirmatory Rasch/1PL model could be employed rather than either of the exploratory 2PL or 3PL models if for some reason there is a philosophical reason for doing so. Otherwise, the 3PL model remains the most theoretically defensible choice.

Regardless of the item-analysis method ultimately chosen, it is important to remember that each flagged item should be evaluated to determine the reason it is functioning improperly. If the reason is correctable (e.g., the correct response was mis-keyed), then the correction should be made and scores should be recalculated. If on the other hand the reason is uncorrectable (e.g., a negative value for Kelley’s D), then all data associated with the item should be removed before final scores are calculated. That said, if something precludes item removal (e.g., institutional resistance due to arguments of valuing content validity or test-design process over internal test reliability), it is recommended that both the 2PL and 3PL models be applied to determine which model fits the data better so that item/test information can be maximized, thereby increasing the validity of examinee scores.

In addition to the creation of a decision tree, the ILTA committee should develop a metric to determine whether the application of an item-analysis method other than the raw-score method is necessary, if for whatever reason the raw-score method is preferred by stakeholders. It is suggested here that magnitude of displacement be the metric of choice, where displacement is a function of only those that would be admitted via the raw-score method (rather than the entire population sample) because it seems to be a more accurate representation of the actual magnitude of displacement. Table 44 illustrates the outcome of such a metric with respect to the full data set analyzed in this study.

Table 44. Displacement as a function of the total number admitted

		Raw Score		Displacement
		Admit	Reject	Percentage
CTT (Kelley)	Admit	383	50	
	Reject	79	1808	27.9%
CTT (PtBis)	Admit	391	51	
	Reject	71	1807	26.4%
Rasch/1PL	Admit	376	41	
	Reject	86	1817	27.5%
IRT 2PL	Admit	369	37	
	Reject	93	1821	28.1%
IRT 3PL	Admit	381	24	
	Reject	81	1834	27.0%

As shown, 26-28% of the 462 examinees admitted under the raw-score method would be displaced if an alternate scoring method were applied to these data. Put another way, more than one of every four examinees offered admission would vary if an alternate scoring method were employed. While there is no absolute criterion by which to decide whether the displacement percentage is substantial enough to justify the use of an alternate scoring method, certainly one in four would seem to meet such a threshold. In this vein, perhaps the ILTA committee could revise its guidelines to state that misfitting items must be removed from the calculation of scores should a certain percentage of examinees be affected by the application of an alternate scoring method.

If stakeholder resistance to item removal were to remain even after a substantial amount of displacement were identified via the suggested analysis, it is recommended the ILTA committee state that either the IRT 2PL or 3PL model be employed because they are exploratory by design, meaning they attempt to fit the data *without item removal*. Because faulty items are also given less weight in IRT estimations of person ability, the retention of all items, faulty or otherwise, could be justified to some degree provided either/both of these models fit the data reasonably well when applied to the full item set.

In conclusion, language-assessment specialists aiming to conduct post-hoc item analysis may face a number of practical challenges. One involves the cultural context in which the testing takes place. The belief that raw scores comprise a legitimate criterion for rank ordering even high-stakes candidates may be entrenched over many generations of testing practice. The use of post-hoc item-analysis methods to identify faulty items therefore may be viewed with suspicion among stakeholders, so careful introduction of any of the methods described above would be a wise first step before effecting any policy changes. A second challenge involves exploratory analyses of existing data sets derived from authentic high-stakes tests. As test designs and examinee populations can be expected to differ across contexts, both unidimensionality and local item independence must be examined each time before IRT modeling is applied to the data, a practical challenge with respect to resource constraints. These concerns duly noted, implementation of any of the methods outlined previously should yield benefits that far outweigh their costs, in the form of substantially more valid test results.

With respect to the university under investigation in this study, it is argued that university administrators have an obligation to ensure classification accuracy is maximized, especially given how critical university entrance is in shaping the future of Japanese citizens. However, they need guidance in order to achieve this goal, so ILTA and other governing testing-bodies have an obligation to expand their guidelines to include advice on how to choose the item-analysis method most appropriate for this context (and many others). Note too that IRT and TRT modeling should be included among the list of choices despite their relative complexity. While IRT modeling software was arcane, expensive, and not readily available even just a decade ago, this is no longer the case. Software applications like Xcalibre 4.1 and SCORIGHT 3.0 are all relatively inexpensive and user-friendly enough that users with little or no experience can learn how to run the analyses and interpret the output simply by following the guidance provided in the corresponding user manual.

Appendix A: Entrance Exam

[I]　次の英文を読み、下記の設問に答えなさい。

Ever since lead was linked to health risks, the U.S. government gradually phased out the use of lead in gasoline and household paint, but it is still present in many products. Makers of china, water faucets, and food supplements have recently gone to great lengths to reduce the amount of lead they use. What is remarkable is that these efforts are not the usual attempts to avoid stiff penalties associated with new public health rules. Instead, they are a response to a California law requiring companies to provide information to the public about products that remain perfectly legal. Corporations all over the country are feeling the effects of an increasingly powerful but unheralded government policy tool: mandatory disclosure.

In 1986 California voters approved by a margin of two to one a ballot initiative that required companies to give 'clear and reasonable warning' whenever they exposed people to cancer-causing chemicals or substances toxic to the reproductive system in amounts above levels set by the state. The 1986 law in fact prompted few such warnings. Faced with public shame if accused of failing to warn consumers, many nationally known companies reduced the public's exposure to lead and other toxic poisons. Ten china companies agreed to cut the amount of lead in their glazes by half. Fourteen major plumbing-supply manufacturers agreed to produce brass pipes that were virtually lead-free. A major food processing company removed lead from all of its canned food products. Many others adopted similar policies about potentially dangerous products.

It is not that the companies accepted the idea that their products posed risks to consumers. On the contrary, they argued that the California law in many cases unfairly emphasized risks that were negligible. They found support in Professor W.K. Viscusi of Harvard Law School, who studied the law closely and concluded that it probably did more harm than good, by giving people a false impression of the real risks. But the companies changed their products anyway. Further, because California amounts to 15 percent of the national market for many goods, they often changed them nationwide. Why did the companies make expensive – and they believed, unnecessary – changes? They were bowing to a newly potent political force: regulation by shaming.

With politicians calling for greater 'transparency' in business and government and complaining that national standards are often costly and ineffective, mandatory disclosure is being used as one way of addressing social problems ranging from persistent pollution to medical errors. Informational approaches to disclosure are less expensive policy tools than government regulations because citizens are informed directly. The Internet provides citizens access to disclosure data and a means of turning it into useful information. Citizen-consumers can then put pressure on offending companies by not buying their products.

Some familiar kinds of disclosure requirements create economic incentives for companies to improve their practices: nutritional labeling on food packaging, for instance, aims to influence which processed foods consumers buy. On-time ranking of airlines is designed to aid travelers in making informed choices. Other requirements amount to corporate shaming. Manufacturers listed among the worst polluters or accused of exploiting workers in developing countries may change their ways out of fear of customer boycotts, increased regulation, or community hostility. The company's reputation, hard to build and easy to destroy, is at stake.

Mandatory disclosure has now taken its rightful place beside the power of the government to tax and the power to define national standards as a means of carrying out public policy. But disclosure is no simple solution. It can itself be costly and ineffective. Requirements on companies should be approached with care. Disclosure requirements are just as difficult to design and enforce as any other government policy.

設問
本文の内容に合うように英文（１－１５）を完成する場合、最も適当なものを、それぞれ下記（a-d）の中から１つ選んで、その記号をマークしなさい。

1. Ever since lead was identified as a health risk, producers have
   a. not been very concerned with the lead reduction policy.
   b. remarked that lead products are now made in China.
   c. tried to reduce the amount of lead in their products.
   d. made considerable efforts to avoid heavy fines.

2. After the enactment of the California law, many corporations
   a. are now required to make powerful tools illegal.
   b. are required to tell consumers about what they are buying.
   c. have heralded the mandatory policy of disclosure.
   d. have had to pay stiff fines for public health policies.

3. The 1986 law in fact
   a. led to many warnings about dangerous chemicals.
   b. forced companies to set reproductive limitations.
   c. warned nationally known companies of toxins.
   d. did not result in many public danger warnings.

4. Faced with exposure unless they warned customers,
   a. well-known companies took action to make their products safer.
   b. many companies faced public shame by accusing the media.
   c. major Chinese companies have cut the lead in their products.
   d. less-well known companies faced the shame of lead customers.

5. One major producer of processed foods
   a. replaced lead with brass in its cans.
   b. reduced the risk of toxic poisoning.
   c. created a major health risk in China.
   d. found brass cans to be poisonous.

6. Although many companies disclosed product information,
   a. they argued with consumers about the unfair risks.
   b. their lawyers emphasized that risks to companies were posed.
   c. they claimed the actual risks were in fact very small.
   d. their lawyers argued with Professor Viscusi of Harvard.

7. Professor Viscusi examined the law carefully, and
   a. got strong support from California lawyers.
   b. showed that it did more good than harm.
   c. was impressed that the risks were really good.
   d. concluded that the law was not a good one.

8. Even though the law was considered faulty,
   a. companies changed their products to avoid possible public criticism.
   b. most companies did not follow the law because it was unnecessary.
   c. products were not changed because 85% of them were sold elsewhere.
   d. many companies decided to sell their products nationwide.

9. Policy makers want more clarity, and believe that
   a. national standards result in best practice.
   b. social problems are solved by strong national standards.
   c. standards are frequently too expensive and not very useful.
   d. mandatory disclosure creates new problems of its own.

10. By relying on media to inform consumers, politicians think that
   a. information about social problems discloses expensive tools.
   b. disclosure information is more cost-effective than national standards.
   c. information disclosure policy leads to errors.
   d. government regulations are tools proven to be effective.

11. Using information technology, consumers can now
   a. directly affect companies by not buying their products.
   b. solve problems with companies using the internet.
   c. offend companies by providing direct information.
   d. inform companies about the usefulness of their products.

12. Companies are motivated by disclosure laws so that
   a. they eventually lead to more processed foods.
   b. the truth about the performance of airlines is hidden.
   c. shameful disclosures help companies gain attention.
   d. disclosures also influence consumer choices.

13 With the risk of corporate shaming through disclosure,
   a. a company's reputation is helped by negative disclosure information.
   b. communities may be hostile if company reputations are regulated.
   c. increased regulation is difficult to build and easy to break.
   d. many companies change their practices to avoid problems.

14. Mandatory disclosure has by now become
   a. a key public policy tool in government.
   b. standard as a means of effective tax policy.
   c. a simple solution to tax and standards problems.
   d. an effective means to set national standards.

15. Mandatory disclosure might be ineffective if
   a. it leads to boycotts.
   b. not designed and implemented carefully.
   c. companies don't agree with it.
   d. approached with too much shaming.

[II]　次の英文の空所（１－２０）に入れるのに最も適当な語を、それぞれ下記（a―d）の中から１つ選んで、その記号をマークしなさい。

Being an educator, and consequently having the souls of children in one's hands, was (1) the Port-Royalists the noblest profession in the eyes of God and the Church. A fundamental part of achieving the general education envisaged by Saint-Cyran (i.e., Jean Duvergier de Hauranne, 1581-1643) was to (2) judgment. It is here that the *Grammaire Générale et Raisonée* (1660), by Charles Lancelot and Antoine Arnauld, and other Port-Royal (3) on language begin to enter the scene as a part of the Port-Royal plan to develop clear thinking, as shown in the prefatory remarks to the *Logique* (1662) by Antoine Arnauld and Pierre Nicole: 'the principal task that one should have is to develop one's judgment and to make it as exact as possible, and it is to this end that the bulk of our studies should be directed.' This emphasis on judgment also implied that one (4) follow reason rather than usage in educational methods. One should strive to find the quickest, (5) methods for teaching, because effort should be spent on working with the subject matter rather than acquiring it. This conception of education (6) in an approach to teaching languages which represented a significant break with (7) methods.

