# Assessing the Feasibility of Using Item Response Theory to Analyse British Columbia's Foundation Skills Assessment 2000

by

Dean Palmi Goodman
B.A., Simon Fraser University

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF ARTS

in the Department of Educational Psychology and Leadership Studies

We accept this thesis as conforming
to the required standard

[redacted]

Dr. J. O. Anderson, Supervisor (Department of Educational Psychology and Leadership Studies)

[redacted]

Dr. W.J.C. Walsh, Departmental Member (Department of Educational Psychology and Leadership Studies)

[redacted]

Dr. R. E. Graves, Outside Member (Department of Psychology)

[redacted]

Dr. I. J. Cameron, External Examiner (Department of Curriculum and Instruction)

© Dean Palmi Goodman, 2001
University of Victoria

All rights reserved. This thesis may not be reproduced in whole or in part, by photocopy or other means, without the permission of the author.

Supervisor: Dr. John O. Anderson

## ABSTRACT

This study assessed the feasibility of using item response theory (IRT) to analyse British Columbia's Foundation Skills Assessment (FSA) 2000, a recent addition to the BC Ministry of Education's provincial learning assessment program. The purposes of this study were to determine the extent to which data from the Grade 7 FSA 2000 tests met the assumptions and properties of IRT and which of several popular item response models best fit Grade 7 FSA 2000 data. The results of this study suggested the reading comprehension and numeracy data sets of FSA 2000 did not meet key assumptions and properties of IRT. Evidence was obtained to suggest the two-parameter logistic model and generalized partial credit model fit the data sets better than other IRT models; however neither fit the data well. Given the discouraging and limited results, further research into the feasibility of using IRT on FSA data was recommended.

[redacted]

Dr. J. O. Anderson, Supervisor (Department of Educational Psychology and Leadership Studies)

[redacted]

Dr. W.J.C. Walsh, Departmental Member (Department of Educational Psychology and Leadership Studies)

[redacted]

Dr. R. E. Graves, Outside Member (Department of Psychology)

[redacted]

Dr. I. J. Cameron, External Examiner (Department of Curriculum and Instruction)

TABLE OF CONTENTS

ABSTRACT ..... ii

TABLE OF CONTENTS ..... iii

LIST OF TABLES ..... vii

LIST OF FIGURES ..... viii

CHAPTER 1: INTRODUCTION ..... 1

Purpose ..... 4

CHAPTER 2: REVIEW OF THE LITERATURE ..... 6

Provincial Assessments in British Columbia–Background ..... 6

Purpose of the Foundation Skills Assessment ..... 8

Comparing Performance on FSA Over Time: Assessment Benchmarks ..... 9

The FSA Instrument ..... 10

The Traditional Approach to Analysing PLAP Assessment Data: Classical Test Theory ..... 13

Limitations of CTT ..... 14

Item Response Theory: An Alternative Approach for Analysing FSA Data ..... 16

Advantages of IRT ..... 17

Disadvantage of IRT: Strong Assumptions ..... 18

Unidimensionality ..... 19

Local Independence ..... 20

Model-Data Fit ..... 21

Nonspeededness ..... 22

Equal Discrimination Indices....22
Lower Asymptote of Zero (Minimal Guessing)....23
Item and Ability Parameter Invariance....24
Popular Item Response Models....25
Dichotomous Item Response Models....26
Polytomous Item Response Models....34
CHAPTER 3: METHOD....43
Data Set....43
Testing the Assumptions of IRT....45
Unidimensionality....45
Local Independence....48
Nonspeededness....49
Equal Item Discrimination....50
Lower Asymptote....51
Item and Ability Parameter Invariance....52
Model-Data Fit....54
CHAPTER 4: RESULTS....56
Unidimensionality....56
Local Independence....59
Local Independence of Reading Comprehension Items....61
Local Independence of Numeracy Items....62
Nonspeededness....63

Reading Comprehension ..... 63
Numeracy ..... 63
Equal Item Discrimination ..... 64
Lower Asymptote ..... 64
Item and Ability Parameter Invariance ..... 67
Difficulties Addressed Prior to Invariance Investigation ..... 67
Problems With Conducting One Parameter Analyses With PARSCALE ..... 67
Sampling Precautions and Estimation Difficulties ..... 67
Results From the Item and Ability Parameter Invariance Investigation ..... 70
Item Parameter Invariance ..... 70
Ability Parameter Invariance ..... 76
Model-Data Fit ..... 84
CHAPTER 5: SUMMARY, DISCUSSION, AND CONCLUSION ..... 89
Summary ..... 89
Discussion ..... 90
Unidimensionality ..... 90
Local Independence ..... 92
Nonspeededness ..... 93
Equal Item Discrimination ..... 94
Lower Asymptote ..... 95
Item and Ability Parameter Invariance ..... 96
Item Parameter Invariance ..... 97

Ability Parameter Invariance .....99

Model-Data Fit .....100

Relative Fit of Dichotomous IRT Models .....101

Relative Fit of Polytomous IRT Models .....102

Conclusion .....103

REFERENCES .....106

APPENDIX A .....111

APPENDIX B .....118

# LIST OF TABLES

Table 1: Percentage of Variance Accounted for the 5 Largest Components of the Reading Comprehension and Numeracy Data Sets....57

Table 2: Poly-DIMTEST Results Testing Departure From Unidimensionality for Reading Comprehension and Numeracy Data Sets....60

Table 3: Summary Point-Biserial and Point-Polyserial Correlation Statistics....65

Table 4: Proportion of Students in the Lowest Decile Who Answered the Five Most Difficult Multiple-Choice Questions Correctly....66

Table 5: Reliability Measures for Total Test and Dichotomous and Polytomous Items (Reading Comprehension)....83

Table 6: Reliability Measures for Total Test and Dichotomous and Polytomous Items (Numeracy)....83

Table 7: Summary of Reading Comprehension Item and Test Fit by IRT Model and Sample Size....85

Table 8: Summary of Numeracy Item and Test Fit by IRT Model and Sample Size....86

Table 9: Differences in Fit Statistics Between Pairs of IRT Models for Reading Comprehension Data Sets of Various Sample Sizes....87

Table 10: Differences in Fit Statistics Between Pairs of IRT Models for Numeracy Data Sets of Various Sample Sizes....88

# LIST OF FIGURES

Figure 1: 1PL item characteristic curves at three levels of difficulty. .....28

Figure 2: 2PL item characteristic curves with three levels of discrimination. .....30

Figure 3: 3PL item characteristic curves with 3 different pseudo guessing parameters. ...33

Figure 4: Category characteristic curves for a polytomous item ($a_i$ = 1.0, $b_{i1}$ = -0.75, $b_{i2}$ = 0.75). .....37

Figure 5: Two four-category polytomous items with $a$ = (1.0, 0.7). .....40

Figure 6: Scree plots of eigenvalues for all components of the reading comprehension and numeracy tests. .....56

Figure 7a: Scatterplots and related correlations of reading comprehension item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 1PL-PCM. .....71

Figure 7b: Scatterplots and related correlations of reading comprehension item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 2PL-PCM. .....72

Figure 7c: Scatterplots and related correlations of reading comprehension item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 2PL-GPCM. .....72

Figure 7d: Scatterplots and related correlations of reading comprehension item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 3PL-PCM. .....73

Figure 7e: Scatterplots and related correlations of reading comprehension item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 3PL-GPCM. .....73

Figure 8a: Scatterplots and related correlations of numeracy item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 1PL-PCM. .....74

Figure 8b: Scatterplots and related correlations of numeracy item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 2PL-PCM. .....74

Figure 8c: Scatterplots and related correlations of numeracy item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 2PL-GPCM. .....75

Figure 8d: Scatterplots and related correlations of numeracy item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 3PL-PCM. .....75

Figure 8e: Scatterplots and related correlations of numeracy item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 3PL-GPCM. .....76

Figure 9a: Scatterplots and related correlations of reading comprehension ability estimates for Odd/Even and Easy/Hard Items using 1PL-PCM. .....77

Figure 9b: Scatterplots and related correlations of reading comprehension ability estimates for Odd/Even and Easy/Hard Items using 2PL-PCM. .....78

Figure 9c: Scatterplots and related correlations of reading comprehension ability estimates for Odd/Even and Easy/Hard Items using 2PL-GPCM.....78

Figure 9d: Scatterplots and related correlations of reading comprehension ability estimates for Odd/Even and Easy/Hard Items using 3PL-PCM. .....79

Figure 9e: Scatterplots and related correlations of reading comprehension ability estimates for Odd/Even and Easy/Hard Items using 3PL-GPCM.....79

Figure 10a: Scatterplots and related correlations of numeracy ability estimates for Odd/Even and Easy/Hard Items using 1PL-PCM. ..... 80

Figure 10b: Scatterplots and related correlations of numeracy ability estimates for Odd/Even and Easy/Hard Items using 2PL-PCM. ..... 80

Figure 10c: Scatterplots and related correlations of numeracy ability estimates for Odd/Even and Easy/Hard Items using 2PL-GPCM. ..... 81

Figure 10d: Scatterplots and related correlations of numeracy ability estimates for Odd/Even and Easy/Hard Items using 3PL-PCM. ..... 81

Figure 10e: Scatterplots and related correlations of numeracy ability estimates for Odd/Even and Easy/Hard Items using 3PL-GPCM. ..... 82

# CHAPTER 1

# INTRODUCTION

As part of its provincial learning assessment program (PLAP), British Columbia's Ministry of Education regularly assesses the extent to which students are attaining the knowledge and skills expected at various levels of the Kindergarten to Grade 12 education system. Since 1976, this program has been defined by a series of subject-specific assessments administered to students in selected grades (usually Grades 4, 7, and 10) on a four or five year cycle. Results of these assessments were generally reported as average scores at the provincial-, district-, and, in most cases, school-levels. In 1999, the PLAP expanded with the implementation of a new assessment, the Foundation Skills Assessment (FSA). The FSA differs from past provincial assessments in a number of ways, including its content (the cross-curricular skills of reading comprehension, writing, and numeracy), its administrative cycle (annual), and its level and method of reporting (reporting provincial, district, school, and, starting in 2000, individual student results in relation to provincial standards of performance).

The unique features of the FSA give rise to some significant practical and technical challenges. As FSA tests will be composed primarily of different questions from year to year (with approximately 20% overlap), steps need to be taken to ensure results on different FSA administrations can be meaningfully compared from one year to the next. The issue of comparability is further heightened with the reporting of FSA results in relation to provincial standards, which connote the expectation of consistent measures across different administrations of the assessment. The provision of individual student results on FSA, and the increased impact these results may have on individual

students' learning, bring a further challenge and increased responsibility to ensure that individual results are calculated in the most accurate manner possible. As a means to address these challenges, this study reviewed the manner in which provincial assessments have traditionally been analysed and explored the feasibility of using a more sophisticated approach to analyse FSA data based on modern test theory.

Past PLAP assessments have relied on a traditional method of analysis based on classical test theory (CTT). In general, CTT is based on the postulate that any observed score on a test ($X$) may be envisioned as the composite of two hypothetical components—a true score ($T$) and a random error component ($E$)—expressed in the form $X = T + E$ (Crocker & Algina, 1986). Performance in the particular domain assessed by a given PLAP assessment was reported as the average observed score obtained on the test, with higher observed scores assumed to represent higher ability in the domain of interest.

While providing an adequate and easy-to-understand foundation for analysing past provincial assessments, key shortcomings of CTT (documented in such literature as Crocker and Algina [1986], Embretson and Reise [2000], Hambleton and Swaminathan [1985], Hambleton, Swaminathan, and Rogers [1991], and Lord [1980]) raise doubts about its continued utility to the FSA. Of primary concern, CTT does not readily enable the comparison of different groups of students on different tests (as would be required when administering different FSA tests over time). A further limitation of CTT results from the calculation and application of a standard error of measurement that is assumed to be equal at all score levels. Hambleton et al. (1991) point out the implausibility of this assumption, arguing that "scores on any test are unequally precise measures for examinees of different ability" (p. 4). Given that individual FSA results will incorporate

a measure of error, procedures to tailor this error to individual student scores should explored.