Traditionally, students in French schools had not (8) French, as this was deemed a subject unworthy of attention. (9) study had been limited to Latin and perhaps Greek, both taught via the 'direct method,' which meant studying the target language (10) the aid of the student's native language. Rather (11) being the fulfillment of any philosophical approach to teaching language, the direct method had (12) the one most commonly used simply by habit. Since it was the language of education, all (13) were in Latin, whether the subject was medicine or religion, or Hebrew or Latin itself. Emphasis was (14) memorization of grammatical rules as well as passages from classical (15), and the composition of strictly traditional themes.

In contrast, Lancelot's efforts ensured that the (16) at the Port-Royal schools was considerably more enlightened. A thorough (17) in French was considered important before beginning the study of a foreign language. This permitted students to make use of their native language in the study of Latin or Greek; written translations were (18) out from the target language to French and vice versa, thereby improving their skills in both languages. Moreover, students were encouraged to seek natural forms of expression in translations into Latin and Greek; the exact (19) of the author were not required. Style was not (20), but language was seen above all as a tool of communication.

1.	a. at	b. in	c. against	d. for
2.	a. dodge	b. stress	c. fare	d. relate
3.	a. boats	b. ideas	c. works	d. generals
4.	a. try	b. keep	c. should	d. never
5.	a. easiest	b. hardest	c. unsurest	d. feudalist
6.	a. found	b. altered	c. spawned	d. resulted
7.	a. easy	b. past	c. hard	d. salty
8.	a. called	b. studied	c. clarified	d. experienced
9.	a. University	b. Optional	c. Language	d. Mathematics
10.	a. without	b. whence	c. from	d. throughout
11.	a. often	b. obviously	c. than	d. carefully
12.	a. seen	b. been	c. taken	d. received
13.	a. prayers	b. missals	c. textbooks	d. films
14.	a. to	b. on	c. up	d. at
15.	a. authors	b. architecture	c. era	d. cities
16.	a. chairs	b. weight	c. kindness	d. situation
17.	a. grinder	b. grain	c. grounding	d. grove
18.	a. lied	b. taking	c. won	d. carried
19.	a. message	b. words	c. natures	d. communication
20.	a. unimportant	b. illustration	c. illegal	d. everlasting

[III]　次の英文を読み、文脈を判断しながら、文中の下線部（１－１５）の意味に最も近いものを、　それぞれ下記（a—d）の中から１つ選んで、その記号をマークしなさい。

Searching for a land of freedom and opportunity, thousands of former slaves left the United States in the 19th century and sailed across the Atlantic to a continent their ancestors had unwillingly left.

Over the (1) decades, freed blacks settled on the west coast of Africa in what is today Liberia. They established a nation on July 26, 1847, and also a relationship with Africans that continues to influence regional politics.

A colony for blacks outside the United States had been proposed several times, beginning in the 1700s, but it was the American Colonization Society's formation in 1817 that provided the (2) impetus to make it a reality.

"Colonization was supposed to be sort of a (3) remedy for slavery and racial inequality in the country," said Claude Clegg, author of "The Price of Liberty: African Americans and the Making of Liberia." "Colonization was believed to be a middle ground -- you rid the nation of slavery but also rid the country of African Americans and the whole issue of race altogether."

The Colonization Society attracted a (4) mixed bag of supporters, Clegg said. Anti-slavery Quakers believed blacks would only find true freedom away from the United States; many slaveholders did not want free blacks in the country; and some freed blacks who wished to live in their ancestral homeland supported the group.

Still, many other freed slaves and anti-slavery activists opposed the idea of colonization, believing that those wishing to go to Africa should stay and fight for freedom in the United States.

"Many blacks said: 'We were born here and we have every right to be here as much as any group' and criticized those who wanted to leave," explained Wynfred Russell, who was born in Liberia and teaches classes on African-American and African studies at the University of Minnesota.

Five years after its formation, the American Colonization Society launched its first ship to Liberia, (5) founding a settlement named Monrovia, after U.S. President James Monroe. Over the decades, the number of blacks sailing to Liberia steadily increased. Settlers built schools, churches and roads and formed a government modeled on the United States.

By the 1840s, many European countries had established colonies surrounding Liberia and were pressuring the colony, the American Colonization Society and even the United States to clarify Liberia's role and identity: Could it (6) broker treaties and trade agreements? Could it (7) levy taxes? And could England or France annex the area if it was not claimed by any other country?

In response, Joseph Jenkins Roberts, who had immigrated to Liberia in 1829, publicly declared the colony an independent republic on July 26, 1847, and was elected president the next year. The declaration created the first black-ruled republic in Africa.

But tensions between the settlers and the (8) indigenous people grew within the new nation. When the first settlers came ashore, they were not completely welcomed, according to Clegg. The new immigrants' takeover of land and the injection of U.S. customs and religion into the culture was (9) resented.

"It was the same issue as this country and any other settler society in which you have a native people and then others who have come to settle," said Clegg, a history professor at Indiana University. "Liberia is a mirror image of the United States and its settlement.

"You have immigrants ... who are settling along the coast and seizing the lands and the labor -- and sometimes the lives -- of African people. They weren't particularly pleased to see the settlers arrive."

This (10) discord, vulnerability to African diseases, and the hardship of creating a Westernized nation disilllusioned many immigrants.

"There was a certain amount of romantic (11) sentiment that comes through some of the letters" (from immigrants), Clegg said. "There are those who believed they had a sort of long-standing connection with Africa.

"Maybe that (12) facilitated the willingness of some people to project upon Africa their hopes and desires. Once they get there, many of them are shocked into a realization that they were very wrong about what Africa was about."

But some did find what they were looking for. William Burke and his family sailed to Liberia in 1853 after they were freed by owner Robert E. Lee, later a (13) celebrated Confederate general, and were enthusiastic about their new home.

"I bless God that ever my lot was cast in this part of the earth," Burke wrote after five years in Liberia, where he became a minister and educator. "God has blessed me abundantly since my residence in Africa, for which I feel that I can never be sufficiently thankful."

"I love Africa and would not exchange it for America," agreed Burke's wife, Rosabella, in a letter to the wife of her former owner Lee.

By 1867, the American Colonization Society had sent more than 13,000 people, according to the Library of Congress. In these immigrants' (14) quest to escape oppression in the United States, they created the same exclusive practices they left behind and sowed the seeds for future civil conflict.

"They took a completely new political system that natives didn't know about and (15) dominated the political system for 150 years," Russell said.

1.	a. obstacles	b. objections	c. seas	d. years
2.	a. financing	b. force	c. structure	d. truth
3.	a. reason	b. protection	c. cure	d. symptom
4.	a. variety	b. minimum	c. maximum	d. bunch
5.	a. discovering	b. establishing	c. invading	d. borrowing
6.	a. break	b. understand	c. arrange	d. obey
7.	a. pay	b. lower	c. avoid	d. collect
8.	a. native	b. poor	c. unfriendly	d. hard-working
9.	a. appreciated	b. disliked	c. completed	d. repeated
10.	a. conflict	b. illness	c. reason	d. agreement
11.	a. logic	b. insistence	c. history	d. feeling
12.	a. distorted	b. erased	c. strengthened	d. weakened
13.	a. famous	b. simple	c. dangerous	d. generous
14.	a. duty	b. hurry	c. search	d. failure
15.	a. destroyed	b. avoided	c. explained	d. controlled

[IV]　次の各文の（a～d）の中で誤っている箇所があれば、その記号を１つマークしなさい。もし誤っている箇所がなければ、（e）をマークしなさい。

1. The (a) term automobile is commonly (b) applies to a four-wheeled vehicle designed (c) to carry two to six passengers and a limited amount of cargo, as (d) contrasted with a truck. (e) no error
2. The (a) processes of nuclear fission (b) was discovered in 1938 (c) by Otto Hahn and Fritz Strassmann and was (d) explained in early 1939 by Lise Meitner and Otto Frisch. (e) no error
3. After (a) Toyotomi Hideyoshi death (1598), Tokugawa Ieyasu (b) became the most powerful *daimyo* by (c) defeating rival barons (d) in the battle of Sekigahara (1600). (e) no error
4. In the USA, each state (a) issues fishing licenses and sets (b) regulations as to the season (c) of which certain fish may be caught, the minimum size, and the number that may be (d) taken per day. (e) no error
5. Sally (a) or Betty were very (b) late for class because they had (c) stayed up late (d) studying the night before. (e) no error
6. When you (a) try to repair a car by (b) oneself, it is important to have all the (c) tools and parts at hand (d) before you begin. (e) no error
7. Summer is an important time (a) for families to go (b) always together on vacation and enjoy some (c) special time together without the (d) stress of everyday life. (e) no error
8. If you are (a) doing the laundry, you (b) should try to wash white things and bright colored things (c) separate, or the colors might (d) ruin the white clothes. (e) no error
9. As the plane (a) began to move (b) from the terminal, the attendants gave a brief (c) safety demonstration and the passengers fastened (d) their seat belts. (e) no error
10. Since the start of the Internet in 1993, the world has become more (a) connect than (b) ever before (c) with e-mail and web pages (d) linking people in almost all nations. (e) no error
11. (a) During the Ashikaga period, the town now named Kobe (b) was an important port in Japan for trading with (c) much countries (d) in Asia. (e) no error
12. (a) If the (b) purpose of the war was to (c) bring world peace, then it was (d) complete unsuccessful. (e) no error
13. The boy rode (a) his bike through the park and (b) onto the street (c) which there were many cars and (d) trucks. (e) no error
14. (a) Driving a car can be very (b) convenience, (c) but it can be very expensive, (d) too. (e) no error
15. When Bill (a) put the steaks (b) on the grill, Mary put the (c) potato salad (d) and drinks on the picnic table. (e) no error
16. As the clouds (a) cover the sun, the weather started (b) to turn cold and Kenji wished he (c) had brought (d) a jacket. (e) no error
17. Before (a) you start to answer the questions (b) in the test, you (c) must put your bag under the desk and (d) turn off your cell phone. (e) no error

18. There (a) seems to be a problem with the video (b) because every time the movie starts, (c) we can't hear (d) anything. (e) no error
19. The sun is the (a) center of the solar system (b) with nine planets (c) revolving around (d) it. (e) no error
20. This is the (a) last exam (b) for the day (c) after you can go home and (d) relax for the evening. (e) no error

Appendix B: Item-Analysis Tables (Full Data Set)

Appendix B.1. CTT (Kelley's Discrimination Index)

Item	UG	LG	Total	*p*	D	Flag
1	486	383	1592	69%	20%	F
2	391	142	958	41%	41%	
3	352	141	903	39%	35%	
4	355	160	930	40%	33%	
5	320	205	938	40%	20%	F
6	398	152	979	42%	41%	
7	495	277	1478	64%	37%	
8	432	148	1013	44%	47%	
9	443	201	1198	52%	41%	
10	346	182	920	40%	28%	
11	513	248	1457	63%	45%	
12	484	273	1405	61%	36%	
13	347	151	829	36%	33%	
14	273	135	743	32%	23%	
15	444	248	1271	55%	34%	
16	309	206	923	40%	18%	F
17	164	57	352	15%	18%	F
18	319	219	1001	43%	18%	F
19	562	386	1755	76%	31%	
20	487	271	1444	62%	37%	
21	523	405	1777	77%	22%	
22	311	172	944	41%	24%	
23	590	469	1977	85%	23%	F
24	545	433	1812	78%	21%	
25	450	288	1352	58%	28%	
26	518	395	1727	74%	23%	
27	347	133	845	36%	36%	
28	586	432	1915	83%	28%	F
29	484	275	1371	59%	36%	
30	279	131	726	31%	25%	
31	449	296	1379	59%	27%	
32	345	217	1010	44%	22%	
33	497	434	1728	74%	13%	F
34	345	250	1062	46%	17%	F
35	307	237	1034	45%	13%	F
36	605	498	2112	91%	21%	F
37	306	258	1056	46%	10%	F
38	330	171	904	39%	27%	
39	562	468	1928	83%	19%	F
40	521	350	1661	72%	30%	
41	209	151	634	27%	11%	F
42	175	97	489	21%	13%	F
43	509	292	1512	65%	37%	
44	509	243	1417	61%	45%	
45	414	194	1081	47%	37%	
46	556	421	1812	78%	25%	
47	364	228	1107	48%	24%	
48	435	277	1288	56%	28%	