A promising alternative for analysing FSA data is through a more modern measurement approach known as item response theory (IRT). This approach to item and test analysis attempts to model the relationship between an unobserved variable (e.g., examinee ability) and the probability of the examinee correctly responding to a given test item (Harris, 1989). While not exhaustive, two features of IRT hold significant promise for FSA. Of primary importance is IRT's ability to place both test items and examinee scores on a common scale of measurement (Baker & Al-Karni, 1991). A second feature of IRT that would be advantageous for FSA is the possible application of a unique error of measurement for different levels of examinee ability (Embretson & Reise, 2000, p. 185).

IRT is increasingly being used by many large-scale assessment developers to resolve practical measurement problems such as those faced with the FSA (see Keeves, 1994, pp. 90-93; Kolen & Brennan, 1995, p. 156; Hambleton & Swaminathan, 1985, p. 8). However, while gaining in popularity, the appropriateness of IRT for analysing assessment data is neither simply determined or guaranteed; a number of significant assumptions and conditions must be considered and met before the benefits of IRT can accrue (Rogers & Hattie, 1987). Of particular importance to IRT models are assumptions about dimensionality (usually, that only one dimension or ability is measured by items that make up the test) and local independence (i.e., after taking into account examinee ability, examinees' responses to any pair of items are statistically independent) (Hambleton, 1989). The ability of an item response model to adequately predict or fit the

test data of interest is also critical to the proper application of IRT; as stated by Rogers and Hattie (1987), "without acceptable fit, the validity of the results cannot be ensured" (p. 47).

Related to the above are several other assumptions and properties that should be considered in any application of IRT. These include non-speeded test administrations (i.e., students should have sufficient time to respond to all items on the test), the distribution of item-test score correlations (i.e., equal or unequal discrimination indices), the extent to which guessing takes place on the test, and, the cornerstone of IRT, the existence item and ability invariance (i.e., the property where "the parameters that characterize an item do not depend on the ability distribution of the examinees and the parameter that characterizes an examinee does not depend on the set of test items" [Hambleton et al., 1991, p. 18]).

## Purpose

Although IRT is widely used in test development and analysis activities, its assumptions and expected properties are not adequately investigated in many applications reported in the literature (Hambleton et al., 1991, p. 53). By investigating the assumptions and properties relevant to the application of item response models to FSA data, this study aimed to avoid these shortcomings and help ensure that any application of IRT to the FSA is based upon sound measurement principles.

To facilitate a comprehensive yet parsimonious analysis, only one of the three grades assessed by FSA 2000 was investigated in the present study (with the intention that findings from this preliminary study could inform future studies on the applicability

of IRT to FSA data). Grade 7 was selected as the focus of this study given its medial position between the other two grades.

Specifically, the purposes of this study were to determine:

1) the extent to which data from the Grade 7 FSA 2000 tests meet assumptions and properties relevant to IRT (i.e., unidimensionality, local independence, nonspeededness, possible equal item discrimination, possible non-zero lower asymptote, and item and ability parameter invariance), and
2) which of several popular item response models best fit Grade 7 FSA 2000 data.

# CHAPTER 2

# REVIEW OF THE LITERATURE

British Columbia's Ministry of Education has routinely administered provincial assessments to selected grades of students (usually Grades 4, 7, and 10) as part of its provincial learning assessment program. In 1999, the Ministry implemented a new assessment called the Foundation Skills Assessment (FSA). While sharing some similarities with past assessments, the FSA marked a significant departure from the types of information traditionally available through provincial assessments. The following section of this paper will provide some background to provincial assessments in British Columbia and details on the FSA, including its purpose and a description of the test instruments. It will also describe how provincial assessments in BC have traditionally been analysed through classical test theory and provide a review of an alternative measurement system known as item response theory that may better suit the needs of the FSA.

## Provincial Assessments in British Columbia–Background

Since 1976, the BC Ministry of Education has administered provincial assessments as a means to assess the effectiveness of educational programs in British Columbia. These assessments have generally been linked to particular subjects in the provincial curriculum (e.g., Language Arts, Mathematics, Science, or Social Studies), with each subject area assessed on a cyclical basis approximately every four or five years. The results of these assessments were usually available at both the provincial and district levels, with school level information provided in many cases.

In 1999, a new provincial assessment, the FSA, was introduced (Ministry of Education, 1999). This assessment tests student ability in reading comprehension, writing, and numeracy on an annual basis. With limited exceptions (e.g., students with special needs or English as a second language students who, even with appropriate adaptations, cannot respond meaningfully to the assessment), all BC students in grades 4, 7, and 10 write the FSA tests.

Several key aspects of the FSA differentiate it from past provincial learning assessments. First, instead of assessing specific subject areas (e.g., science, social studies, language arts, mathematics), the FSA focuses on the key skills embedded across the provincial curricula. Second, the FSA will be administered annually, rather than every four or five years. Third, in addition to provincial, district, and school results, individual students results will be available for all students who write the FSA in 2000 and annually thereafter (Ministry of Education, 1999, p. 28).

A further difference between the FSA and other provincial learning assessments is the manner in which test results are reported. Prior to 1999, provincial learning assessment results were communicated as mean scores on a given test. For each FSA, however, all results (i.e., provincial, district, school, and, starting in 2000, individual student results) will be reported in relation to provincial standards of performance. These standards will be based on the recommendations of provincial standard-setting panels, comprising grade-level teachers from across the province, that carefully review the tests before assigning two scores they feel represent the following levels of performance (Ministry of Education, 2000a):

1. *Meets expectations* provincial standard, defined as "the level of performance where a student meets or exceeds the widely held expectations for the grade on this test. With no other information, this is the level below which a teacher would want to know more about the reasons for a student's low performance." (p. 16)

2. *Exceeds expectations* provincial standard, defined as "the level of a student's performance that is beyond that at which a teacher would say the student has fully met the expectations of the grade on this test. Student performance would be considered excellent for the grade on this test." (p. 16)

The BC Ministry of Education reviews the recommendations of the teachers in light of actual student performance on the test and sets the provincial standards against which provincial, district, school, and individual student FSA results will be reported. These standards will enable student performance on the FSA to be classified into one of three categories: not yet within expectations (i.e., falling below the meets expectations standard of performance), meets expectations, or exceeds expectations.

## Purpose of the Foundation Skills Assessment

Although its content, administration cycle, and method of reporting differ from past provincial learning assessments, the primary purpose of the FSA is consistent with the purposes of past assessments: to help the province, school districts, and schools evaluate existing programs and make plans for improvement (Ministry of Education, 2000a, p. 1). The FSA will provide the province, school districts, and schools annual information about their students' reading comprehension, writing, and numeracy skills that can be monitored over time. This information should help inform program decisions

at the school, district, and provincial levels, and provide an ongoing means to evaluate the impact of such decisions or programs already in place.

The FSA's secondary purpose, which is in effect starting with the 2000 administration, differentiates it from all past provincial learning assessments administered in British Columbia. In addition to providing the provincial, district, and school level information outlined above, all annual FSAs from 2000 onward will give teachers, students, and parents/guardians an external source of information about individual students' performance in important foundation skill areas (Ministry of Education, 2000a, p. 1). This information, collected, scored, and reported in a standardized way across the province, will give teachers, students, and parents/guardians a unique opportunity to see how individual students are performing in relation to provincial standards (Ministry of Education, 2000a, p. 1).

## Comparing Performance on FSA Over Time: Assessment Benchmarks

To help monitor student performance at the provincial, district, and school levels over time, it is critical to establish a common reference point for reporting from one year to the next (i.e., assessment benchmarks). Unfortunately, differences between administrative and reporting procedures make the comparison of 1999 and 2000 FSA results problematic. In 1999, the FSA was administered under a significantly different context than in 2000 (Ministry of Education, 2000a, p. 6). With no school or individual level results being provided in 1999, schools and individual students could not be held accountable for their results on the tests. Consequently, the stakes attached to the 1999 tests were significantly less than for 2000 and beyond. Furthermore, unpredictable world

events occurring at the time of the 1999 assessment (e.g., multiple shootings in schools in Littleton, Colorado, and Taber, Alberta, by students who were considered to be "loners" by their classmates) put a focus on issues which were considered potentially sensitive to students and schools. Test booklets containing items that coincidentally touched on these issues (e.g., a story about someone famous who was considered to be a loner by his classmates) were consequently withdrawn from the 1999 assessment, resulting in the removal of a significant portion of reading items and all writing items at Grades 7 and 10. Finally, in response to the large number of students who were unable to finish some components of FSA 1999, students were given additional time to complete all components of FSA 2000 (Ministry of Education, 2000a, p. 6).

Since it is expected that FSA 2000 will share comparable administrative and reporting procedures with future administrations of the FSA, 2000 should, for all intents and purposes, be considered the inaugural year of the FSA. As such, FSA 2000 will play an important role in defining assessment benchmarks for reading comprehension, writing, and numeracy that can be applied to future FSAs (Ministry of Education, 2000a, p. 5). Finding a way to effectively apply common benchmarks across annual administrations of the FSA will be critical to satisfy the primary purpose of the program.

## The FSA Instrument

To provide a unique measure tailored to British Columbia's education system, the FSA instruments are developed by the BC Ministry of Education with significant involvement of many individuals from BC's educational community (Ministry of Education, 2000a, p. 2). The guiding specifications for each test, derived from the

prescribed learning outcomes of British Columbia's curriculum, are developed by Ministry of Education staff with input from subject-area and measurement specialists. Practising teachers from across the province are involved in writing, reviewing, pilot-testing, and revising test items and the final instruments. Content validity of the instruments is promoted through further review by subject-area and measurement specialists, ensuring the tests are composed of a suitable proportion of items and tasks of different types and that the appropriate content areas have been addressed. Representatives from BC's provincial educational organizations (i.e., teachers, administrators, parents, superintendents, and trustees) provide direction to the Ministry on how the FSA should be administered in schools, and how results should be reported and communicated.

The reading comprehension and numeracy test instruments of FSA 2000 consisted of 32 to 38 multiple-choice and 4 to 6 open-ended questions across the three grades (Ministry of Education, 2000a, pp. 43 & 69). The writing test instrument of FSA 2000 consisted of one extended writing task based on one of two topics (Ministry of Education, 2000a, p. 61).

In total, FSA 2000 was expected to take 4 hours to complete, and was expected to be administered over several sessions. The reading comprehension and numeracy tests were both split into two parts. Each part was designed to take most students 40-50 minutes to complete. The writing assessment was administered in one 60 minute session. To help ensure the tests were not a hardship for any one student, invigilators of the assessment were allowed to provide students (particularly those with special needs or on modified programs) additional time as required (Ministry of Education, 2000b, p. 1).

The multiple-choice items on the reading comprehension and numeracy components of FSA 2000 consisted of a prompt followed by four response options. The scoring of these items was dichotomous in nature, with responses being scored as either correct (assigned a value of 1) or incorrect (assigned a value of 0). The open-ended items on the reading comprehension and numeracy components required students to respond to prompts using written answers. These written answers were, for the most part, scored using a polytomous key, where one of a number of marks was assigned depending on the extent to which a student satisfied specific criteria (outlined in Ministry of Education [2000a, pp. 45-59 & p. 71]). For reading comprehension, open-ended questions were generally assigned maximum values of four marks (two open-ended questions per grade were assigned a maximum value of one mark); for numeracy, all open-ended questions were assigned a maximum value of four marks. Overall results on the reading comprehension and numeracy components of the FSA were based on the total marks awarded to each student on both multiple-choice and open-ended questions (Ministry of Education, 2000a, p. 7).

Responses to the extended writing task on the writing component of FSA 2000 typically involved more than one paragraph of text. As outlined in Ministry of Education (2000a, pp. 62-66; 2000c, p. 60), student work was scored using a ten-point scale (i.e., 0 to 9, with 9 signifying high proficiency) that represented writing skills typically demonstrated by students from Kindergarten through Grade 12. Grades 4, 7, and 10 students' responses on the writing component of the FSA were all scored using the same scale. Each paper was scored by two markers, resulting in a combined writing score that was placed on a nineteen-point scale (i.e., a half-point scale that ranged from a combined

score of 0-0, 0-1, 1-1, 1-2, 2-2, etc., to a combined score of 9-9). Overall results on the writing component of the FSA were based on the combined scores assigned to each paper by the two markers (Ministry of Education, 2000c, p. 60).