49	456	320	1423	61%	24%	
50	586	420	1916	83%	30%	F
51	425	233	1251	54%	33%	
52	453	268	1345	58%	32%	
53	215	41	397	17%	29%	F
54	444	275	1357	58%	30%	
55	544	383	1728	74%	29%	
56	522	295	1536	66%	39%	
57	379	245	1203	52%	24%	
58	268	170	803	35%	17%	F
59	450	146	1096	47%	50%	
60	453	151	1072	46%	50%	
61	576	419	1855	80%	29%	
62	472	241	1264	54%	39%	
63	397	223	1133	49%	30%	
64	318	107	755	33%	35%	
65	425	268	1284	55%	28%	
66	349	202	1042	45%	25%	
67	540	363	1701	73%	31%	
68	406	204	1125	48%	34%	
69	211	103	573	25%	18%	F
70	387	138	936	40%	41%	

Appendix B.2. CTT (Point-Biserial Correlations)

Item	b	rpb	Flag
1	-0.84	0.18	F
2	0.37	0.34	
3	0.48	0.30	
4	0.43	0.26	
5	0.41	0.17	F
6	0.33	0.33	
7	-0.61	0.32	
8	0.27	0.38	
9	-0.08	0.32	
10	0.45	0.21	
11	-0.57	0.38	
12	-0.47	0.30	
13	0.62	0.28	
14	0.8	0.21	
15	-0.21	0.26	
16	0.44	0.17	F
17	1.83	0.21	
18	0.29	0.15	F
19	-1.22	0.31	
20	-0.54	0.31	
21	-1.27	0.23	
22	0.4	0.19	F
23	-1.87	0.28	
24	-1.36	0.24	

25	-0.36	0.25	
26	-1.15	0.21	
27	0.6	0.30	
28	-1.67	0.31	
29	-0.4	0.30	
30	0.84	0.23	
31	-0.42	0.23	
32	0.28	0.18	F
33	-1.15	0.13	F
34	0.18	0.14	F
35	0.23	0.12	F
36	-2.45	0.31	
37	0.19	0.11	F
38	0.48	0.22	
39	-1.71	0.23	
40	-0.99	0.29	
41	1.04	0.11	F
42	1.41	0.15	F
43	-0.68	0.32	
44	-0.49	0.36	
45	0.14	0.30	
46	-1.37	0.25	
47	0.09	0.21	
48	-0.24	0.23	
49	-0.5	0.20	
50	-1.68	0.34	
51	-0.18	0.26	
52	-0.35	0.26	
53	1.68	0.29	
54	-0.38	0.24	
55	-1.16	0.28	
56	-0.73	0.32	
57	-0.09	0.18	F
58	0.68	0.15	F
59	0.11	0.39	
60	0.15	0.40	
61	-1.49	0.29	
62	-0.2	0.30	
63	0.04	0.25	
64	0.78	0.29	
65	-0.24	0.21	
66	0.21	0.20	
67	-1.11	0.29	
68	0.05	0.27	
69	1.18	0.17	F
70	0.41	0.33	

Appendix B.3. Rasch/1PL

		Infit		Exact Match		
Item	b	IN.MSQ	IN.ZSTD	Obs%	Exp%	Flag
1	-0.84	1.04	2.18	68.2	69.6	F
2	0.37	0.95	-3.75	65.8	63.0	
3	0.48	0.98	-1.73	65.0	64.3	
4	0.43	1.00	0.04	64.0	63.6	
5	0.41	1.05	3.82	61.6	63.4	F
6	0.33	0.96	-3.24	65.6	62.6	
7	-0.61	0.97	-2.20	67.2	65.9	
8	0.27	0.93	-5.94	66.9	62.0	
9	-0.08	0.97	-3.15	63.8	60.9	
10	0.45	1.03	2.05	62.8	63.8	F
11	-0.57	0.93	-4.85	68.4	65.3	
12	-0.47	0.98	-1.37	64.4	64.0	
13	0.62	0.98	-1.21	69.6	66.2	
14	0.80	1.02	1.28	69.1	69.0	
15	-0.21	1.00	0.02	62.2	61.5	
16	0.44	1.06	4.16	60.8	63.8	F
17	1.83	0.99	-0.26	84.8	84.8	
18	0.29	1.07	5.46	57.1	62.2	F
19	-1.22	0.96	-1.51	76.6	75.9	
20	-0.54	0.97	-2.05	66.8	65.0	
21	-1.27	0.99	-0.32	77.3	76.8	
22	0.40	1.04	3.16	59.7	63.3	F
23	-1.87	0.96	-1.10	85.4	85.3	
24	-1.36	0.99	-0.20	78.6	78.3	
25	-0.36	1.01	0.78	61.4	62.8	
26	-1.15	1.01	0.54	74.9	74.8	
27	0.60	0.98	-1.65	67.9	65.8	
28	-1.67	0.95	-1.53	82.8	82.7	
29	-0.40	0.98	-1.56	63.4	63.1	
30	0.84	1.01	0.63	69.1	69.6	
31	-0.42	1.02	1.48	62.3	63.4	
32	0.28	1.05	3.99	59.3	62.0	F
33	-1.15	1.05	2.16	75.0	74.9	F
34	0.18	1.07	6.20	57.5	61.3	F
35	0.23	1.09	7.38	55.5	61.7	F
36	-2.45	0.93	-1.18	91.1	91.0	
37	0.19	1.09	7.97	54.9	61.4	F
38	0.48	1.02	1.72	63.4	64.2	
39	-1.71	0.98	-0.45	83.4	83.3	
40	-0.99	0.97	-1.25	73.4	72.2	

41	1.04	1.07	3.04	72.3	73.0	F
42	1.41	1.03	1.08	79.2	79.0	
43	-0.68	0.97	-2.16	68.3	67.0	
44	-0.49	0.94	-4.15	67.3	64.3	
45	0.14	0.98	-1.82	63.8	61.1	
46	-1.37	0.98	-0.52	78.6	78.4	
47	0.09	1.04	3.22	59.5	60.9	F
48	-0.24	1.02	1.79	59.8	61.7	
49	-0.50	1.04	2.61	62.5	64.4	F
50	-1.68	0.93	-1.89	83.1	82.8	
51	-0.18	1.00	0.14	61.9	61.3	
52	-0.35	1.00	0.15	62.8	62.7	
53	1.68	0.96	-1.16	82.9	82.9	
54	-0.38	1.01	1.03	62.5	62.9	
55	-1.16	0.98	-0.93	75.1	74.9	
56	-0.73	0.96	-2.25	69.0	67.8	
57	-0.09	1.05	4.36	57.8	60.9	F
58	0.68	1.06	3.58	64.7	67.0	F
59	0.11	0.93	-6.77	66.1	61.0	
60	0.15	0.92	-7.13	68.0	61.2	
61	-1.49	0.97	-0.96	80.2	80.2	
62	-0.20	0.98	-1.77	62.9	61.4	
63	0.04	1.01	0.86	60.7	60.8	
64	0.78	0.98	-1.41	69.5	68.6	
65	-0.24	1.03	2.64	59.8	61.7	F
66	0.21	1.04	3.32	58.9	61.5	F
67	-1.11	0.97	-1.21	74.8	74.2	
68	0.05	1.00	-0.05	60.0	60.9	
69	1.18	1.03	1.32	74.9	75.4	
70	0.41	0.96	-3.07	66.2	63.3	

Appendix B.4. IRT 2PL

Item	a	b	Flag
1	0.242	-1.947	F
2	0.472	0.475	
3	0.390	0.725	
4	0.313	0.778	
5	0.208	1.091	F
6	0.439	0.449	
7	0.437	-0.845	
8	0.534	0.305	
9	0.423	-0.121	

10	0.254	0.979	F
11	0.564	-0.651	
12	0.394	-0.714	
13	0.371	0.978	
14	0.269	1.675	F
15	0.328	-0.384	
16	0.192	1.264	F
17	0.373	2.852	
18	0.176	0.899	F
19	0.482	-1.550	
20	0.413	-0.790	
21	0.338	-2.173	
22	0.223	0.991	F
23	0.531	-2.185	
24	0.361	-2.203	
25	0.304	-0.691	
26	0.303	-2.170	
27	0.400	0.873	
28	0.563	-1.867	
29	0.391	-0.608	
30	0.293	1.625	F
31	0.272	-0.879	F
32	0.203	0.742	F
33	0.209	-3.034	F
34	0.165	0.584	F
35	0.146	0.846	F
36	0.809	-2.096	
37	0.139	0.709	F
38	0.258	1.037	F
39	0.413	-2.455	
40	0.423	-1.408	
41	0.187	3.037	F
42	0.264	2.983	F
43	0.451	-0.918	
44	0.500	-0.613	
45	0.375	0.212	
46	0.395	-2.047	
47	0.238	0.207	F
48	0.270	-0.515	F
49	0.238	-1.186	F
50	0.637	-1.715	
51	0.372	-0.584	
52	0.570	-1.655	
53	0.627	1.270	
54	0.422	-1.211	
55	0.448	-1.780	
56	0.455	-0.978	
57	0.201	-0.240	F
58	0.195	1.893	F

59	0.561	0.117	
60	0.590	0.159	
61	0.485	-1.881	
62	0.391	-0.312	
63	0.305	0.069	
64	0.396	1.150	
65	0.245	-0.563	F
66	0.227	0.501	F
67	0.426	-1.562	
68	0.336	0.082	
69	0.259	2.555	F
70	0.444	0.548	

Appendix B.5 IRT 3PL

Item ID	a	b	c	Flag
1	0.278	-0.895	0.214	F
2	0.704	0.888	0.173	
3	0.583	1.189	0.174	
4	0.580	1.333	0.210	
5	0.405	1.893	0.227	
6	0.748	0.913	0.197	
7	0.539	-0.242	0.202	
8	0.872	0.715	0.188	
9	0.584	0.470	0.198	
10	0.553	1.575	0.232	
11	0.687	-0.196	0.190	
12	0.507	-0.030	0.208	
13	1.052	1.265	0.227	
14	0.535	2.023	0.192	
15	0.457	0.400	0.210	
16	0.355	2.162	0.223	
17	0.938	2.345	0.118	
18	0.278	2.030	0.209	F
19	0.534	-1.045	0.204	
20	0.488	-0.189	0.196	
21	0.366	-1.512	0.204	
22	0.345	1.858	0.199	
23	0.564	-1.788	0.203	
24	0.393	-1.565	0.207	
25	0.383	0.152	0.206	
26	0.339	-1.362	0.211	
27	0.744	1.257	0.191	
28	0.605	-1.472	0.203	
29	0.515	0.105	0.215	
30	0.606	1.910	0.191	
31	0.336	0.054	0.206	
32	0.351	1.707	0.220	
33	0.225	-1.920	0.218	F
34	0.262	1.858	0.216	F
35	0.236	2.177	0.212	F

36	0.835	-1.881	0.200	
37	0.221	2.157	0.213	F
38	0.455	1.686	0.206	
39	0.439	-1.938	0.206	
40	0.472	-0.846	0.200	
41	0.710	2.762	0.230	K
42	0.791	2.575	0.171	
43	0.537	-0.354	0.198	
44	0.627	-0.086	0.197	
45	0.545	0.810	0.194	
46	0.434	-1.454	0.205	
47	0.343	1.168	0.206	
48	0.358	0.425	0.209	
49	0.295	-0.101	0.211	F
50	0.676	-1.379	0.199	
51	0.457	0.090	0.198	
52	0.630	-1.226	0.204	
53	1.097	1.346	0.123	
54	0.496	-0.600	0.204	
55	0.504	-1.210	0.211	
56	0.551	-0.386	0.206	
57	0.267	0.955	0.205	F
58	0.407	2.521	0.220	K
59	0.715	0.511	0.164	
60	0.860	0.553	0.178	
61	0.539	-1.362	0.211	
62	0.546	0.376	0.212	
63	0.422	0.832	0.198	
64	0.665	1.465	0.165	
65	0.314	0.439	0.203	
66	0.328	1.458	0.201	
67	0.476	-0.988	0.204	
68	0.475	0.786	0.201	
69	0.637	2.478	0.180	
70	0.667	0.985	0.176	

Appendix B.6. TRT

Item	a	b	c	Flag
1	0.343	-1.209	0.206	
2	0.669	0.899	0.059	
3	0.524	1.251	0.049	
4	0.550	1.657	0.127	
5	0.826	2.976	0.304	
6	0.954	1.167	0.163	
7	0.665	-0.824	0.066	
8	1.007	0.746	0.109	
9	0.642	0.107	0.060	
10	0.756	2.294	0.240	

11	0.899	-0.620	0.060	
12	0.723	-0.309	0.133	
13	1.723	1.776	0.238	
14	0.475	2.572	0.090	
15	0.445	-0.180	0.060	
16	0.491	2.883	0.202	
17	1.164	2.895	0.085	
18	0.300	1.896	0.099	
19	0.759	-1.627	0.051	
20	0.617	-0.769	0.046	
21	0.500	-2.447	0.036	
22	0.323	2.011	0.085	
23	0.822	-2.423	0.043	
24	0.547	-2.452	0.037	
25	0.508	-0.406	0.081	
26	0.462	-2.327	0.045	
27	0.997	1.396	0.147	
28	0.916	-1.955	0.057	
29	1.203	0.400	0.291	
30	0.665	2.258	0.125	
31	0.381	-0.837	0.047	
32	0.711	2.282	0.248	
33	0.468	-1.299	0.291	
34	0.248	1.543	0.081	F
35	0.208	1.735	0.059	F
36	1.201	-2.558	0.047	
37	0.303	3.167	0.166	F
38	0.503	1.772	0.111	
39	0.564	-3.032	0.037	F
40	0.666	-1.488	0.044	
41	1.579	3.323	0.246	F
42	0.956	3.459	0.134	F
43	0.646	-0.980	0.044	
44	0.889	-0.334	0.117	
45	0.639	0.549	0.072	
46	0.651	-2.070	0.065	
47	0.371	0.991	0.104	
48	0.386	-0.223	0.071	
49	0.308	-1.277	0.053	
50	0.961	-1.965	0.050	
51	0.593	-0.038	0.066	

52	0.578	-0.369	0.068	
53	1.091	2.034	0.038	
54	0.499	-0.536	0.052	
55	0.732	-1.481	0.074	
56	0.819	-0.784	0.070	
57	0.319	0.091	0.051	
58	0.412	3.358	0.139	F
59	1.055	0.247	0.046	
60	1.421	0.457	0.123	
61	0.947	-1.452	0.162	
62	0.981	0.293	0.181	
63	0.557	0.384	0.068	
64	0.909	1.347	0.086	
65	0.388	-0.347	0.046	
66	0.409	0.904	0.061	
67	0.768	-1.367	0.058	
68	0.610	0.440	0.080	
69	0.797	2.826	0.136	
70	0.908	0.753	0.069	

Appendix C: Item-Analysis Tables (Random 80%)

Appendix C.1. CTT (Kelley's Discrimination Index)

Item	UG	LG	Total	*p*	D	Flag
1	380	300	1273	69%	19%	F
2	306	119	767	41%	39%	
3	276	114	730	39%	34%	
4	272	130	750	40%	30%	
5	248	160	748	40%	20%	
6	317	124	797	43%	41%	
7	385	220	1181	64%	36%	
8	343	121	817	44%	47%	
9	339	164	968	52%	37%	
10	261	149	750	40%	24%	
11	394	199	1173	63%	42%	
12	371	212	1124	61%	35%	
13	280	124	676	36%	33%	
14	201	104	589	32%	21%	
15	345	200	1016	55%	32%	
16	234	163	741	40%	16%	F
17	122	45	281	15%	16%	F
18	249	180	808	44%	16%	F
19	436	305	1403	76%	30%	
20	374	218	1148	62%	34%	
21	393	330	1422	77%	16%	F
22	237	137	760	41%	22%	
23	450	376	1580	85%	19%	F
24	413	346	1445	78%	17%	F
25	351	224	1071	58%	28%	
26	393	318	1381	74%	18%	F
27	266	111	673	36%	33%	
28	451	348	1536	83%	24%	F
29	369	222	1089	59%	32%	
30	212	105	584	31%	23%	
31	348	233	1101	59%	26%	
32	269	175	815	44%	21%	
33	380	342	1391	75%	11%	F
34	266	206	852	46%	14%	F
35	238	187	832	45%	12%	F
36	462	392	1682	91%	18%	F
37	233	208	844	45%	7%	F
38	253	148	726	39%	23%	
39	432	371	1544	83%	16%	F
40	392	276	1326	71%	26%	
41	168	122	519	28%	11%	F
42	140	74	392	21%	14%	F
43	386	239	1216	66%	32%	
44	392	194	1137	61%	42%	
45	326	153	861	46%	37%	
46	422	338	1440	78%	20%	
47	275	179	886	48%	21%	
48	330	212	1022	55%	26%	

49	342	259	1123	61%	19%	F
50	447	339	1530	82%	25%	F
51	327	187	996	54%	30%	
52	341	211	1067	57%	29%	
53	162	35	312	17%	26%	
54	341	225	1092	59%	26%	
55	415	312	1381	74%	24%	
56	402	240	1244	67%	35%	
57	291	205	979	53%	20%	
58	206	139	643	35%	15%	F
59	349	118	880	47%	48%	
60	343	119	859	46%	47%	
61	439	336	1483	80%	24%	
62	368	192	1014	55%	38%	
63	297	182	895	48%	25%	
64	247	83	612	33%	34%	
65	329	213	1031	56%	26%	
66	272	167	840	45%	23%	
67	410	298	1371	74%	26%	
68	312	149	889	48%	35%	
69	167	78	461	25%	19%	F
70	298	109	736	40%	40%	

Appendix C.2. CTT (Point-Biserial Correlations)

Item	b	rpb	Flag
1	-0.84	0.20	
2	0.37	0.34	
3	0.46	0.30	
4	0.41	0.26	
5	0.42	0.18	F
6	0.30	0.33	
7	-0.60	0.31	
8	0.25	0.38	
9	-0.10	0.32	
10	0.41	0.20	
11	-0.59	0.38	
12	-0.47	0.30	
13	0.59	0.28	
14	0.81	0.20	
15	-0.21	0.27	
16	0.44	0.16	F
17	1.83	0.21	
18	0.27	0.14	F
19	-1.21	0.31	
20	-0.52	0.30	
21	-1.27	0.20	
22	0.39	0.20	
23	-1.86	0.27	
24	-1.35	0.23	

25	-0.34	0.26	
26	-1.15	0.20	
27	0.60	0.29	
28	-1.68	0.31	
29	-0.38	0.29	
30	0.83	0.22	
31	-0.41	0.24	
32	0.26	0.19	F
33	-1.17	0.14	F
34	0.18	0.14	F
35	0.22	0.12	F
36	-2.40	0.32	
37	0.19	0.10	F
38	0.47	0.20	
39	-1.72	0.24	
40	-0.98	0.29	
41	1.01	0.12	F
42	1.40	0.17	F
43	-0.70	0.30	
44	-0.50	0.36	
45	0.15	0.31	
46	-1.33	0.24	
47	0.09	0.20	
48	-0.22	0.24	
49	-0.47	0.19	F
50	-1.66	0.32	
51	-0.17	0.26	
52	-0.33	0.26	
53	1.70	0.28	
54	-0.40	0.22	
55	-1.15	0.27	
56	-0.77	0.32	
57	-0.12	0.17	F
58	0.67	0.15	F
59	0.10	0.40	
60	0.15	0.39	
61	-1.49	0.28	
62	-0.21	0.31	
63	0.07	0.24	
64	0.75	0.29	
65	-0.25	0.21	
66	0.19	0.19	F
67	-1.13	0.27	
68	0.08	0.29	
69	1.18	0.18	F
70	0.44	0.33	

Appendix C.3. Rasch/1PL

		Infit		Exact Match		
Item	b	IN.MSQ	IN.ZSTD	Obs%	Exp%	Flag
1	-0.84	1.03	1.33	68.4	69.5	
2	0.37	0.95	-3.35	65.9	62.9	
3	0.46	0.98	-1.56	64.7	64.0	
4	0.41	1.00	-0.09	63.5	63.4	
5	0.42	1.05	3.19	61.6	63.5	F
6	0.30	0.96	-3.24	66.2	62.2	
7	-0.60	0.97	-1.90	67.2	65.7	
8	0.25	0.93	-5.57	67.2	61.8	
9	-0.10	0.97	-2.76	64.1	60.8	
10	0.41	1.03	2.01	62.2	63.4	F
11	-0.59	0.93	-4.46	69.1	65.5	
12	-0.47	0.98	-1.57	64.7	63.9	
13	0.59	0.98	-1.20	69.3	65.7	
14	0.81	1.03	1.32	69.1	69.2	
15	-0.21	1.00	-0.08	61.4	61.4	
16	0.44	1.06	3.83	59.5	63.7	F
17	1.83	0.99	-0.17	84.7	84.9	
18	0.27	1.07	5.35	56.9	62.0	F
19	-1.21	0.96	-1.53	76.7	75.8	
20	-0.52	0.97	-1.67	65.8	64.6	
21	-1.27	1.01	0.21	77.1	76.8	
22	0.39	1.04	2.61	59.4	63.1	
23	-1.86	0.96	-0.83	85.3	85.2	
24	-1.35	1.00	-0.08	78.4	78.0	
25	-0.34	1.00	0.03	61.5	62.4	
26	-1.15	1.01	0.55	75.0	74.8	
27	0.60	0.98	-1.04	67.6	65.9	
28	-1.68	0.95	-1.29	83.0	82.9	
29	-0.38	0.98	-1.20	62.7	62.8	
30	0.83	1.02	0.73	68.9	69.5	
31	-0.41	1.01	0.78	62.5	63.2	
32	0.26	1.04	3.19	59.6	61.9	
33	-1.17	1.05	1.71	75.3	75.2	
34	0.18	1.07	5.73	57.4	61.2	
35	0.22	1.08	6.35	55.2	61.5	
36	-2.40	0.93	-1.11	90.6	90.6	
37	0.19	1.09	7.21	54.6	61.3	
38	0.47	1.03	2.18	63.1	64.1	
39	-1.72	0.98	-0.47	83.4	83.4	
40	-0.98	0.97	-1.23	73.2	72.0	

41	1.01	1.06	2.62	71.7	72.4	
42	1.40	1.02	0.75	79.1	78.9	
43	-0.70	0.97	-1.54	68.5	67.1	
44	-0.50	0.94	-3.68	67.2	64.3	
45	0.15	0.97	-2.23	64.4	61.1	
46	-1.33	0.99	-0.32	78.2	77.8	
47	0.09	1.04	2.89	58.8	60.8	
48	-0.22	1.01	0.90	60.2	61.4	
49	-0.47	1.04	2.76	61.9	63.8	
50	-1.66	0.94	-1.53	82.9	82.7	
51	-0.17	1.00	0.08	61.7	61.1	
52	-0.33	1.00	0.14	62.5	62.3	
53	1.70	0.96	-0.95	83.2	83.2	
54	-0.40	1.02	1.33	62.0	63.0	
55	-1.15	0.99	-0.51	74.9	74.8	
56	-0.77	0.96	-1.88	70.0	68.3	
57	-0.12	1.05	4.34	57.6	60.9	
58	0.67	1.06	3.26	65.0	66.9	
59	0.10	0.92	-6.40	66.3	60.9	
60	0.15	0.92	-6.26	67.8	61.1	
61	-1.49	0.97	-0.80	80.0	80.2	
62	-0.21	0.97	-2.06	63.2	61.3	
63	0.07	1.02	1.43	60.1	60.8	
64	0.75	0.98	-1.12	68.7	68.2	
65	-0.25	1.03	2.24	59.8	61.7	
66	0.19	1.04	3.41	58.9	61.4	
67	-1.13	0.98	-0.75	74.8	74.6	
68	0.08	0.99	-0.97	60.3	60.8	
69	1.18	1.03	1.00	74.8	75.2	
70	0.44	0.96	-2.89	66.7	63.7	