## The Traditional Approach to Analysing PLAP Assessment Data:

### Classical Test Theory

The 1999 FSA and all previous provincial assessments have relied on a traditional method of analysis based on classical test theory (CTT). Under the classical test model, it is hypothesized that an examinee's observed score ($X$) on a given test is composed of a true score ($T$) and a random error component ($E$), expressed in the form $X = T + E$ (Crocker & Algina, 1986, p. 107). True score, under this model, may be defined as the average observed score the examinee would hypothetically earn over many repeated administrations of a given instrument or strictly parallel instruments (Crocker & Algina, 1986, p. 111).

Within the classical framework, test scores on related components of a given PLAP assessment were given appropriate weights (as necessary) and summed to create an overall test score for each student. The overall test score for each student was then converted to an overall percent-correct score, which was used as the estimate of the student's true score on the assessment. Errors in this estimate were assumed to be the same for all students who wrote the assessment.

Test items on past provincial assessments were primarily characterized by two indices: item difficulty and item discrimination. Under the classical framework, item difficulty on provincial assessments was expressed as the proportion ($p$-value) of students who answered the item correctly (Crocker & Algina, 1986, p. 311). For example, if 65

percent of examinees answered a given item correctly, the difficulty (*p*-value) of that item was reported as 0.65; this item would be considered to be more difficult than an item with a *p*-value of 0.80 (i.e., where 80 percent of examinees answered the item correctly). As stated by Anderson (1997), item difficulty estimates are specific to a particular test and group of examinees; another administration of the test to a new group of examinees may result in different difficulty estimates.

Item discrimination, the extent to which an item discriminates between high and low ability students, was expressed as a correlation between student performance on the test item and the total score received on the test (Crocker & Algina, 1986, p. 317). This statistic is based on the premise that high ability students (i.e., those receiving high test scores) should tend to get the test item correct (i.e., high item score) and low ability students (i.e., those receiving low test scores) should tend to get the test item incorrect (i.e., low item score), resulting in a positive correlation (Anderson, 1997). Consistent with guidelines reported in Crocker and Algina (1986, p. 315), items with item-total score correlations of 0.30 or greater in previous administrations or field-tests of the provincial assessment were considered to have suitable levels of discrimination for inclusion in new assessments. As with item difficulty, however, item discrimination values are specific to a given test and group of examinees and may fluctuate from one test administration to another.

## Limitations of CTT

Although CTT has provided a useful way to analyse and report past provincial assessment data, a number of shortcomings have brought its continued use as the primary

model for analysing FSA data into question. Given the need for monitoring change in performance over time, of primary concern is the fact that CTT-based test and item indices for the FSA will change as the examinee context changes (making them group dependent), and that examinee ability estimates will change as the test context changes (making them test dependent) (Hambleton et al., 1991, p. 3). Due to this circular characteristic of CTT, "it is very difficult to compare examinees who take different tests and very difficult to compare items whose characteristics are obtained using different groups of examinees" (p. 3). Since different tests will be created for each administration of the FSA, and different groups of examinees will be tested from one year to the next (albeit in the same grades, i.e., 4, 7 and 10), comparability of test results from one administration of the FSA to another is problematic under a CTT framework.

Another limitation of using CTT in the analysis of FSA data arises with the release of individual student results in 2000 and beyond. Under CTT, it is generally assumed that the standard error of measurement is constant for all individuals who write a test (Embretson & Reise, 2000, pp. 15-18). As noted by R. Graves (personal communication, January 20, 2001), a method for creating CTT-based confidence intervals that are not constant across individuals is available (see Crawford and Howell [1998]), however any differences in the confidence intervals become insignificant with large sample sizes (i.e., greater than 100). This means that under a CTT-framework the same error of measurement would be applied to students who are close to the *meets expectations* cut-score on a given FSA test as would be applied to students who are close to the *exceeds expectations* cut-score. However, as Hambleton et al. (1991) point out, "scores on any test are unequally precise measures for examinees of different ability.

Hence the assumption of equal errors of measurement for all examinees is implausible." (p. 4). The consequence of this for students writing the FSA is that, if analysed using CTT, their reported results would not adequately take into account the actual error accompanying their observed score, possibly resulting in a misrepresentation of students' actual abilities (e.g., due to a less precise error of measurement that extends over two standard categories, students close to a cut-score may be reported as falling into a grey-zone of meeting/non-meeting expectations under CTT; with the more precise error of measurement available through IRT, these same students may be classified into one specific category rather than overlapping categories). Given this limitation of CTT, alternative methods for analysing and reporting individual student results should be explored.

## Item Response Theory: An Alternative Approach for Analysing FSA Data

An alternative approach for item and test analysis that has been used in many large-scale assessment programs to resolve practical measurement problems such as those faced with the FSA is item response theory (IRT) (see Keeves, 1994, pp. 89-93; Kolen & Brennan, 1995, p. 156; Hambleton & Swaminathan, 1985, p. 8). This modern measurement theory is "a general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test" (Hambleton & Jones, 1993, p. 40). IRT and its associated models rest on two basic postulates:

(a) The performance of an examinee on a test item can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; and

(b) the relationship between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an *item characteristic function* or *item characteristic curve* (ICC). This function specifies that as the level of the trait increases, the probability of a correct response to an item increases. (Hambleton et al., 1991, p. 7)

## Advantages of IRT

IRT has three primary advantages over traditional CTT models, as outlined by Hambleton (1989):

1. An examinee's estimated ability is not dependent on the particular sample of test items chosen from the calibrated pool of items. (However, the precision of the estimate is.)
2. The statistical descriptors of a test item (for example, item difficulty and discrimination indexes) are independent of the particular sample of examinees drawn from the population of examinees for whom the test is intended.
3. A statistic indicating the precision of the estimated ability is provided for each examinee. This statistic is dependent on examinee ability and the number and the statistical properties of the items and the test. (p. 160)

The advantages associated with IRT have significant relevance to the challenges faced with the FSA. Through the first and second advantages stated above, IRT enables

the comparison of examinees who have taken different tests, as well as the comparison of test results for different groups of examinees (Crocker & Algina, 1986, p. 339). These features make IRT a promising approach for enabling the performance of different students, and even the same student, on different FSA tests to be monitored over time. Furthermore, as per the third stated advantage, IRT will allow unique estimates of standard error to be calculated for each examinee, thereby providing an increased level of precision in estimating the abilities of individual examinees; if applied to the FSA, this feature would help ensure that individual student results are calculated in a more precise manner than is available under a classical framework.

## Disadvantage of IRT: Strong Assumptions

Although IRT holds promise for addressing the challenges faced by the FSA, its suitability to the analysis of FSA data is not assured. As outlined in Embretson and Reise (2000, p. 45), Hambleton et al. (1991, p. 9), and Kolen and Brennan (1995, p. 157), most commonly used item response models make a number of strong assumptions about the test data and underlying latent trait(s), including unidimensionality, local independence, and model-data fit. Furthermore, assumptions about non-speeded test administrations, equal discrimination indices, a lower asymptote of zero (i.e., minimal guessing), and the existence of item and ability invariance should also be considered in the selection and use of IRT analyses.

It is generally recognized that the strong assumptions underlying IRT applications are not likely to hold precisely in real testing situations (Kolen & Brennan, 1995, p. 156). Even if the assumptions of IRT are not completely met by a set of test data, however, the

benefits of IRT may accrue to a certain extent; as suggested by Embretson and Reise (2000, p. 226), the advantages available through the application of a given item response model will be directly proportional to the degree to which the test data are consistent with the assumptions underlying its application. Consequently, investigating the extent to which the test data of interest fit the assumptions and predictions of a given item response model is, as suggested by Kolen and Brennan (1995, p. 156), a crucial part of IRT applications.

## Unidimensionality

It is assumed in IRT that responses to the items on a test "can be accounted for by latent traits that are fewer in number than the test items" (Crocker & Algina, 1986, p. 339). Most IRT models (the most common of which will be outlined later in this paper) assume that examinee ability is described by a single latent trait. This implies that the construct being measured by the test is unidimensional. In practical terms, the unidimensionality assumption requires that the tests measure only one ability (Kolen & Brennan, 1995, p. 158).

As suggested by Hambleton and Swaminathan (1985, p. 17), the assumption of unidimensionality cannot be strictly met because a number of cognitive, personality, and test-taking factors (e.g., motivation, test anxiety, language, and ability to work quickly) always affect test performance to some extent. Consequently, Hambleton and Swaminathan (1985, p. 17) argue that unidimensionality is adequately met by the presence of a dominant component or factor that underlies test performance; this dominant factor is referred to as the ability measured by the test. The position taken by

Hambleton and Swaminathan (1985) is reflected in the less-restrictive notion of "essential unidimensionality," which Stout (1987; 1990) proposed as a sufficient indicator of unidimensionality. As described by Embretson and Reise (2000, p. 230), a data set will be essentially unidimensional when its dominant factor is so strong that examinee trait level estimates are robust to the presence of smaller factors and influences (i.e., the examinee trait levels obtained through an IRT model will be reliable estimates of the dominant factor).

## Local Independence

Related to unidimensionality is the assumption of local independence. As defined by Kolen and Brennan (1995), "local independence means that, after taking into account examinee ability, examinee responses to the items are statistically independent. [This assumption] implies that there are no dependencies among items other than those that are attributable to latent ability" (p. 158) (i.e., an examinee's responses to different items on a test are statistically independent except for the underlying trait in question). As stated by Hambleton and Swaminathan (1985, p. 23), for local independence to be true an examinee's performance on one item must not influence his or her performance on any other item in the test. For example, local independence would not hold in situations where an item provides clues to other items in the test or where a correct response on one item requires a correct response to a previous item. As outlined by Yen (1993), local dependence may also occur in instances where several items are attached to the same passage or setting (a common occurrence in many reading tests), where a subset of items

in a test measure unique content (e.g., clock reading items in a grade 3 mathematics concepts test), or where items share a common response format (e.g., written response).

## Model-Data Fit

A critical consideration in the appropriate use of IRT is the extent to which an IRT model adequately predicts or explains the test data set of interest (Harris, 1989; Lord, 1980, p. 15). Simply put, good model-data fit is achieved when examinee responses match the predicted responses of a given IRT model. While no item response model is likely to fit examinee responses perfectly (especially given sufficient amounts of data) (van der Linden & Hambleton, 1997, p. 16), the fit between the predicted responses of an item response model and actual examinee responses must be sufficiently close for the advantages of IRT to be realized. As argued by Rogers and Hattie (1987), "without acceptable fit, the validity of results cannot be ensured" (p. 47).

Given that different item response models will provide different predicted responses, Harris (1989) advises assessing the fit of various models and selecting the simplest model that accurately represents the observed data. If one is not satisfied with the model-data fit, Harris (1989) suggests a different model may be used or, if appropriate, examinees or items that clearly do not fit the model may be deleted to improve the fit (however, the latter procedure, as pointed out by Harris [1989] and argued by others [e.g., Traub, 1983; Rogers & Hattie, 1987], may be strongly questioned on pedagogical grounds).

As Hambleton and Swaminathan (1985, chap. 8) and Hambleton et al. (1991, chap. 4) outline, model-data fit should be assessed in a variety of ways. Statistical tests

such as the chi-square test of fit are available on most IRT computer programs. However, these statistical tests are usually sensitive to sample size and, given a large enough sample, will likely lead to a rejection of any null hypothesis that a given model fits the data (Hambleton & Swaminathan,1985, p. 153). Consequently, a thorough assessment of model-data fit should also include investigations into whether a data set satisfies the assumptions and expected properties of IRT.

## Nonspeededness

Related to unidimensionality is the assumption of nonspeededness. As stated by Hambleton and Swaminathan (1985, p. 30), all commonly used item response models assume that the test to which they are fitted are not administered under speeded conditions. In general, a test is speeded when a portion of the examinee population does not have sufficient opportunity to attempt all questions on the test within the allocated time (Oshima, 1994). Speed in answering questions introduces another ability that would underlie test scores, precluding the important assumption of unidimensionality. Nonspeededness means that examinees have sufficient time to attempt all items on the test and thus incorrect responses or omissions are due to limited ability rather than lack of the opportunity to attempt the questions.