Appendix C.4. IRT 2PL

Item	a	b	Flag
1	0.273	-1.736	F
2	0.470	0.472	
3	0.382	0.705	
4	0.317	0.741	
5	0.217	1.056	F
6	0.450	0.391	
7	0.431	-0.849	
8	0.547	0.277	
9	0.420	-0.157	
10	0.253	0.899	F
11	0.576	-0.662	

12	0.407	-0.697	
13	0.375	0.911	
14	0.265	1.717	F
15	0.331	-0.380	
16	0.195	1.219	F
17	0.381	2.796	
18	0.172	0.854	F
19	0.497	-1.504	
20	0.402	-0.780	
21	0.302	-2.397	
22	0.234	0.911	F
23	0.505	-2.261	
24	0.354	-2.214	
25	0.328	-0.603	
26	0.299	-2.185	F
27	0.381	0.916	
28	0.557	-1.897	
29	0.377	-0.601	
30	0.287	1.631	F
31	0.289	-0.817	F
32	0.217	0.657	F
33	0.222	-2.927	F
34	0.165	0.561	F
35	0.155	0.754	F
36	0.816	-2.039	
37	0.143	0.686	F
38	0.238	1.091	F
39	0.426	-2.396	
40	0.432	-1.372	
41	0.198	2.767	F
42	0.287	2.746	F
43	0.433	-0.972	
44	0.496	-0.630	
45	0.394	0.213	
46	0.376	-2.080	
47	0.239	0.199	F
48	0.291	-0.443	F
49	0.225	-1.151	F
50	0.601	-1.775	
51	0.374	-0.570	
52	0.551	-1.678	
53	0.608	1.315	
54	0.405	-1.265	
55	0.420	-1.872	
56	0.443	-1.052	
57	0.193	-0.360	F
58	0.198	1.847	F
59	0.567	0.105	
60	0.582	0.153	
61	0.478	-1.897	
62	0.401	-0.316	
63	0.285	0.124	F

64	0.389	1.131	
65	0.250	-0.571	F
66	0.220	0.476	F
67	0.401	-1.675	
68	0.365	0.117	
69	0.275	2.395	F
70	0.444	0.587	

Appendix C.5 IRT 3PL

Item	a	b	c	Flag
1	0.32	-0.75	0.22	
2	0.73	0.91	0.19	
3	0.59	1.21	0.19	
4	0.58	1.32	0.21	
5	0.41	1.88	0.23	
6	0.78	0.88	0.20	
7	0.54	-0.20	0.21	
8	0.91	0.70	0.19	
9	0.59	0.47	0.21	
10	0.54	1.57	0.23	
11	0.70	-0.20	0.20	
12	0.53	-0.01	0.21	
13	1.05	1.24	0.23	
14	0.54	2.10	0.20	
15	0.46	0.42	0.22	
16	0.36	2.16	0.23	
17	0.93	2.40	0.12	
18	0.29	2.05	0.22	F
19	0.56	-0.98	0.21	
20	0.49	-0.12	0.21	
21	0.33	-1.60	0.22	
22	0.36	1.81	0.21	
23	0.55	-1.80	0.21	
24	0.39	-1.53	0.22	
25	0.43	0.23	0.22	
26	0.34	-1.33	0.22	
27	0.74	1.32	0.20	
28	0.61	-1.46	0.21	
29	0.51	0.16	0.22	
30	0.62	1.94	0.20	
31	0.36	0.09	0.21	
32	0.38	1.62	0.23	
33	0.24	-1.82	0.22	F
34	0.27	1.88	0.23	F
35	0.26	2.07	0.22	F
36	0.85	-1.80	0.21	
37	0.24	2.14	0.22	F
38	0.45	1.80	0.22	
39	0.46	-1.84	0.21	
40	0.49	-0.79	0.21	

41	0.75	2.62	0.24	K
42	0.80	2.50	0.17	
43	0.52	-0.35	0.21	
44	0.62	-0.08	0.20	
45	0.59	0.81	0.20	
46	0.42	-1.41	0.21	
47	0.35	1.19	0.21	
48	0.39	0.45	0.21	
49	0.29	0.04	0.22	F
50	0.65	-1.39	0.21	
51	0.47	0.14	0.21	
52	0.62	-1.20	0.21	
53	1.16	1.39	0.13	
54	0.48	-0.59	0.21	
55	0.48	-1.23	0.22	
56	0.54	-0.42	0.21	
57	0.26	0.95	0.22	F
58	0.43	2.51	0.23	K
59	0.75	0.53	0.18	
60	0.86	0.57	0.18	
61	0.54	-1.34	0.22	
62	0.59	0.40	0.22	
63	0.41	0.97	0.21	
64	0.64	1.51	0.17	
65	0.33	0.46	0.21	
66	0.33	1.50	0.21	
67	0.46	-1.02	0.21	
68	0.53	0.79	0.21	
69	0.67	2.41	0.18	
70	0.71	1.04	0.19	

Appendix D: Item-Analysis Tables (Middle 80%)

Appendix D.1. CTT (Kelley's Discrimination Index)

Item	UG	LG	Total	*p*	D	Flag
1	352	321	1271	68%	12%	F
2	277	133	750	40%	33%	
3	250	136	720	39%	27%	
4	242	134	715	39%	25%	
5	226	175	737	40%	14%	F
6	263	126	746	40%	31%	
7	358	253	1190	64%	27%	
8	281	130	773	42%	34%	
9	304	191	953	51%	27%	
10	236	150	708	38%	21%	
11	381	232	1186	64%	36%	
12	343	245	1118	60%	25%	
13	201	134	610	33%	17%	F
14	175	126	577	31%	13%	F
15	316	211	1016	55%	26%	
16	232	180	742	40%	14%	F
17	93	46	250	13%	11%	F
18	231	197	813	44%	11%	F
19	421	350	1425	77%	21%	
20	356	251	1179	64%	27%	
21	406	366	1467	79%	15%	F
22	229	163	766	41%	17%	F
23	447	406	1611	87%	16%	F
24	405	368	1465	79%	15%	F
25	321	250	1087	59%	20%	
26	388	339	1399	75%	17%	F
27	215	124	644	35%	22%	
28	443	383	1564	84%	20%	F
29	338	246	1090	59%	24%	
30	199	107	570	31%	21%	
31	323	275	1113	60%	15%	F
32	235	172	784	42%	16%	F
33	366	366	1394	75%	7%	F
34	256	221	848	46%	11%	F
35	234	212	850	46%	9%	F
36	459	443	1730	93%	11%	F
37	228	222	859	46%	5%	F
38	229	144	703	38%	21%	
39	427	407	1568	84%	12%	F
40	378	324	1350	73%	17%	F
41	127	121	484	26%	3%	F
42	107	84	372	20%	6%	F
43	375	268	1229	66%	28%	
44	375	221	1145	62%	37%	
45	301	170	858	46%	31%	
46	413	370	1469	79%	16%	F
47	261	189	877	47%	19%	F
48	316	245	1035	56%	20%	

49	327	280	1139	61%	15%	F
50	437	379	1571	85%	19%	F
51	306	220	1015	55%	22%	
52	321	237	1077	58%	22%	
53	134	41	282	15%	20%	
54	331	248	1108	60%	22%	
55	399	346	1394	75%	18%	F
56	384	272	1242	67%	29%	
57	271	223	977	53%	14%	F
58	193	158	644	35%	10%	F
59	323	164	896	48%	37%	
60	312	151	843	45%	37%	
61	422	360	1496	81%	20%	F
62	323	212	993	54%	27%	
63	289	200	918	49%	23%	
64	207	107	586	32%	23%	
65	312	244	1048	56%	19%	F
66	258	182	838	45%	19%	F
67	397	329	1383	75%	20%	
68	285	189	892	48%	24%	
69	137	79	439	24%	14%	F
70	256	136	731	39%	28%	

Appendix D.2. CTT (Point-Biserial Correlations)

Item	b	rpb	Flag
1	-0.80	0.10	F
2	0.40	0.27	
3	0.47	0.22	
4	0.48	0.19	F
5	0.43	0.10	F
6	0.40	0.23	
7	-0.61	0.21	
8	0.34	0.28	
9	-0.07	0.21	
10	0.49	0.16	F
11	-0.60	0.29	
12	-0.44	0.20	
13	0.73	0.14	F
14	0.82	0.11	F
15	-0.20	0.21	
16	0.42	0.11	F
17	1.91	0.13	F
18	0.25	0.09	F
19	-1.24	0.19	F
20	-0.58	0.23	
21	-1.37	0.16	F
22	0.36	0.13	F
23	-1.94	0.18	F
24	-1.36	0.16	F

25	-0.36	0.14	F
26	-1.16	0.15	F
27	0.65	0.18	F
28	-1.74	0.21	
29	-0.37	0.19	F
30	0.84	0.19	F
31	-0.43	0.12	F
32	0.32	0.12	F
33	-1.14	0.06	F
34	0.18	0.08	F
35	0.17	0.07	F
36	-2.68	0.17	F
37	0.15	0.06	F
38	0.51	0.17	F
39	-1.75	0.13	F
40	-1.01	0.16	F
41	1.07	0.04	F
42	1.42	0.06	F
43	-0.70	0.22	
44	-0.50	0.27	
45	0.15	0.24	
46	-1.38	0.14	F
47	0.11	0.15	F
48	-0.24	0.16	F
49	-0.48	0.13	F
50	-1.77	0.20	
51	-0.21	0.18	F
52	-0.34	0.18	F
53	1.77	0.22	
54	-0.41	0.17	F
55	-1.14	0.17	F
56	-0.73	0.24	
57	-0.11	0.12	F
58	0.65	0.09	F
59	0.06	0.30	
60	0.18	0.29	
61	-1.48	0.22	
62	-0.15	0.22	
63	0.01	0.19	F
64	0.79	0.21	
65	-0.28	0.16	F
66	0.19	0.15	F
67	-1.13	0.20	
68	0.07	0.17	F
69	1.20	0.13	F
70	0.43	0.23	

Appendix D.3. Rasch/1PL

		Infit		Exact Match		
Item	b	MnSq	Zstd	Obs%	Exp%	Flag
1	-0.80	1.02	1.21	68.5	68.5	
2	0.40	0.97	-2.83	62.8	60.9	
3	0.47	0.98	-1.26	62.6	62.0	
4	0.48	1.00	-0.36	61.9	62.2	
5	0.43	1.03	2.20	60.4	61.4	F
6	0.40	0.98	-1.60	62.2	61.0	
7	-0.61	0.99	-0.94	64.8	64.4	
8	0.34	0.96	-3.48	63.4	60.1	
9	-0.07	0.99	-1.11	59.6	57.8	
10	0.49	1.01	0.54	61.7	62.4	
11	-0.60	0.96	-2.89	65.6	64.2	
12	-0.44	0.99	-0.47	61.5	61.4	
13	0.73	1.01	0.69	67.0	67.0	
14	0.82	1.02	1.04	68.8	68.8	
15	-0.20	0.99	-1.06	59.9	58.6	
16	0.42	1.02	1.99	59.4	61.2	
17	1.91	1.00	-0.01	86.5	86.5	
18	0.25	1.03	3.31	55.4	59.1	F
19	-1.24	0.99	-0.38	76.9	76.8	
20	-0.58	0.98	-1.30	64.9	63.9	
21	-1.37	0.99	-0.16	79.1	79.1	
22	0.36	1.02	1.75	57.5	60.4	
23	-1.94	0.99	-0.28	86.9	86.9	
24	-1.36	1.00	-0.06	78.9	78.9	
25	-0.36	1.01	1.22	58.4	60.3	
26	-1.16	1.00	0.09	75.4	75.4	
27	0.65	1.00	-0.16	65.9	65.3	
28	-1.74	0.98	-0.46	84.4	84.4	
29	-0.37	1.00	-0.20	59.8	60.4	
30	0.84	0.99	-0.36	69.3	69.2	
31	-0.43	1.02	1.73	59.9	61.2	
32	0.32	1.02	2.08	58.1	59.8	F
33	-1.14	1.03	1.21	75.1	75.1	
34	0.18	1.04	4.05	56.1	58.4	F
35	0.17	1.04	4.58	54.4	58.3	F
36	-2.68	0.98	-0.20	93.2	93.2	
37	0.15	1.05	4.92	53.5	58.2	F
38	0.51	1.00	0.21	62.6	62.7	
39	-1.75	1.00	0.05	84.6	84.6	
40	-1.01	1.00	-0.03	72.7	72.7	

41	1.07	1.04	1.62	73.9	73.9	
42	1.42	1.03	0.80	79.9	79.9	
43	-0.70	0.98	-1.09	66.6	66.3	
44	-0.50	0.96	-2.65	64.3	62.4	
45	0.15	0.98	-2.42	61.3	58.2	
46	-1.38	1.00	0.04	79.3	79.3	
47	0.11	1.01	1.34	57.1	58.0	
48	-0.24	1.01	1.02	57.2	58.9	
49	-0.48	1.02	1.29	61.9	62.2	
50	-1.77	0.98	-0.42	84.8	84.8	
51	-0.21	1.00	0.00	59.2	58.6	
52	-0.34	1.00	0.07	60.4	60.0	
53	1.77	0.97	-0.58	84.8	84.8	
54	-0.41	1.00	0.37	60.6	61.0	
55	-1.14	1.00	-0.14	75.2	75.2	
56	-0.73	0.98	-1.25	67.0	67.0	
57	-0.11	1.02	2.75	55.9	58.0	F
58	0.65	1.03	1.74	65.0	65.3	
59	0.06	0.96	-5.14	62.6	57.8	
60	0.18	0.96	-4.33	63.7	58.4	
61	-1.48	0.98	-0.62	80.8	80.8	
62	-0.15	0.99	-1.48	60.0	58.2	
63	0.01	1.00	-0.42	59.1	57.7	
64	0.79	0.99	-0.71	68.3	68.3	
65	-0.28	1.01	0.98	58.1	59.3	
66	0.19	1.01	1.46	57.1	58.5	
67	-1.13	0.99	-0.50	74.9	74.9	
68	0.07	1.00	0.45	56.6	57.8	
69	1.20	1.01	0.27	76.2	76.2	
70	0.43	0.98	-1.53	63.2	61.4	

Appendix D.4. IRT 2PL

Item	a	b	Flag
1	0.177	-2.559	F
2	0.340	0.680	
3	0.261	1.028	F
4	0.206	1.306	F
5	0.149	1.560	F
6	0.280	0.826	F
7	0.264	-1.351	F
8	0.348	0.571	

9	0.228	-0.185	F
10	0.194	1.417	F
11	0.393	-0.942	
12	0.238	-1.083	F
13	0.206	1.994	F
14	0.185	2.448	F
15	0.235	-0.518	F
16	0.147	1.545	F
17	0.303	3.644	F
18	0.130	1.024	F
19	0.313	-2.342	
20	0.285	-1.210	F
21	0.265	-3.002	F
22	0.157	1.248	F
23	0.388	-3.011	F
24	0.267	-2.964	F
25	0.169	-1.223	F
26	0.240	-2.783	F
27	0.241	1.533	F
28	0.387	-2.711	
29	0.215	-0.992	F
30	0.240	1.989	F
31	0.151	-1.584	F
32	0.145	1.188	F
36	0.539	-3.110	F
38	0.203	1.397	F
39	0.302	-3.397	F
40	0.248	-2.375	F
43	0.312	-1.342	
44	0.343	-0.885	
45	0.274	0.302	F
46	0.274	-2.945	F
47	0.173	0.322	F
48	0.170	-0.825	F
49	0.169	-1.623	F
50	0.404	-2.659	
51	0.239	-0.943	F
52	0.427	-2.227	
53	0.481	1.734	
54	0.314	-1.680	
55	0.297	-2.580	F
56	0.315	-1.390	
57	0.133	-0.496	F

58	0.158	2.252	F
59	0.398	0.082	
60	0.398	0.260	
61	0.393	-2.288	
62	0.258	-0.358	F
63	0.230	0.011	F
64	0.275	1.656	F
65	0.185	-0.880	F
66	0.163	0.621	F
67	0.302	-2.210	
68	0.212	0.160	F
69	0.252	2.714	F
70	0.286	0.870	F

Appendix D.5 IRT 3PL

Item	a	b	c	Flag
1	0.207	-0.696	0.275	F
2	0.604	1.312	0.225	
3	0.483	1.814	0.235	
4	0.440	2.162	0.250	
5	0.405	2.636	0.287	K
6	0.505	1.600	0.234	
7	0.357	-0.110	0.259	
8	0.581	1.250	0.225	
9	0.356	1.071	0.254	
10	0.444	2.255	0.257	
11	0.515	-0.115	0.248	
12	0.331	0.269	0.261	
13	0.544	2.515	0.250	
14	0.510	3.060	0.254	F
15	0.369	0.737	0.257	
16	0.377	2.745	0.285	F
17	0.838	3.359	0.138	F
18	0.279	2.624	0.271	F
19	0.374	-1.272	0.266	
20	0.370	-0.054	0.259	
21	0.302	-1.798	0.270	
22	0.317	2.604	0.263	K
23	0.460	-2.094	0.263	
24	0.308	-1.750	0.269	
25	0.240	0.640	0.264	F
26	0.275	-1.444	0.271	F
27	0.509	2.193	0.235	
28	0.455	-1.826	0.263	
29	0.307	0.496	0.264	
30	0.588	2.367	0.228	
31	0.199	0.547	0.266	F
32	0.303	2.601	0.265	

36	0.616	-2.445	0.259	
38	0.413	2.351	0.249	
39	0.343	-2.323	0.269	
40	0.290	-1.068	0.267	F
43	0.409	-0.292	0.255	
44	0.454	0.057	0.251	
45	0.471	1.221	0.245	
46	0.315	-1.761	0.269	
47	0.296	1.800	0.256	F
48	0.253	0.967	0.263	F
49	0.220	0.307	0.267	F
50	0.463	-1.847	0.262	
51	0.326	0.396	0.259	
52	0.514	-1.410	0.259	
53	0.918	1.861	0.151	
54	0.394	-0.618	0.259	
55	0.352	-1.459	0.266	
56	0.410	-0.322	0.261	
57	0.209	1.628	0.261	F
58	0.496	3.014	0.284	F
59	0.658	0.782	0.240	
60	0.670	0.909	0.232	
61	0.466	-1.427	0.261	
62	0.398	0.800	0.256	
63	0.361	1.228	0.251	
64	0.587	2.106	0.217	
65	0.269	0.806	0.263	F
66	0.298	2.083	0.258	F
67	0.359	-1.128	0.263	
68	0.358	1.410	0.255	
69	0.668	2.836	0.202	K
70	0.543	1.575	0.235	

Appendix E: Xcalibre 4.0 Input & Output Specifications

File specifications

Specification	Value	Specification	Value
Number of examinees	2320	Total Items	70
Calibrated Items	70	Pretest Items	0
Excluded Items	0	Number of domains	1
Classic Data Header	Yes	Delimited input	No
Delimiter for input	N/A	Number of ID columns	17
ID begins in column	1	Responses begin in column	18
Omit character	N	Not Admin character	N
Save item parameters	Yes	Item parameter format	N/A
Save data matrix	Yes	Include omit codes in matrix	No
Include Not Admin codes in matrix	No	Score Not Admin as omits	No
Plot the IRFs	Yes	Save the IRFs and IIFs	No
Produce the fit line	Yes	# Groups for Plot	15
Type of score groups	Equally sized	Max score group size	200
# Groups for Chi-square	15	Perform classification	No
Classify using	N/A	Two-group cutpoint	N/A
Low group label	N/A	High group label	N/A

IRT Calibration Specifications

Specification	Value	Specification	Value
IRT Specification	Dichotomous	Model constant	1.0
Polytomous IRT Model	N/A	Dichotomous IRT Model	1-parameter
Center the boundary locations	No	Centered value	N/A
Floating priors	Yes	a parameter prior mean (sd)	1.000 (0.300)
b parameter prior mean (sd)	0.000 (1.000)	c parameter prior mean (sd)	0.250 (0.030)
Theta estimation method	MLE	Bayesian prior mean (sd)	N/A
Maximum E-M loops	50	Convergence criterion	0.010
Quadrature points	25	Center dich item parameters on	b
Acceptable P range	0.00 to 1.00	Acceptable item-corr range	0.00 to 1.00
Acceptable item mean range	0.00 to 15.00	Correct for spuriousness	Yes
Fit statistic critical alpha	0.050	Minimum a	0.05
Maximum a	4.00	Minimum b	-4.00
Maximum b	4.00	Minimum c	0.00
Maximum c	0.70	Minimum theta	-7.00
Maximum theta	7.00	Treat scored items as poly	No
Center poly parameters on theta	No	Test for DIF	No
Group status column	N/A	Ability levels for DIF Test	N/A
Group 1 code	N/A	Group 2 code	N/A
Group 1 label	N/A	Group 2 label	N/A
Exclude items with low N	No	Minimum valid N	N/A

Compute scaled scores	No	Mean (SD) of scaled scores	N/A
Save delimited output	Yes	Delimiter	Comma
Save scores output	Yes	Delimiter	Comma
Save delimited output	Yes	Delimiter	Comma

Flag Specifications

Specification	Value	Specification	Value
Low a Flag Bound	0.30	High a Flag Bound	4.00
Low b Flag Bound	-3.00	High b Flag Bound	3.00
Low c Flag Bound	0.00	High c Flag Bound	0.40
Key Flag	K	Fit Flag	F
Low a Flag	La	High a Flag	Ha
Low b Flag	Lb	High b Flag	Hb
Low c Flag	Lc	High c Flag	Hc

References

Adams, R., Griffin, P., & Martin, L. (1987). A latent trait method for measuring a dimension in second language proficiency. *Language Testing, 4*(1), 9-27.

Adedoyin, O. (2010). Investigating the Invariance of Person Parameter Estimates Based on Classical Test and Item Response Theories. *International Journal of Education Science, 2*(2), 107-113.

Anderson, J. (1999). Does Complex Analysis (IRT) Pay Any Dividends in Achievement Testing? *The Alberta Journal of Educational Research, XLV*(4), 344-352.

Ayala, R., & Hertzog, M. (1991). The Assessment of Dimensionality for Use in Item Response Theory. *Multivariate Behavioral Research, 26*(4), 765-792.

Bachman, L. (1982). The Trait Structure of Cloze Test Scores. *TESOL Quarterly, 16*(1), 61-70.

Bachman, L., & Palmer, A. (1981a). The construct validation of the FSI oral interview. *Language Learning, 31*, 67-86.

Bachman, L., & Palmer, A. (1981b). A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading. In A. Palmer & P. Groot (Eds.), *The Construct Validation of Tests of Communicative Competence*. Washington, DC: TESOL.

Bachman, L., & Palmer, A. (1982). The construct validation of some components of communicative competence. *TESOL Quarterly, 16*, 449-465.

Bachman, L., & Palmer, A. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*, 14-29.

Baker, F. (2001). *The Basics of Item Response Theory* (2 ed.): ERIC Clearinghouse on Assessment and Evaluation.

Barnes, L., & Wise, S. (1991). The Utility of a Modified One-Parameter IRT Model with Small Samples. *Applied Measurement in Education, 4*(2), 143-157.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101-118.

Bejar, I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17*(4), 283-296.

Bejar, I. (1988). An approach to assessing unidimensionality revisited. *Applied Psychological Measurement, 12*(4), 377-379.

Betebenner, D., Shang, Y., Xiang, Y., & Zhao, Y. (2008). The Impact of Performance Level Misclassification on the Accuracy and Precision of Percent at Performance Level Measures. *Journal of Educational Measurement, 45*(2), 119-137.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian Random Effects Model for Testlets. *Psychometrika, 64*(2), 153-168.

Braeken, J., Tuerlinckx, F., & De Boeck, P. (2007). Copula functions for residual dependency. *Psychometrika, 72*(3), 393-411.

Bramley, T. (2010). What can be Inferred about Classification Accuracy from Classification Consistency? *Educational Research, 52*(3), 325-330.

Breyer, F., & Lewis, C. (1994). Pass-Fail Reliability for Tests with Cut Scores: A Simplified Method. Princeton, NJ: Educational Testing Service.