## Equal Discrimination Indices

While not required by all item response models, equal item discrimination may be assumed in certain IRT applications. As suggested by Hambleton et al. (1991, p. 56), when the distribution of classical item-test score correlations (e.g., point-biserial

correlations) calculated for each item are reasonably homogeneous, the selection of an item response model that assumes equal item discrimination may be appropriate. Although no absolute criterion is available, a general guideline provided by Hambleton and Swaminathan (1985, p. 157) for checking the viability of this assumption is to look at the percentage of item-test score correlations falling outside some acceptable range (e.g., the average item-test score correlation ± 0.15).

Lower Asymptote of Zero (Minimal Guessing)

Another assumption required by some item response models is that examinees of very low ability have zero probability of answering an item correctly (Hambleton et al., 1991, p. 13). As will be demonstrated graphically in a later section of this paper, for these models the mathematical expression relating probability of providing a correct response to underlying ability will have a lower asymptote of zero. A practical interpretation of this assumption is that guessing is assumed to be minimal (Hambleton et al, 1991, p. 57), which implies the assumption is viable if the lower asymptote is *close to* zero.

For item response models that assume a lower asymptote of zero, the presence of guessing will negatively influence model-data fit, likely preventing the benefits of IRT to be realized. In general, guessing is most likely to be a factor in test performance on selected-response (e.g., multiple-choice or true-false) items or on tests that are speeded. However, as suggested by Hambleton et al. (1991, p. 16), the assumption of minimal guessing often can be met with selected-response items when the test is not too difficult for examinees (as might be expected on well-designed competency test administered after

effective instruction). In situations where the proportion of examinees with very low abilities is greater than zero (as would be expected on tests where guessing can occur), the application of an item response model that allows for a possible non-zero lower asymptote may be appropriate.

Item and Ability Parameter Invariance

If the relevant assumptions of an item response model are satisfied and reasonable model-data fit is achieved, a fundamental property of IRT may be achieved: invariance of item and ability parameters. A cornerstone of IRT, and its major distinction from CTT, this property implies that item parameter estimates do not depend on the sample of examinees to which a test is administered (item parameter invariance) and examinee parameter estimates do not depend on the set of test items (ability parameter invariance) (Hambleton et al., 1991, p. 18). Thus, under this property the item parameter estimates obtained from one group of examinees (e.g., students who wrote FSA 2000) should remain stable across other groups (e.g., students who write future FSA tests). Similarly, the ability parameter estimates obtained from one set of test items (e.g., items in FSA 2000) should also remain stable across different sets of test items (e.g., items in future FSA tests). As argued by Fan and Ping (1999), "the importance of the invariance property can never be overstated because without this important property the complexity of IRT models can hardly be justified on either theoretical or practical grounds" (p. 6).

As outlined in Lord (1980, p. 36), parameter invariance only holds as long as the origin and unit of the arbitrarily chosen ability scale are fixed. The common procedure of choosing a scale so that the mean and standard deviation of ability are 0 and 1 assists in

this regard. However, as Lord (1980, p. 36) describes, a fixed scale does not mean the parameter estimates from the two different groups of examinees or items will be identical but rather that they will have a linear relation to each other. In simple terms, item parameter estimates obtained from two groups of examinees or ability parameter estimates obtained from two sets of items will differ by only a constant.

Item and ability parameter invariance enables equating of different test forms, something that will be necessary to compare FSA results from one year to another. As defined by Kolen and Brennan (1995), "equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. Equating adjusts for differences in difficulty among forms that are built to be similar in difficulty and content" (p. 2). Kolen and Brennan (1995) outline a number of procedures to equate the results of different tests. Of relevance to FSA is the common-item nonequivalent groups design, a design in which two tests have a set of items in common and different groups of examinees are administered the two tests (Kolen & Brennan, 1995, p. 18). By virtue of parameter invariance, the estimates of item parameters on common items (commonly referred to as "anchor" items) can be used to estimate the linear transformation required to place parameter estimates from the two tests on the same scale. Readers interested in a detailed review of the technical and practical considerations of using IRT equate different tests are referred to Kolen and Brennan (1995, chap. 6 & 8).

## Popular Item Response Models

In this section, a number of item response models commonly used to analyse data obtained from achievement tests like the FSA are outlined. Using a framework proposed by Bejar (1983, p. 9), these models are classified by the type of data to which they may be applied and the number of parameters used to describe test items. The first three models described are appropriate for dichotomous response data (i.e., data that are scored into two categories to represent success or failure). The last two models described are appropriate for polytomous response data (i.e., data that are scored into three or more response categories). All models presented assume the test to be analysed is unidimensional, and therefore are classified as unidimensional item response models.

## Dichotomous Item Response Models

The three most commonly used dichotomous item response models are the one-, two-, and three-parameter logistic models, so named because of the number of parameters used to describe the test items and the logistic function used to relate examinee ability and the item parameter(s) to the probability of correctly responding to an item (Hambleton et al., 1991, p. 12; Harris, 1989). Descriptions for each model and their defining characteristics and assumptions are provided below.

One-parameter logistic model. The one-parameter logistic model (or Rasch model, as it is often called) is the simplest of all item response models. A distinguishing characteristic of the one-parameter logistic model (1PL) is that difficulty, or the *b*-parameter, is the only item parameter estimated. All items are assumed to be equally

discriminating in the 1PL model and guessing is assumed to be minimal (Hambleton & Swaminathan, 1985, p. 46).

The 1PL model may be mathematically specified as:

$$P_i(\theta) = \frac{e^{-1.7(\theta - b_i)}}{1+e^{-1.7(\theta - b_i)}} \qquad i = 1, 2, \ldots, n \qquad (1)$$

where $P_i(\theta)$ is the probability that a randomly chosen examinee with ability $\theta$ will answer item $i$ correctly, $e$ is the base of the natural logarithm system (whose value is 2.718), $b_i$ is the difficulty parameter for item $i$, $n$ is the number of items in the test, and 1.7 (often labelled D) is a scaling factor introduced to make the results comparable (±.01) to a cumulative normal distribution (Crocker & Algina, 1986, p. 353; Hambleton et al., 1991, pp. 12-13).

To illustrate the attributes of the 1PL model, it is useful to look at the item characteristic curves (ICCs) derived from the 1PL equation. Three 1PL ICCs with differing levels of difficulty ($b$= -1, 0, and 1) are presented in Figure 1. Several features of these ICCs are worth noting at this time. First, as described by Lord (1980, p. 12), the $b$ parameter represents the location of the curve's inflection point along the ability scale, with more difficult items having curves shifted further to the right. When guessing is assumed to not be a factor (as is the case with the 1PL model), $b$ is the ability level where the probability of a correct response is 0.5. Using the interpretations provided by Hambleton et al. (1991, p. 13), very easy items would have values of $b$ near -2.0 and very difficult items would have values of $b$ near +2.0.

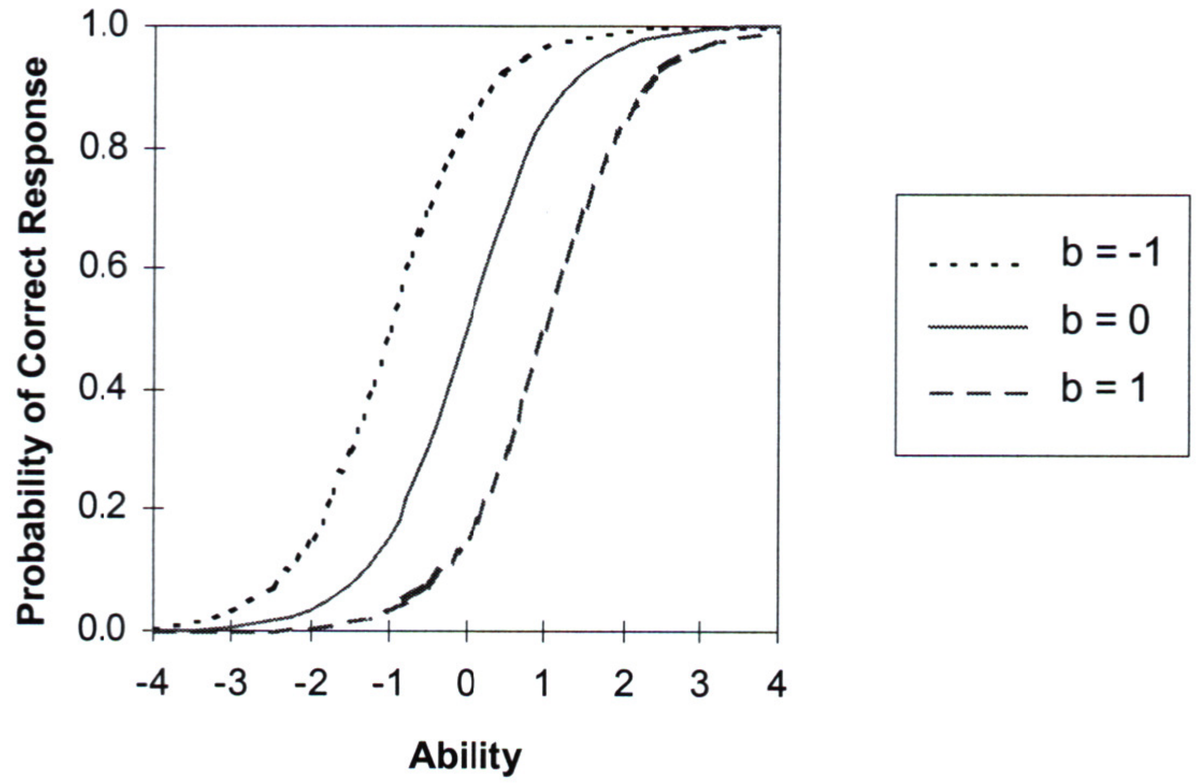



Figure 1. 1PL item characteristic curves at three levels of difficulty.

A second feature to note in Figure 1 is that the ICCs are parallel to one another, differing only in respect to their location on the ability scale (Wainer & Mislevy, 1990, p. 69). This is a defining characteristic of the 1PL model, which assumes that item difficulty is the only item characteristic that affects examinee performance (Hambleton et al., 1991, p. 13). Consistent with this assumption is a third feature worth noting in Figure 1: the lower asymptotes of each ICC shown are zero. As previously discussed, this specifies that examinees of very low ability have a zero probability of answering these items correctly (Hambleton et al., 1991, p. 13).

The simplicity of the 1PL is both a strength and weakness of the model. Given only one parameter is estimated in the 1PL, the model is relatively easy to work with (Hambleton & Swaminathan, 1985, p. 47) and does not require as large a sample size as other item response models for parameter estimation(Harris, 1989). Harris (1989) also suggests the 1PL model fits nicely with number-right scoring procedures since the total

test score is a sufficient statistic for estimating $\theta$ and the number of examinees who correctly answer an item is a sufficient statistic for estimating $b$. For these reasons, practitioners whose data are adequately modeled by the 1PL model are strongly encouraged to use it (Hulin, Drasgow, & Parsons, 1983, p. 39). Wright (1999) also strongly advocates the 1PL model, arguing that it provides a stable measure of the construct (i.e., the parallel curves ensure that the hierarchy of relative item difficulty will not change across ability levels, providing the stability expected in a sound measurement model).

The simple form of the 1PL model, however, can be its chief weakness when applied to real test data (Hulin et al. 1983, p. 39). As argued by Traub (1983) and Hulin et al. (1983, p. 39), test items do not necessarily all have equal discriminating power and guessing is likely to be a factor in many testing situations, particularly when multiple-choice items are involved. Consequently, the 1PL model will likely be appropriate only for carefully pre-tested sets of items that have been selected to meet the assumptions that underlie the model (Hulin et al., 1983, p. 39).

Two-parameter logistic model. The two-parameter logistic (2PL) model generalizes the 1PL model by allowing items to vary in their capacity to discriminate between high and low ability examinees, as well as in their difficulty (Nunnally & Bernstein, 1994, p. 404). Like the 1PL, the 2PL model assumes that examinees of low ability are almost certain to answer an item incorrectly and those high in ability are almost certain to answer the item correctly (Nunnally & Bernstein, 1994, p. 404). Guessing is assumed not to occur.

The form of the 2PL model may be expressed as:

$$P_i(\theta) = \frac{e^{-1.7\,a_i(\theta - b_i)}}{1 + e^{-1.7\,a_i(\theta - b_i)}} \qquad i = 1, 2, \ldots, n \qquad (2)$$

which is equivalent to Equation 1 except for one subtle but important difference: the addition of the item discrimination parameter, $a_i$ (Wainer & Mislevy, p. 69; Hambleton et al., 1991, p. 15).