Brown, J., & Ahn, R. (2011). Variables that affect the dependability of L2 pragmatics tests. *Journal of Pragmatics, 43*(1), 198-217.

Brown, J., & Bailey, K. (2008). Language Testing Courses: What are they in 2007? *Language Testing, 25*(3), 349-383.

Carr, N. (2006). The Factor Structure of Test Task Characteristics and Examinee Performance. *Language Testing, 23*(3), 269-289.

Childs, R., & Oppler, S. (1999). Practical Implications of Test Dimensionality for Item Response Theory Calibration of the Medical College Admission Test. Washington, DC: American Institutes for Research.

Choi, I. (1989). *An Application of Item Response Theory to Language Testing: Model-data Fit Studies.* University of Illinois at Urbana-Champaign, Urbana, IL.

Choi, I. (1992). An Application of Item Response Theory to Language Testing. In S. Belasco (Ed.), *Theoretical Studies in Second Language Acquisition* (Vol. 2). New York.

Choi, I., & Bachman, L. (1992). An Investigation into the Adequacy of Three IRT Models for Data from Two EFL Reading Tests. *Language Testing, 9*, 51-78.

Chong, H. (2011). A simple guide to Item Response Theory (IRT) and Rasch Modeling Retrieved January 4, 2010, from http://www.creative-wisdom.com

Cook, K., Kallen, M., & Amtmann, D. (2009). Having a Fit: Impact of Number of Items and Distribution of Data on Traditional Criteria for Assessing IRT's Unidimensionality Assumption. *Qual Life Res, 18*, 447-460.

Coombs, C., Milholland, J., & Womer, F. (1956). The assessment of parital knowledge. *Educational and Psychological Measurement, 16*, 13-37.

Cutts, R. (1997). *An Empire of Schools: Japan's Universities and the Molding of a National Power Elite*. Armonk, NY: ME Sharpe, Inc.

Davey, T., & Hirsch, T. (1990). *Examinee discrimination as a measure of test data dimensionality*. Paper presented at the Annual Meeting of the Pyschometric Society, Princeton, NJ.

Davidson, F. (1985). *The factor structure of the Fall 1984 ESLPE*. paper. University of California at Los Angeles. Los Angeles.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367-396.

Dawadi, B. (1999). Robustness of the Polytomous IRT Model to Violations of the Unidimensionality Assumption. Montreal, Quebec.

De Champlain, A. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*, 109-117.

De Jong, J., & Glas, C. (1987). Validation of listening comprehension tests using item response theory. *Language Testing, 4*(2), 170-194.

di Gennaro, K. (2009). Investigating differences in the writing performance of international and Generation 1.5 students. *Language Testing, 26*(4), 533-559.

Dinero, T., & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement, 1*(581-592).

Dorans, N., & Kingston, N. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating on the GRE verbal scale. *Journal of Educational Measurement, 22*(4), 249-262.

Douglas, K., & Mislevy, R. (2010). Estimating Classification Accuracy for Complex Decision Rules Based on Multiple Scores. *Journal of Educational and Behavioral Statistics, 35*(3), 280-306.

Downing, S., & Haladyna, T. (1996). A model for evaluating high-stakes testing programs: why the fox should not guard the chicken coop. *Educational Measurement: Isssues and Practice, 15*(1), 5-12.

Drasgow, F., & Parsons, C. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*(2), 189-199.

Dressel, P., & Schmid, P. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement, 13*, 574-595.

Dwyer, C. (1996). Cut Scores and Testing: Statistics, Judgment, Truth, and Error. *Psychological Assessment, 8*(4), 360-362.

Edelen, M., & Reeve, B. (2007). Applying Item Response Theory (IRT) Modeling to Questionnaire Development, Evaluation, and Refinement. *Qual Life Res, 16*, 5-18.

Embretson, S., & Reise, S. (2000). *Item Response Theory for Psychologists: Multivariate Applications*. Mahwah, NJ: Lawrence Erlbaum.

Ercikan, K., & Julian, M. (2002). Classification Accuracy of Assigning Student Performance to Proficiency Levels: Guidelines for Assessment Design. *Applied Measurement in Education, 15*(3), 269-294.

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educational and Psychological Measurement, 58*, 357-381.

Farhady, H. (1983). On the plausibility of the unitary language proficiency factor. In J. Oller (Ed.), (pp. 11-28). Rowley, MA: Newbury House.

Folk, V., & Green, B. (1989). Adaptive estimation when the unidimensionality of assumption of IRT is violated. *Applied Psychological Measurement, 13*(4), 373-389.

Fouly, K., Bachman, L., & Cziko, G. (1990). The divisibility of language competence: a confirmatory approach. *Language Learning, 40*, 1-21.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading items difficulty: Implications for construct validity. *Language Testing, 10*(2), 133-170.

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

Gibbons, R., Immekus, J., & Bock, R. (2007). The Added Value of Multidimensional IRT Models *Multi-Dimensional and Hierarchial Modeling Monograph*.

Grabe, W., & Stoller, F. (2002). *Teaching and Researching: Reading*. Harlow, UK: Longman.

Guo, F. (2006). Expected Classification Accuracy using the Latent Distribution. *Practical Assessment, Research & Evaluation, 11*(6).

Hakstian, A., & Kansup, W. (1975). A Comparison of Several Methods of Assessing Partial Knowledge in Multiple-Choice Tests: II. Testing Procedures. *Journal of Educational Measurement, 12*(4), 231-239.

Hall, T., Reise, S., & Haviland, M. (2007). An Item Response Theory Analysis of the Spiritual Assessment Inventory. *The International Journal for the Psychology of Religion, 17*(2), 157-178.

Hambleton, R., Crocker, L., Masters, G., van der Linden, W., & Wright, B. (1992). Commentary under Thesis 5. *Rasch Measurement Transactions* Retrieved February 27, 2010, from http://www.rasch.org/rmt/rmt62d.htm

Hambleton, R., & Rovinelli, R. (1986). Assessing the Dimensionality of a Set of Test Items.

Hambleton, R., & Slater, S. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education, 10*(1), 19-38.

Hambleton, R., & Swaminathan, H. (1985). *Item Response Theory: Principles & Applications*. Boston: Kluwer-Nijhoff.

Hambleton, R., Swaminathan, H., Cook, L., Eignor, D., & Gifford, J. (1978). Developments in latent trait theory: models, technical issues, and applications. *Review of Educational Research, 48*(4), 467-510.

Hambleton, R., & Traub, R. (1971). Information curves and efficiency of three logistic test models. *British Journal of Mathematical and Statistical Psychology, 24*, 273-281.

Hambleton, R., & Traub, R. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology, 26*, 195-211.

Hartig, J., & Hohler, J. (2009). Multidimensional IRT Models for the Assessment of Competencies. *Studies in Educational Evaluation, 35*, 57-63.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49-78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-163.

Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing, 1*, 237-241.

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research*. Boston: Heinle & Heinle.

Henning, G. (1988). The Influence of Test and Sample Dimensionality on Latent Trait Person Ability and Item Difficulty Calibrations *Educational Testing Service* (pp. 83-99).

Henning, G. (1989). Meanings and Implications of the Principle of Local Independence *Educational Testing Service* (pp. 95-108).

Henning, G. (1992). Dimensionality and Construct Validity of Language Tests. *Language Testing, 9*(1), 1-11.

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality. *Language Testing, 2*, 141-154.

Hernandez, R. (2009). Comparison of the Item Discrimination and Item Difficulty of the Quick-Mental Aptitude Test Using CTT and IRT Methods. *The International Journal of Educational and Psychological Assessment, 1*(1), 12-18.

Hoffman, G., & Wise, L. (2000). Establishing the Reliability of Student Proficiency Classifications: The Accuracy of Observed Classifications. Alexandria, VA: Human Resources Research Organization.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local item dependences among test items. *Psychological Methods, 2*, 261-277.

Hulin, C., Lissak, R., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*(3), 249-260.

Huynh, H. (1976). On the Reliability of Decisions in Domain-Referenced Testing. *Journal of Educational Measurement, 13*(4), 253-264.

Ip, E. (2000). Adjusting for information inflation due to local dependence in moderately large item clusters. *Psychometrika, 65*, 73-91.

Ip, E. (2002). Locally dependent latent trait model and the Dutch identity revisited. *Psychometrika, 67*, 367-386.

Ip, E. (2010). Empirically Indistinguishable Multidimensional IRT and Locally Dependent Unidimensional Item Response Models. *British Journal of Mathematical and Statistical Psychology, 63*, 395-416.

Ip, E., Smits, D., & De Boeck, P. (2009). Locally dependent linear logistic test model with person covariates. *Applied Psychological Measurement, 33*(7), 555-569.

Kansup, W., & Hakstian, A. (1975). A Comparison of Several Methods of Assessing Partial Knowledge in Multiple-Choice Tests: I. Scoring Procedures. *Journal of Educational Measurement, 12*(4), 219-230.

Keller, L., Swaminathan, H., & Sireci, S. (2003). Evaluating Scoring Procedures for Context-Dependent Item Sets. *Applied Measurement in Education, 16*(3), 207-222.

Kelley, T. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*(1), 17-24.

Kim, Y. (2009). An investigation into native and non-native teachers' judgments of oral English performance. *Language Testing, 26*(2), 187-217.

Kolen, M., Zeng, L., & Hanson, B. (1992). Conditional Standard Errors of Measurement for Scale Scores Using IRT. *Journal of Educational Measurement, 33*(2), 129-140.

Kozaki, Y. (2010). An alternative decision-making procedure for performance assessments: using the multifaceted Rasch model to generate cut estimates. *Language Assessment Quarterly, 7*(1), 75-95.

Lawson, S. (1991). One parameter lateent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), *Advances in Educational Research: Substantive Findings, Methodological Developments* (pp. 159-168). Greenwhich, CT: JAI.

Lee, W., Brennan, R., & Wan, L. (2009). Classification Consistency and Accuracy for Complex Assessments Under the Compound Multinomial Model. *Applied Psychological Measurement, 33*(5), 374-390.

Lee, W., Hanson, B., & Brennan, R. (2000). Procedures for Computing Classification Consistency and Accuracy Indices with Multiple Categories. Iowa City, IA: American College.

Lee, W., Hanson, B., & Brennan, R. (2002). Estimating Consistency and Accuracy Indices for Multiple Classifications. *Applied Psychological Measurement, 26*(4), 412-432.

Lee, Y. (2004). Examining Passage-Related Local Item Dependence and Measurement Construct Using Q3 Statistics in an EFL Reading Comprehension Test. *Educational Testing Service, 21*(1), 74-100.

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing, 26*(2), 245-274.

Li, S., & Sireci, S. (2005). *Evaluating the accuracy of proficiency classifications using item response theory*. Paper presented at the National Council on Measurement in Education, Montreal, Quebec, Canada.

Linacre, J. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. (2005). Virtual equating. *Transactions of the Rasch Measurement SIG, 19*(3), 1025-1032.

Linn, R. (1990). Has Item Response Theory Increased the Validity of Achievement Test Scores? *Applied Measurement in Education, 3*(2), 115-141.

Lissitz, R., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448.

Livingston, S., & Lewis, C. (1995). Estimating the Consistency and Accuracy of Classifications Based on Test Scores. *Journal of Educational Measurement, 32*(2), 179-197.

Livingston, S., & Wingersky, M. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement, 16*, 247-260.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: a multitrait-multimethod approach. *Language Testing, 24*, 489-515.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom of English proficiency based on teacher judgments *Educational Measurment: Issues and Practices*: National Council on Measurement in Education.

Lord, F., & Novick, M. (1968). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lunzer, E., Waite, M., & Dolan, T. (1979). Comprehension and comprehension tests. In E. Lunzer & K. Gardner (Eds.), *The Effective Use of Reading*. London: Heinemann Educational Books.

Lynch, B., Davidson, F., & Henning, G. (1988). Person dimensionality and language test validation. *Language Testing, 5*(2), 206-219.

Matthews, T. (1992). *Development of a Foreign Language Placement Test Using Item Response Scoring on a Multiple-choice Cloze Test.* University of Delaware.

McCall, M. (2002). *Analysis of the Characteristics of a Large-Scale Reading Assessment.* (Dissertation), Portland State University.