Technically, $a_i$ is defined as the slope of the ICC at the point of inflection and can range in value from -∞ to +∞ (although on most achievement tests the range will usually be from 0 to 2) (Hambleton & Swaminathan, 1985, p. 36; Harris, 1989). Figure 2 illustrates the variations that occur across ICCs of items with different levels of discrimination.

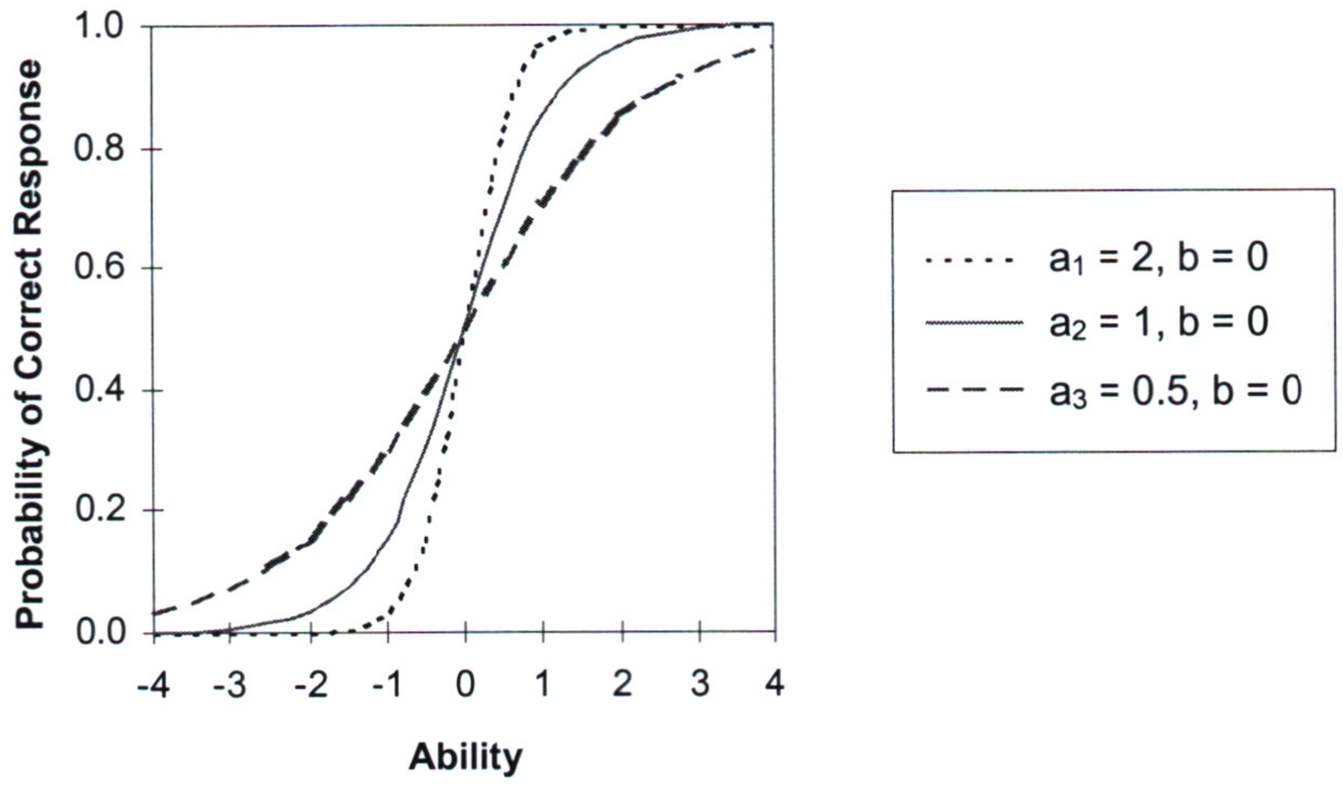



Figure 2. 2PL item characteristic curves with three levels of discrimination.

Figure 2 shows the ICCs of items that have high discrimination ($a = 2$), average discrimination ($a = 1$), and lower than average discrimination ($a = 0.5$) (Wainer & Mislevy, 1990, p. 69). Unlike the 1PL ICCs provided in Figure 1, 2PL ICCs are not parallel. Clearly, variations in the values of $a_i$ alter the shape of an ICC, with higher values of $a_i$ corresponding to steeper curves and increased power in discriminating between examinees of different levels of ability. The three ICCs in Figure 2 also show that the $a_i$ parameter controls the steepness of the slope around the $b_i$ parameter (held constant at $b_i = 0$ in this case). When the curve is steep (e.g., $a_1 = 2$), there is a substantial difference between the probabilities of a correct response for an examinee whose ability is slightly below the item's $b_i$ parameter and an examinee whose ability is slightly above the item's $b_i$ parameter; this difference in the probabilities of a correct response becomes less pronounced as the slope of the curve is reduced (e.g., $a_2 = 1$ or $a_3 = 0.5$).

It is also worth noting that, like the 1PL ICCs depicted in Figure 1, all of the 2PL ICCs in Figure 2 have a lower asymptote of zero. Additionally, as was the case with the 1PL model, item difficulty in the 2PL model (held constant at $b_i = 0$ in Figure 2) represents the ability level at which the probability of success is 0.5.

Given a wealth of empirical evidence that suggests items from achievement tests are likely to vary in the degree to which they correlate with the underlying trait (Traub, 1983), it can be expected the 2PL model is likely to be more appropriate for tests such as the FSA than the 1PL model. The 2PL model's assumption of a zero lower asymptote, however, can pose problems when modelling response data from tests where guessing may occur (e.g., multiple-choice tests).

Three-parameter logistic model. To account for the non-zero performance of low-ability examinees on test items, a third parameter, denoted $c_i$, is added to the item response model. This three-parameter logistic (3PL) model is defined mathematically as:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{-1.7\,a_i(\theta - b_i)}}{1 + e^{-1.7\,a_i(\theta - b_i)}} \qquad i = 1, 2, \ldots, n \qquad (3)$$

where $P_i(\theta)$, $b_i$, and $a_i$ are defined as before (Hambleton et al., 1991, p. 17). The parameter $c_i$ is typically referred to as the pseudo guessing (Crocker & Algina, 1986, p. 354) or pseudo-chance-level (Hambleton et al., 1991, p. 17) parameter. As described by Hambleton et al. (1991, p. 17), this parameter provides a (possibly) non-zero lower asymptote for the ICC and represents the probability of examinees with low ability getting an item correct. Theoretically, $c_i$ ranges from 0.0 to 1.0, but typically assumes values less than 0.3 (Harris, 1989).

The impact of the addition of the $c_i$ parameter on ICCs can be seen in Figure 3. Values of $c_i$ greater than zero (e.g., $c = 0.2$ and $c = 0.1$) result in a vertical shift of the probability functions, with higher values representing a higher probability that examinees with low abilities will answer an item correctly. When $c_i$ is set at zero (as is the case with the third ICC in Figure 3), the ICC takes a form characteristic of the 2PL model. It should also be noted that item difficulty takes on a different meaning in the 3PL model. Although item difficulty still occurs at the inflection point of the ICC, it is no longer the ability level where the probability of a correct response is 0.5. When there is a non-zero

lower asymptote (as illustrated in Figure 3 by the ICCs with $c_i$ parameter values of 0.2 and 0.1), item difficulty is the ability level where the probability of a correct response is halfway between $c_i$ and 1.0 (Lord, 1980, p. 13).

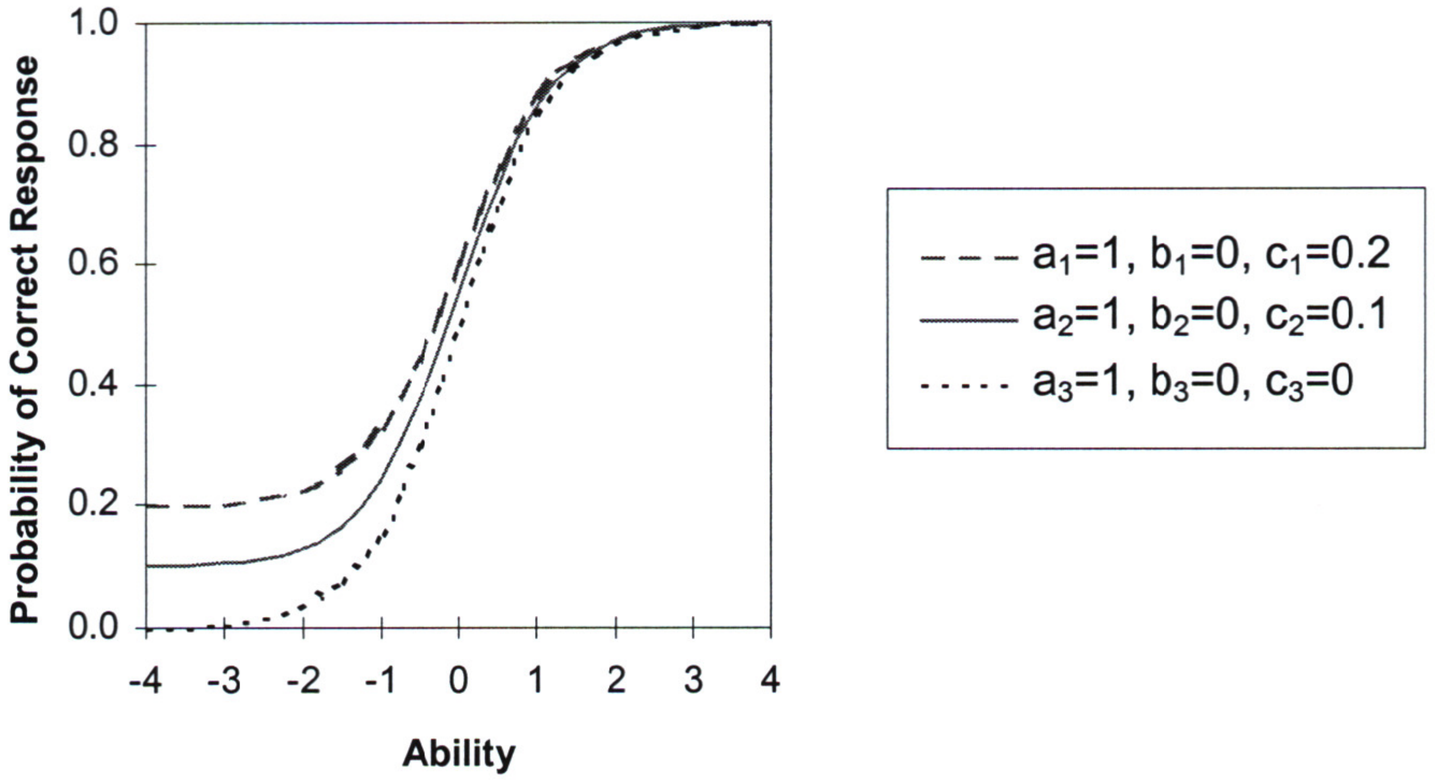



Figure 3. 3PL item characteristic curves with three different pseudo guessing parameters.

Given the 3PL model's accommodation of a possible non-zero lower asymptote and unequal item discrimination, it may appear that the 3PL model is the dichotomous IRT model of choice for analysing multiple-choice achievement data. However, making this assumption without first investigating the extent to which various item response models fit the data would be a significant mistake. As suggested by Harris (1989), possible difficulties in estimating the $c_i$ parameter in real data sets may make the 2PL model more suitable for practical applications. Also, in addition to requiring the estimation of parameters that really do not need to be estimated, use of an unnecessarily complex model may result in less accurate estimates than if a simple but adequate model was used (Crocker & Algina, 1986, p. 361). Consequently, in line with the

recommendations of Harris (1989) and Crocker and Algina (1986, p. 362), conscientious users of IRT should evaluate the fit of various models to the test data in a variety of ways and select the model with the least number of parameters that accurately represents the observed data.