McDonald, R. (1982). Linear versus non-linear models in item response theory. *Applied Psychological Measurement, 6*, 379-396.

McNamara, T. (1990). Item Response Theory and the Validation of an ESP Test for Health Professionals. *Language Testing, 7*(1), 52-75.

McNamara, T. (1991). Test Dimensionality: IRT Analysis of an ESP Listening Test. *Language Testing, 8*(2), 139-159.

Meara, K., Robin, F., & Sireci, S. (2000). Using Multidimensional Scaling to Assess the Dimensionality of Dichotomous Item Data. *Multivariate Behavioral Research, 35*(2), 229-259.

Messick, S. (1965). Personality measurement and the ethics of assessment. *American Psychologist, 20*, 136-142.

Messick, S. (1980). Test Validity and the Ethics of Assessment. *American Psychologist, 35*(11), 1012-1027.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education.

Mislevy, R., & Bock, R. (1982). BILOG: Item Analysis and Test Scoring with Binary Logistic Models. Chicago: Scientific Software.

Newton, P. (2009). The reliability of results from national curriculum testing in England. *Educational Research, 51*(2), 181-212.

Novick, M. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1-18.

Nystrom, P. (2004). Reliability of Educational Assessments: The case of classification accuracy. *Scandinavian Journal of Educational Research, 48*(4), 427-440.

Oller, J. (1976). Evidence for a general language proficiency factor: an expetancy grammar. *Die Neueren Sprachen, 75*, 165-174.

Oller, J. (1983). Evidence for a general language proficiency factor: an expectancy grammar. In J. Oller (Ed.), *Issues in Language Testing Research* (pp. 3-10). Rowley, MA: Newbury House.

Oshima, T., & Miller, M. (1990). Multidimensionality and IRT-based item invariance indexes: the effect of between-group variation in trait correlation. *Journal of Educational Measurement, 27*(3), 273-283.

Partchev, I. (2004). *A Visual Guide to Item Response Theory*.

Perkins, K., & Miller, L. (1984). Comparative analyses of English as a second language reading comprehension data. *Language Testing, 1*, 21-32.

Prieto, G., Delgado, A., Perea, M., & Ladera, V. (2010). Scoring Neuropsychological Tests Using the Rasch Model: An Illustrative Example with the Rey-Osterrieth Complex Figure. *The Clinical Neuropsychologist, 24*, 45-56.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 4*(3), 207-230.

Reckase, M. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36.

Reckase, M., Ackerman, T., & Carlson, J. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*(3), 193-203.

Reckase, M., Carlson, J., & Ackerman, T. (1985). When Unidimensional Data are not Unidimensional. Arlington, VA: Office of Naval Research.

Reeve, B., Hays, R., Chang, C., & Perfetto, E. (2007). Applying Item Response Theory to Enhance Health Outcomes Assessment. *Qual Life Res, 16*, 1-3.

Reise, S., Ainsworth, A., & Haviland, M. (2005). Fundamentals, Applications, and Promise in Psychological Research. *Current Directions in Psychological Science, 14*(2), 95-101.

Reise, S., & Haviland, M. (2005). Item Response Theory and the Measurement of Clinical Change. *Journal of Personality Assessment, 84*(3), 228-238.

Reise, S., & Henson, J. (2003). A Discussion of Modern Versus Traditional Psychometrics as Applied to Personality Assessment Scales. *Journal of Personality Assessment, 81*(2), 93-103.

Rijmen, F. (2009). Three Multidimensional Models for Testlet-Based Tests: Formal Relations and an Empirical Comparison. Princeton, NJ: ETS.

Rijmen, F. (2010). Formal Relations and an Empirical Comparison Among the Bi-Factor, the Testlet, and a Second-Oder Multidimensional IRT Model. *Journal of Educational Measurement, 47*(3), 361-372.

Ross, S. (2011). [Personal communication].

Rudner, L. (2001). Measurement Decision Theory. Washington, DC: National Inst. on Student Achievement, Curriculum, and Assessment.

Rudner, L. (2002). *An Examination of Decision-Theory Adaptive Testing Procedures*. Paper presented at the American Educational Research Association (AERA), New Orleans, LA.

Rudner, L. (2003). *The classification accuracy of measurement decision theory*. Paper presented at the National Council on Measurement in Education (NCME), Chicago, IL.

Rudner, L. (2005). Expected Classification Accuracy. *Practical Assessment, Research & Evaluation, 10*(13).

Sang, F., Schmitz, B., Vollmer, H., Baumert, J., & Roeder, P. (1986). Models of second language competence: a structural equation approach. *Language Testing, 3*, 54-79.

Sawaki, Y. (2003). *A Comparison of Summarization and Free Recall as Reading Comprehension Tasks in Web-Based Assessment of Japanese as a Foreign Language.* (PhD), UCLA, Los Angeles.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: reporting a score profile and a composite. *Language Testing, 24*, 355-390.

Sawaki, Y., Stricker, L., & Oranje, A. (2009a). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*, 5-30.

Sawaki, Y., Stricker, L., & Oranje, A. (2009b). Factor Structure of the TOEFL Internet-based Test. *Language Testing, 26*(1), 5-30.

Schaeffer, B. (2010, September 10, 2010). FairTest Reacts to the 2010 SAT Scores Retrieved March 18, 2011, from http://fairtest.org/press-release-2010-sat-scores

Schedl, M., Gordon, A., Carey, P., & Tang, K. (1996). An Analysis of the Dimensionality of TOEFL Reading Comprehension Items. Princeton, NJ: Educational Testing Service.

Scherbaum, C., Finlinson, S., Barden, K., & Tamanini, K. (2006). Applications of Item Response Theory to Measurement Issues in Leadership Research. *The Leadership Quarterly, 17*, 366-386.

Schulz, E., Kolen, M., & Nicewander, W. (1999). A Rationale for Defining Achievement Levels Using IRT-Estimated Domain Scores. *Applied Psychological Measurement 23*(4), 347-362.

Shin, D. (1999). *Construct Validation of a Diagnostic L2 Listening Test: An Operational Model Utilized and Multidimensionality Issues Revisited.* (PhD), University of Illinois, Urbana Champagne.

Shin, S. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing, 22*, 31-57.

Silva, S. (1985). A Comparison of Traditional Approaches and Item Response Approaches to the Problem of Item Selection for Criterion-Referenced Measurement. Chicago, IL: American Educational Research Association.

Sireci, S. (1991). *"Sample-Independent" Item Parameters? An Investigation of the Stability of IRT Item Parameters Estimated from Small Data Sets.* Ellenville, NY.

Sireci, S., Thissen, D., & Wainer, H. (1991). On the Reliability of Testlet-Based Tests. *Journal of Educational Measurement, 28*(3), 237-247.

Skorupski, W., & Carvajal, J. (2010). A Comparison of Approaches for Improving the Reliability of Objective Level Scores. *Educational and Psychological Measurement, 70*(3), 357-375.

Spolsky, B. (1978). Introduction: Linguists and language testers *Advances in Language Testing: Series 2, Approaches to Language Testing*. Arlington, VA: Center for Applied Linguistics.

Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Braley & D. Stevenson (Eds.), *Practice and Problems in Language Testing* (Vol. 1, pp. 3-50).

Stone, C., Weissman, A., & Lane, S. (2005). The Consistency of Student Proficiency Classifications Under Competing IRT Models. *Educational Assessment, 10*(2), 125-146.

Stone, C., Ye, F., Zhu, X., & Lane, S. (2010). Providing Subscale Scores for Diagnostic Information: A Case Study When the Test is Essentially Unidimensional. *Applied Measurement in Education, 23*, 63-86.

Stout, W. (1984). A statistical procedure for assessing test dimensionality. *Measurement Series, 84*(2).

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Streiner, D. (2010). Measure for Measure: New Developments in Measurement and Item Response Theory. *Research Methods in Psychiatry, 55*(3), 180-186.

Subkoviak, M. (1976a). Estimating Reliability from a Single Administration of a Criterion-Referenced Test. *Journal of Educational Measurement 13*(4), 265-276.

Subkoviak, M. (1976b). *Estimating Reliability from a Single Administration of a Mastery Test.* The University of Wisconsin.

Subkoviak, M. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55.

Taylor, C., & Lee, Y. (2010). Stability of Rasch scales over time. *Applied Measurement in Education, 23*, 87-113.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: a use of multiple-categorical-response models. *Journal of Educational Measurement, 26*, 247-260.

Thorndike, R. (1982). The improvement of measurement in education and psychology: Contributions of laten trait theory. In D. Spearritt (Ed.), *Educational Measurement: Theory and Practice* (pp. 3-13). Princeton, NJ: ERIC Clearinghouse of Tests, Measurements, and Evaluations.

Turner, C. (1989). The Underlying Factor Structure of L2 Close Test Performance in Francophone, University-Level Students: Casual Modelling as an Approach to Construct Validation *L2 Cloze Test Performance* (pp. 172-197).

Unick, G., & Stone, S. (2010). State of Modern Measurement Approaches in Social Work Research Literature. *Social Work Research, 34*(2), 94-101.

van de Vijver, F. (1986). The robustness of Rasch measurements. *Applied Psychological Measurement, 10*, 45-57.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287-307.

von Davier, M. (2009). Is there need for the 3PL model? Guess what? *Measurement: Interdisciplinary Research and Perspectives, 7*(2), 110-114.

Wainer, H., Bradlow, E., & Wang, X. (2007). *Testlet Response Theory and Its Applications*. New York: Cambridge University Press.

Wainer, H., & Kiely, G. (1986). CATs, Testlets, and Test Construction: A Rationale for Putting Test Developers Back into CAT (pp. 46). Princeton, NJ: ETS.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: a case for testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Wainer, H., & Thissen, D. (1985). Estimating ability with the wrong model (pp. 80). Brooks AFB, TX: Educational Testing Service.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12*(4), 339-368.

Walker, L., & Beretvas, N. (2000). *Using multidimensional versus unidimensional ability estimates to determine student proficiency in mathematics*. Paper presented at the American Educational Research Association, New Orleans.

Wang, W., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*(4), 296-318.

Wang, X., Bradlow, E., & Wainer, H. (2002). A General Bayesian Model for Testlets: Theory and Applications. *Applied Psychological Measurement, 26*(1), 109-128.

Wang, X., Bradlow, E., & Wainer, H. (2005). SCORIGHT (Version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis *Research & Development (RR-04-49)*. Princeton, NJ: Educational Testing Service.

Wilcox, R. (1981). A review of the beta-binomial model and its extensions. *Journal of Educational Statistics, 6*, 3-32.

Wood, R., & Lord, F. (1976). A User's Guide to LOGIST. Princeton, NJ: Educational Testing Service.

Wright, B., & Stone, M. (1979). *Best test design*. Chicago: MESA Press.

Wyse, A., & Mapuranga, R. (2009). Differential Item Functioning Analysis Using Rasch Item Information Functions. *International Journal of Testing, 9*, 333-357.

Yen, S., & Walker, L. (2007). *Multidimensional IRT Models for Composite Scores.* Chicago.

Yen, W. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17*, 297-311.

Yen, W. (1993). Scaling Performance Assessments: Strategies of Managing Local Item Dependence. *Journal of Educational Measurement, 30*(3), 187-213.

Yen, W., & Fitzpatrick, A. (2006). Item Response Theory. In R. Brennan (Ed.), *Educational Measurement: Issues and Practices* (Fourth ed.). Phoenix: Oryx Press.

Zenisky, A., Hambleton, R., & Robin, F. (2004). DIF Detection and Interpretation in Large-Scale Science Assessments: Informing Item Writing Practices. *Educational Assessment, 9*(1 & 2), 61-78.

Zhang, B. (2008). Application of Unidimensional Item Response Models to Tests with Items Sensitive to Secondary Dimensions. *The Journal of Experimental Education, 77*(2), 147-166.

Zhang, B. (2010). Assessing the Accuracy and Consistency of Language Proficiency Classification Under Competing Measurement Models. *Language Testing, 27*(1), 119-140.

Zhang, J. (2004). Comparison of Unidimensional and Multidimensional Approaches to IRT Parameter Estimation.