## Polytomous Item Response Models

While the 1PL, 2PL, and 3PL models are useful for analysing dichomotously-scored items, they are not suitable for analysing items with multiple response categories. Given FSA 2000 has open-ended items where a range of scores are possible, alternative models will need to be explored to enable the application of IRT to all item formats used in the FSA. Two popular unidimensional item response models that have been successfully applied to polytomously-scored achievement items like those found on FSA 2000 are described in this section. They are the Partial Credit Model (Masters, 1982; Masters & Wright, 1997) and the Generalized Partial Credit Model (Muraki, 1992; 1993; 1997).

Partial credit model (PCM). The PCM was originally developed for analysing test items that require multiple steps and for which partial credit was assigned for completing a number of steps in the solution process (Embretson & Reise, 2000, p. 105). The PCM shares the same algebraic form as the 1PL model, generalized to model polytomous data. Like the 1PL model, the PCM estimates only one item parameter, the difficulty parameter ($b$). In the case of the PCM, $b$ is conceptualized in terms of the difficulty of each successive step in the polytomous item (i.e., going from one possible score to another)

(Masters, 1982). Through the PCM, the difficulty of each step may be calculated and plotted.

The linkage between the 1PL and the PCM becomes clear when the models are viewed as the conditional probabilities of an examinee responding in one possible category (e.g., $k$) rather than an adjacent category (e.g., $k$-1) given $m$+1 possible categories of a particular item (Muraki & Bock, 1997, p. 14).

In the dichotomous 1PL model, where $m$+1 = 2, the following holds (Muraki & Bock, 1997, p. 14):

$$P_{i0}(\theta) + P_{i1}(\theta) = 1. \tag{4}$$

Given the above, the probability of scoring 1 rather than 0 on a dichotomous item is expected to increase with ability (Masters & Wright, 1997, p. 102). In the 1PL, this expectation is modeled as

$$\frac{P_{ij1}}{P_{ij0}+P_{ij1}} = \frac{\exp(\theta_j - b_i)}{1+\exp(\theta_j - b_i)} \tag{5}$$

where, as defined by Masters and Wright (1997),

> $P_{ij1}$ is the probability of person $j$ scoring 1 on item $i$, $P_{ij0}$ is the probability of person $j$ scoring 0, $\theta_j$ is the ability of person $j$, and $[b_i]$ is the difficulty of item $i$ defined as the location on the measurement variable at which a score of 1 on item $i$ is as likely as a score of 0. The model is written here as a conditional probability

to emphasize that it is a model for the probability of person *j* scoring 1 *rather than* 0. (p. 102)

(The term exp is equivalent to the constant *e* used in previous equations. With the inclusion of a scaling constant of -1.7, Equation 5 will mirror the 1PL form in expressed in Equation 1.)

When an item has more than two response categories (as is the case with the open-ended reading comprehension and numeracy items on FSA 2000 with five ordered categories scored 0, 1, 2, 3, and 4), Equation 4 may be recurrently applied to each pair of adjacent binary categories (equal to *k*) and expressed as follows (Muraki & Bock, 1997, pp. 14-15 & 18):

$$P_{i0}(\theta) + P_{i1}(\theta) + \cdots + P_{im_i}(\theta) = \sum_{k=0}^{m_i} P_{ik}(\theta) = 1 . \qquad (6)$$

As described by Masters and Wright (1997, p. 103), this process is conventionally rewritten as the unconditional probability of each possible outcome 0, 1, …, $m_i$ of person *j*'s attempt at item *i*, assuming that person *j* is assigned one of the $m_i + 1$ available scores for that item. This general expression for the partial credit model (adapted from Masters and Wright [1997, p. 103], with a scaling constant of -1.7 incorporated) may be written as

$$P_{ijx}(\theta) = \frac{\exp \sum_{k=0}^{x} (-1.7)(\theta_j - b_{ik})}{\sum_{h=0}^{m_j} \exp \sum_{k=0}^{h} (-1.7)(\theta_j - b_{ik})} \qquad x = 0, 1, \ldots, m_i \qquad (7)$$

where $b_{i0} \equiv 0$, $P_{ijx}(\theta)$ is the probability of examinee $j$ with ability $\theta$ scoring $x$ on item $i$, $b_{ik}$ is the difficulty of the $k$'th step in item $i$ (i.e., the $k$'th transition from one category to another), and $h$ is each possible score outcome 0, 1, …, $m_i$ for item $i$ (Masters, 1982). It is worthy to note that this model reduces to the dichotomous 1PL when $m_i = 1$ and $k = 0, 1$ (Muraki & Bock, 1997).

For each item $i$ scored 0, 1,…, $m_i$, the PCM estimates a set of $m_i$ item parameters ($b_{i1}, b_{i2,}…, b_{im}$) (Masters & Wright, 1997, p. 103). These estimates are substituted into Equation 7 and the estimated probabilities of scoring 0, 1, …, $m_i$ on item $i$ are calculated for any specified ability $\theta$ (Masters & Wright, 1997, p. 103). An example outcome of this process is displayed in Figure 4, where category characteristic curves are plotted for an open-ended item with $m_i = 2$.

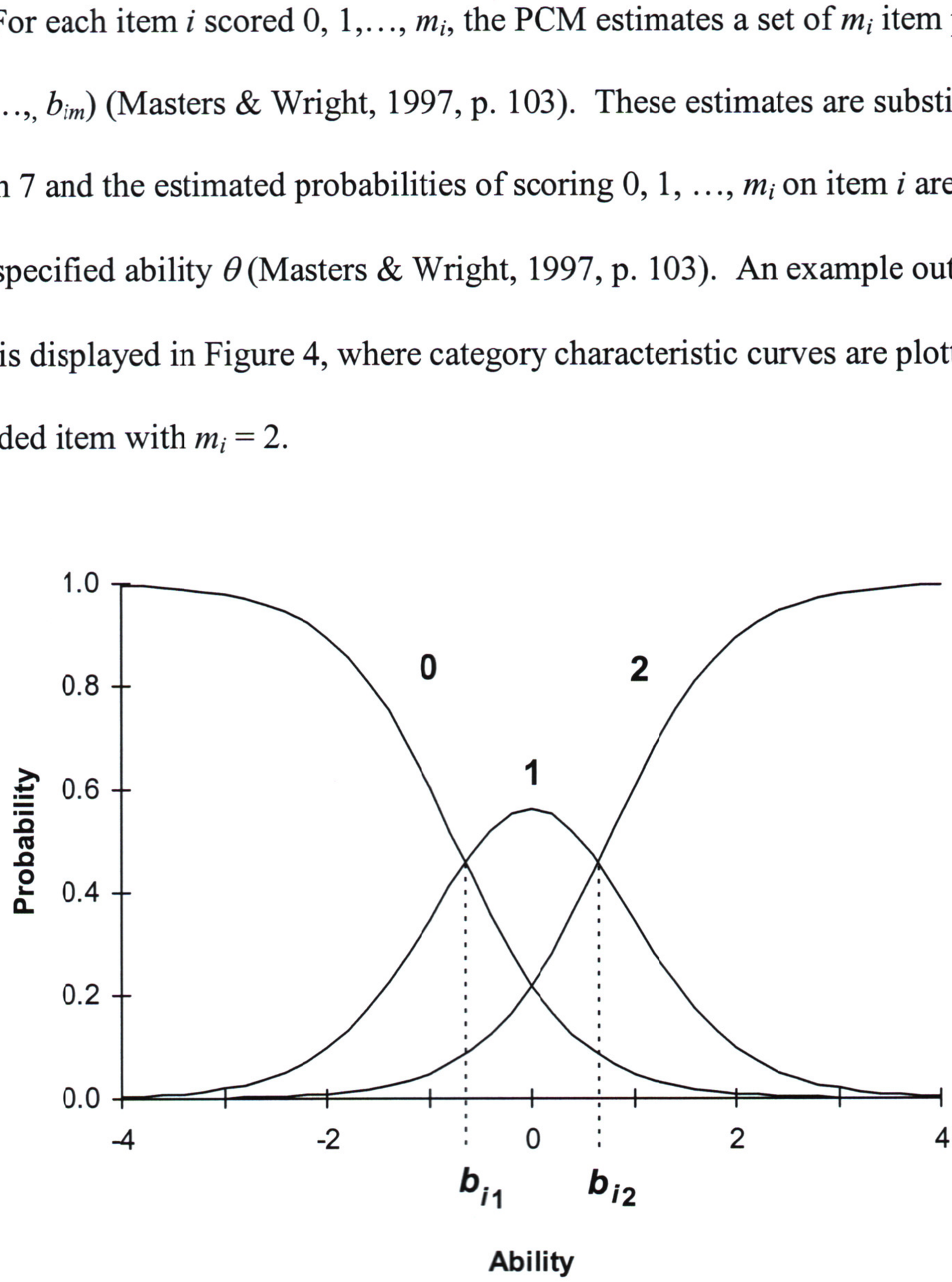



Figure 4. Category characteristic curves for a polytomous item ($a_i = 1.0$, $b_{i1} = -0.75$, $b_{i2} = 0.75$).

A number of important features of the PCM can be noted in Figure 4. Immediately, one can see that the PCM defines $m_i + 1$ (3 in the case of the item presented, where $m_i = 2$) category characteristic curves for a polytomous item $i$. The curves labelled 0, 1, and 2 represent the probability of receiving scores of 0, 1, and 2 (respectively) out of 2 at a given ability level. One should also note that for examinees of increasing ability, the probability of obtaining a score in the lowest category (0) decreases as the probability of obtaining a higher score (1 or 2) increases. The probability of obtaining a score in the middle category (1) increases with ability up to a point and then decreases as the probability of obtaining a higher score (2) increases. The relative difficulties of the $m_i$ steps correspond to the points on the ability scale at which two adjacent categories are equally likely (i.e., at $b_{i1}$ and $b_{i2}$). In graphical terms, the $b_i$ parameters of the PCM are the points on the ability scale at which the characteristic curves for two adjacent categories intersect, representing the step "difficulties" associated with moving from one category to the next (Embretson & Reise, 2000, p. 106).

In addition to the obvious advantage of being able to analyse polytomous response data, another advantage of the PCM (one it shares with the 1PL) is that the raw scale score derived from the model "is a sufficient statistic for estimating examinee trait level; examinees with the same raw score on a set of items that fit the PCM are estimated to have equivalent positions on the latent trait" (Embretson & Reise, 2000, p. 107). However, as was raised earlier with the 1PL, this property has the drawback that all items in a test must be equally related to the underlying ability (i.e., have equal powers of discrimination), a property that Embretson and Reise (2000, p. 107) state may be difficult to produce in practice.

Generalized partial credit model (GPCM). Another unidimensional item response model that can be used to analyse polytomous items is Muraki's (1992) GPCM. The GPCM extends the PCM by allowing the analysis of items that are unequally discriminating. This is accomplished through the addition of a discriminating or slope parameter ($a_i$) to the PCM equation. Given the estimation of a second parameter, the GPCM is considered to be a two-parameter model for polytomous items. If used on dichotomously-scored items, the GPCM reduces to the 2PL model described earlier.

Extending the PCM equation (Equation 7) adapted from Masters and Wright (1997, p. 103), the GPCM may be mathematically defined as

$$P_{ijx}(\theta) = \frac{\exp \sum_{k=0}^{x} (-1.7) a_i (\theta_j - b_{ik})}{\sum_{h=0}^{m_j} \exp \sum_{k=0}^{h} (-1.7) a_i (\theta_j - b_{ik})} \qquad x = 0, 1, \ldots, m_i \qquad (8)$$

where $b_{i0} \equiv 0$, $P_{ijx}(\theta)$ is the probability of examinee $j$ with ability $\theta$ scoring $x$ on item $i$, $a_i$ is the discrimination parameter for item $i$, $b_{ik}$ is the difficulty of the $k$'th step in item $i$, and $h$ is each possible score outcome 0, 1, …, $m_i$ for item $i$.

The two panels in Figure 5 illustrate the impact the $a$-parameter has on the category characteristic curves of polytomous items. Figure 5(a) shows a four-category polytomous item with $a_i = 1.0$, $b_{i1} = -0.5$, $b_{i2} = 0.0$, $b_{i3} = 2.0$ (since $a_i = 1.0$, this figure would reflect curves obtained through the PCM) . If the discrimination parameter is reduced to 0.7 and $b$-parameter values are held constant at $b_i = (-.5, 0.0, 2.0)$, as illustrated in Figure 5(b), the category characteristic curves become flatter, with no

change in the location of the intersection points of the curves (Muraki & Bock, 1997, p. 17).

By providing more information about the characteristics of test items than the PCM (Muraki, 1997, p. 153) and the possibility of improvements in model-data fit (e.g., Embretson & Reise, 2000, p. 115), the GPCM is a viable alternative to the PCM for achievement tests containing polytomously-scored items. The inclusion of the *a*-parameter, however, is not without cost. As suggested by Embretson and Reise (2000, p. 124), estimation of an additional parameter in the GPCM requires a larger sample size and makes the raw scale score an insufficient statistic for examinee trait level (complicating the reporting of results). Overall, as with all IRT models, the extent to which the GPCM accurately represents actual test data will be an critical consideration in its possible application to the FSA.

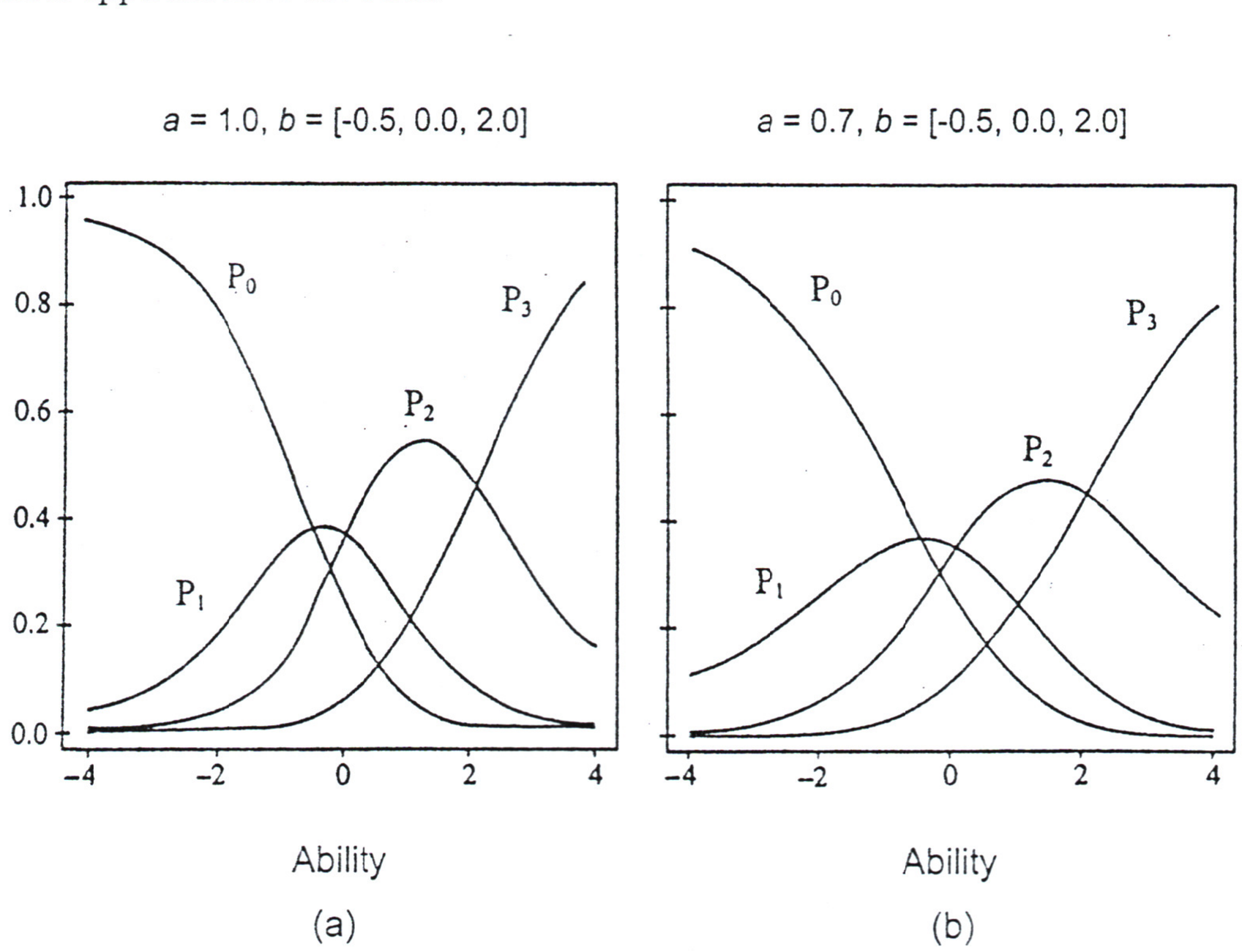



Figure 5. Two four-category polytomous items with $a = (1.0, 0.7)$ (Muraki, 1992).

No 3-parameter model for polytomous data. While the links between dichotomous and polytomous models are clear for one- and two-parameters, it is worthy to note a comparable polytomous link has not been established with 3PL dichotomous model. For test items that require students to formulate a response (a common characteristic of polytomous items) rather than simply select from a number of possible responses (as is the case with multiple-choice questions), the probability of low ability students getting an item correct is very low. Consequently, there is little need for a polytomous IRT model that allows for a non-zero lower asymptote.

## Summary

In 2000, British Columbia's Ministry of Education fully implemented its Foundation Skills Assessment (FSA), an annual assessment designed to assess the reading comprehension, writing, and numeracy skills of grades 4, 7, and 10 students throughout the province. A review of the literature showed a number of significant differences between the FSA and past provincial assessments, including content (the cross-curricular skills of reading comprehension, writing, and numeracy versus the curriculum-specific assessments administered in the past), administrative cycle (annual versus every four or five years), and level and method of reporting (reporting provincial, district, school, and, for the first time, individual student results in relation to provincial standards of performance). Two significant challenges arise from the unique features of the FSA: the need to ensure the comparability of FSA test results from one year to the next and, with the introduction of student-level results, the duty to ensure individual results are calculated in the most accurate manner possible. To help address these

challenges, an alternative to the traditional approach of analysing provincial assessments in British Columbia was proposed for the FSA: item response theory (IRT).

A review of the literature identified ways IRT can help solve measurement problems like those faced by the FSA. Of particular interest is the ability of IRT to place different tests and groups of examinees on a common scale of measurement and to obtain a more precise measure of individual student ability than is possible through traditional test analyses. While IRT appears to be very promising, its proper application is neither simple or guaranteed. A review of the literature identified a number of significant assumptions and conditions that must be considered and met before the benefits of IRT can accrue. Issues such as unidimensionality, local independence, model-data fit, nonspeededness, the distribution of item-test score correlations (i.e., equal or unequal discrimination indices), a lower asymptote of zero, and item and ability invariance were discussed, as were a number of popular item response models.

To help ensure decisions regarding the application of IRT to the FSA are based on sound measurement principles, two purposes for the study were identified: first, to determine the extent to which data from the Grades 7 FSA 2000 tests meet the assumptions and conditions of IRT; and second, to identify which of several popular item response models best fit Grade 7 FSA 2000 data.

# CHAPTER 3

# METHOD

To facilitate a thorough investigation of the feasibility of using IRT with FSA 2000 data, this study utilizes a framework outlined in Hambleton et al. (1991, chap. 4) and Hambleton and Swaminathan (1985, chap. 8) to check the principal assumptions underlying popular IRT models (i.e., unidimensionality, local independence, nonspeededness, equal item discrimination, lower asymptote, item and ability parameter invariance, and model-data fit). Procedures taken from these two respected resources (Nunnally & Bernstein, 1994) were supplemented by methods discussed by Embretson and Reise (2000) and related investigations by Klinger (2000) and Boughton (1998) on the application of IRT to British Columbia Grade 12 provincial examination data sets.

## Data Set

This study made use of response data from FSA tests administered to Grade 7 students throughout British Columbia in May 2000. The responses of 10,000 students were included in the study, drawn from the responses of 94% of British Columbia's Grade 7 student population (approximately 47,000 out of 50,000 students enrolled across the province).

Response data from the reading comprehension and numeracy components of the Grade 7 FSA 2000 were used. Due to an insufficient number of test items to support IRT analyses (only one item), the remaining component of FSA 2000 (writing) was not included in the analyses.

The reading comprehension data set consisted of 38 multiple-choice items (worth one mark each) and 6 open-ended items (two of which were worth one mark and four of which were worth four marks). Given their dichotomous nature, the two open-ended reading comprehension items worth one mark were combined with the dichotomously-scored multiple-choice data set. The numeracy data set consisted of 34 multiple-choice items (worth one mark each) and 4 open-ended items (worth four marks each).

The population data sets were provided to the author of this study by the Data Management and Student Certification Branch of British Columbia's Ministry of Education. Protection of privacy was assured by the removal of student names and school identifiers from the data sets prior to receipt. Prior to analyses, the population data sets were cleaned to remove students for whom overall results were not generated on the actual assessment (e.g., students who were missing one of the components [multiple-choice or open-ended] in the skill area tested, those who did not complete at least 50% of the items in both components, or those who did not score at least 15% of the total score on the given test). These data cleaning criteria were developed and utilized by the ministry to ensure that any reported result from the assessment was based on a reasonable amount of information and would not be hampered by missing booklets or spurious student responses.

To facilitate IRT analyses, students with perfect scores or zero scores on either component were also removed from the population data set (affecting 1,101 students in the reading comprehension data set and 1,714 students in numeracy). A preliminary review of the open-ended data showed sufficient numbers of responses to enable IRT

analyses across available categories, removing the potential need to collapse categories with few responses.

The samples used in this study were drawn by assigning random numbers to each student in the cleaned population file using EXCEL 2000 (Microsoft, 1999). The students were sorted by this random number and the first 10,000 students from the reading comprehension and numeracy data sets were selected to form the final samples to be analysed in this study. With a sample of 10,000 students, it was hoped that the study would reasonably reflect the large-scale realities of FSA without the added time and computer resource burdens that would be associated with analysing the complete data set.

## Testing the Assumptions of IRT

In order to determine the extent to which IRT can be appropriately applied to FSA 2000 data, the following assumptions and expected properties of item response models were investigated: unidimensionality, local independence, nonspeededness, equal item discrimination indices, lower asymptote, and item and ability parameter invariance. In effort to determine which of several item response models (i.e., 1PL, 2PL, 3PL, PCM, GPCM) best fit the FSA 2000 data, these investigations were complemented by further analyses of model-data fit. The specific procedures used in each investigation are outlined in this section.

### Unidimensionality

A common assumption of the popular item response models investigated in this study is unidimensionality, defined as "the existence of one latent trait underlying the

data" (Hattie, 1985, p. 157). Despite the importance of this assumption to these models, relevant reviews of literature pertaining to unidimensionality (e.g., Hambleton & Rovineli, 1986; Hattie, 1984, 1985; Nandakumar, 1994) indicate there is no one agreed upon and absolute way to assess for its presence or absence in a set of test data. A further hindrance to the evaluation of this assumption is the general recognition that data from educational assessments will likely never be strictly unidimensional (Embretson & Reise, 2000, p. 230; Hambleton & Swaminathan, 1985, p. 17; Traub, 1983, p. 59). Consequently, the present study tested the assumption of unidimensionality in a variety of ways using the less restrictive notion of essential unidimensionality (which is satisfied with the presence of one *dominant* component that underlies test performance).

Two traditional procedures to test for unidimensionality through principal components factor analysis were first employed. These included reviewing the percentage of total variance explained by the first principal component and the relative position of the various components' eigenvalues on a scree plot to determine the presence or absence of a dominant component.

All factor analyses were conducted on combined multiple-choice and open-ended data sets for reading comprehension and numeracy using SPSS 10.0 (SPSS Inc., 1999). Using a criterion provided by Reckase (1979), essential unidimensionality was assumed if the first component accounted for at least 20% of the test variance. Reckase's (1979) findings showed that acceptable IRT parameter estimates may be obtained for the first principal component when this criterion holds true. To help further illustrate the possible presence of a dominant component, the eigenvalues for each component of the data sets were plotted. While the values obtained through this procedure are open to interpretation,

clear dominance of the first component was assumed if the largest eigenvalue in a data set was approximately five times larger than the second largest eigenvalue and the second largest and remaining eigenvalues had similar values (Hambleton et al, 1991, p. 62).

To obtain further evidence to assist in any judgement regarding the unidimensionality of the data sets, the traditional heuristics available through factor analysis were augmented by a more formal and sophisticated procedure using the computer program Poly-DIMTEST (William Stout Institute for Measurement, 1999).

Poly-DIMTEST employs a statistical procedure designed to assess essential unidimensionality of either dichotomous or polytomous data. The procedure involves several steps, as outlined in Nandakumar, Yu, Li, and Stout (1998, pp. 101-103). First, an *N*-item test is split into three subtests labelled AT1, AT2, and PT. A relatively small set of *M* items (not less than 4 and no more than 25% of *N*), is selected for subtest AT1 in such a way that the items appear to measure the same dominant ability and, at the same time, seem dimensionally distinct from the rest of the test. This can be achieved either through expert opinion (where experts familiar with the content of the test select relatively homogenous items) or exploratory factor analysis (where items with highest loadings on the second factor are selected). A second set of *M* items is then selected by Poly-DIMTEST for the AT2 subtest so that the items in AT2 have a similar item difficulty distribution as the items in AT1. The remaining items ($n = N - 2M$) are placed into PT, the partitioning subtest, and are used to partition examinees into subgroups based on their PT scores. From these subtests, a statistical index *T* used to assess departure from unidimensionality can be calculated (the interested reader is referred to Nandakumar et al. [1998, pp. 101-103] for a detailed discussion on how the *T* statistic is derived). In

general, the null hypothesis of unidimensionality is rejected "when $T \geq Z_{\alpha}$, where $Z_{\alpha}$ is the upper 100(1 - $\alpha$) percentile of the standard normal distribution, and $\alpha$ is the desired level of significance" (Nandakumar et al., 1998, p. 103).

The procedure used in Poly-DIMTEST is similar to the one used in a version designed exclusively for dichotomous items (DIMTEST). Although recommended as a useful method for determining essential unidimensionality (Embretson & Reise, 2000, p. 245; Hattie, Krakowski, Rogers, & Swaminathan, 1996), the procedure used in DIMTEST (and, likewise, Poly-DIMTEST) is not without faults. One criticism that has been put forward is that different conclusions may be derived depending on which items are selected to comprise AT1 (see Linacre, 1994; Wang, 1994). The need to validate Poly-DIMTEST on real data sets has also been recognized (Li & Stout, 1995; Nandakumar et al., 1998). In an effort to obtain information regarding the dimensionality of FSA 2000 data in a manner that would address possible limitations of the software, different AT1 samples deemed to be homogeneous were used in the calculation of the *T* statistic. The AT1 samples used were selected through expert opinion (obtained through discussions with the test developers at the Ministry of Education) or factor analysis procedures (selecting items with the highest loadings on the second factor).

## Local Independence

As has been previously discussed, "the local independence assumption implies that there are no dependencies among items other than those that are attributable to latent ability" (Kolen & Brennan, 1995, p. 158). The absence of local independence can be described using factor analysis terminology: "if item scores are locally dependent,

then…they will have nonzero residual correlations after removal of the first factor" (Yen, 1993, p. 188). Consistent with this description, local independence in the data sets was assessed by reviewing the residual correlation matrix with the first factor removed produced through a principal axis factor analysis in SPSS 10.0 (SPSS Inc., 1999). This matrix represents the difference between the observed and reproduced correlational matrices, with the residuals representing partial correlations between pairs of variables with the effect of the first factor removed (Tabachnick and Fidell, 1996, p. 673). Using general guidelines derived from Tabachnick and Fidell (1996, p. 673) and McDonald (1999, p. 172), residuals of 0.10 or higher for a pair of items were considered to indicate the presence of another factor. These large residuals were identified on both residual correlation matrices. To help in the identification of potential problem items, moderate residuals with absolute values between 0.05 and 0.1 were also noted and the number of moderate and large residuals obtained for each item were summed. Consistent with the recommendation of McDonald (1999, p. 172), item pairs with large residuals after the first factor was removed were reviewed in the context of the actual test questions to see if any pattern emerged. For example, two items that share a large residual might also share obvious links in the specific type of content they assess, the passage on which they are based, their location in the test, or other characteristics.

## Nonspeededness

As was discussed earlier, appropriate use of IRT assumes that students have enough time to complete all items on the test. When a test is speeded, the assumption of

unidimensionality cannot be met because test performance would then be influenced by a second factor (speed in completing test items) rather than strictly ability.

Nonspeededness was measured by examining the percentage of examinees who did not complete items on each test. Adapting a procedure used by Klinger (2000), nonspeededness was assumed if 90% of the students completed the last five items on Parts 1 and 2 of the reading comprehension and numeracy tests. If less than 90% of the students completed these items, the omission rates of other items in that test were examined to determine if non-response was due to time constraints or item difficulty. If comparable omission rates were found for other items in the test, the omission rates of the last five items were assumed to be due to item difficulty rather than lack of time (Klinger, 2000). If comparable omission rates were not found elsewhere on the test, speed could be assumed to be a factor in test performance.

## Equal Item Discrimination

To help inform which item response models could be used to analyse FSA 2000 test data, the equal item discrimination assumption of the 1PL and PCM was tested. This was achieved by calculating traditional item discrimination indices (point-biserial correlations for multiple-choice items and the point-polyserial correlation for open-ended items) in PARSCALE 3.5 (Muraki & Bock, in press) and reviewing their distributions. As suggested by Hambleton et al. (1991, p. 56), if the distribution of item discriminations indices are relatively homogeneous, it may be viable to use an item response model that assumes equal item discrimination. A general heuristic proposed by Hambleton & Swaminathan (1985, p. 157) and employed by Choi and Bachman (1992) was used to

determine if the equal discrimination assumption was tenable. This included identifying the percentage of item-test score correlations that fall outside some acceptable range, specifically the average item-test score correlation ± 0.15. While no absolute criterion was reported to evaluate the percentage of item-test correlation falling outside this range, Choi and Bachman (1992) considered 17% to be a clear violation of the assumption of equal item discrimination.

Judgments regarding the equal item discrimination assumption were also supported by later analyses evaluating the relative fit of the various IRT models. Since a one parameter IRT model differs from its two parameter counterpart by only the latter's accommodation for possible unequal item discrimination, any increase in fit going from a one-parameter model to a two-parameter model was assumed to provide evidence against equal item discrimination in the data sets.

## Lower Asymptote

To further inform the selection of appropriate item response models, an investigation into the presence of non-zero lower asymptote was carried out. Given a non-zero lower asymptote was not expected to be an issue for polytomous item response models, only multiple-choice items were included in this investigation. Combining procedures outlined in Hambleton et al. (1991) and Klinger (2000), the lower asymptote of each data set will be assessed by checking the item level performance of examinees in the lowest decile on the five most difficult multiple-choice items as defined by the five lowest *p*-values. If the performance levels of these examinees on these items were close to zero, guessing was assumed not to be a factor in student performance. However, if the

performance levels of low ability students on these items were close to 25% (i.e., the probability of getting a multiple-choice item with four alternatives correct simply by chance), a non-zero lower asymptote was assumed to exist.

As pointed out by Hambleton and Swaminathan (1985, p. 161), this procedure is not without faults. The results could be misleading if students in the lowest decile are only relatively low in ability in relation to other examinees but not low ability in any absolute sense. Consequently, the results obtained through this procedure were complemented with later analyses that compared the fit between the 2PL model (where the lower asymptote is fixed at zero) and the 3PL model (which allows for a non-zero lower asymptote).

## Item and Ability Parameter Invariance

Item and ability parameter invariance for each item response model was assessed using procedures outlined in Hambleton et al. (1991, pp. 62-65).

Item parameter invariance in the reading comprehension and numeracy data sets were assessed separately using three general steps. First, two randomly equivalent groups of 3,000 students were drawn from the each sample of 10,000 students using a random sampling procedure in EXCEL. Item *b*-parameter estimates were calculated for the two groups using the default expected a posteriori estimation procedure in PARSCALE for five popular IRT models (1PL, 2PL, 3PL, PCM, GPCM). These results were correlated and *b*-parameter estimates for all items were placed on a scatterplot. Second, two distinct ability groups were drawn from the original samples of 10,000 students (3,000 students from the top half of the total score distribution and 3,000 students from the bottom half of

the total score distribution). Item *b*-parameter estimates were then calculated for these high and low performing groups and, as before, the results were correlated and placed on a scatterplot. Finally, the results of the first and second steps were compared. Item parameter invariance was demonstrated for the relevant IRT model if the scatterplots for the randomly equivalent groups and high and low performing groups were similar and had a linear relation to each other (Hambleton et al., 1991, pp. 63-65; Lord, 1980, p. 36).

To assess invariance of ability parameters across different samples of items and IRT models, a comparable procedure outlined by Hambleton et al. (1991, pp. 63-65) was followed. First, each test was divided into two randomly equivalent forms (consisting of odd- or even-numbered items). The ability estimates of 3,000 randomly chosen students on these equivalent forms were calculated for each IRT model using PARSCALE, correlated, and displayed in a scatterplot to demonstrate any variability due to sampling of test items. Next, each test was divided into two forms consisting of the easiest and hardest items from the test. Ability estimates were calculated using these easy and hard subtests, and the results were correlated and plotted. Finally, as before, the two sets of reading comprehension and numeracy results were compared. If a similar linear relation was observed, ability parameter invariance over tests of varying difficulty was established for a given IRT model.

To further assist the interpretation of ability parameter invariance under various IRT models, classical reliability indices were also calculated for the each complete data set as well as the odd/even and easy/hard subtests. These include Cronbach's alpha and split-half Spearman-Brown for the complete data sets and, comparable to the IRT invariance analyses, Pearson correlations for odd/even and easy/hard item groupings.

Low values on these reliability indices (e.g., less than 0.70) would suggest the items within each test are performing in an inconsistent manner (something that would help explain possible problems in the invariance of IRT ability parameter estimates).

## Model-Data Fit

A critical part of any successful application of IRT to real test data is ensuring that the appropriate item response model is used to characterize the data. As part of this study, five popular item response models (the 1PL, 2PL, 3PL, PCM, and GPCM) were applied to the Grade 7 reading comprehension and numeracy data sets. PARSCALE was used to estimate the item and ability parameters for each data set and evaluate the extent to which each model fit the data. Likelihood-ratio chi-square statistics, the standard fit statistics provided in PARSCALE, were used to evaluate fit at the item and test levels. High chi-square values reflected item or test misfit. The significance of this misfit was tested at two levels: $\alpha$=0.01 and $\alpha$=0.05.

Like all statistics, the chi-square statistic is influenced by sample size, with larger sample sizes generally resulting in larger amounts of misfit (see Hambleton & Murray, 1983, pp. 73-74). Consequently, based on the recommendation of M. Marshall (personal communication, March 13, 2001), item and test fit were assessed using a range of sample sizes (500, 1,500, 3,000, and 10,000). While not based on a strict criterion, a conscious effort was made to select sample sizes of diverse sizes. The relative item and test fit of different IRT models were compared across the various sample sizes and any obvious patterns that emerged were used to make a general conclusion of relative fit across models.

To obtain a more concrete measure of relative fit, a procedure recommended by Embretson and Reise (2000, pp. 73-74) and reported in a number of studies investigating relative model fit (e.g., Choi & Bachman, 1992; Lane, Stone, Ankenmann, & Liu, 1995) was also used. With this procedure, the -2 log likelihood statistic obtained at the end of the parameter estimation process in PARSCALE was compared across models. The difference between the likelihood statistic for models varying by only one parameter (1PL-2PL, 2PL-3PL, and PCM-GPCM) was evaluated as a chi-square distribution, with the degrees of freedom equaling the number of additional item parameters estimated under the less restrictive model.

In all cases, the likelihood statistic of the less restrictive model of the pair was subtracted from the more restrictive model. As per Choi and Bachman (1992), a significant ($p<0.01$) positive difference indicated an increase in model-data fit by going to the less restrictive model. A significant negative difference offered evidence that the more restrictive model achieved better relative fit. To help control possible artifacts relating to sample size, the -2 log likelihood method was applied to the same samples used to evaluate model-data fit in the earlier procedure.

Consistent with the recommendations of Hambleton (1993, p. 173), the results obtained through the statistical procedures to assess model-data fit were complemented by the results of other analyses comprising this study. After considering all relevant information, a general position regarding relative misfit was reported.

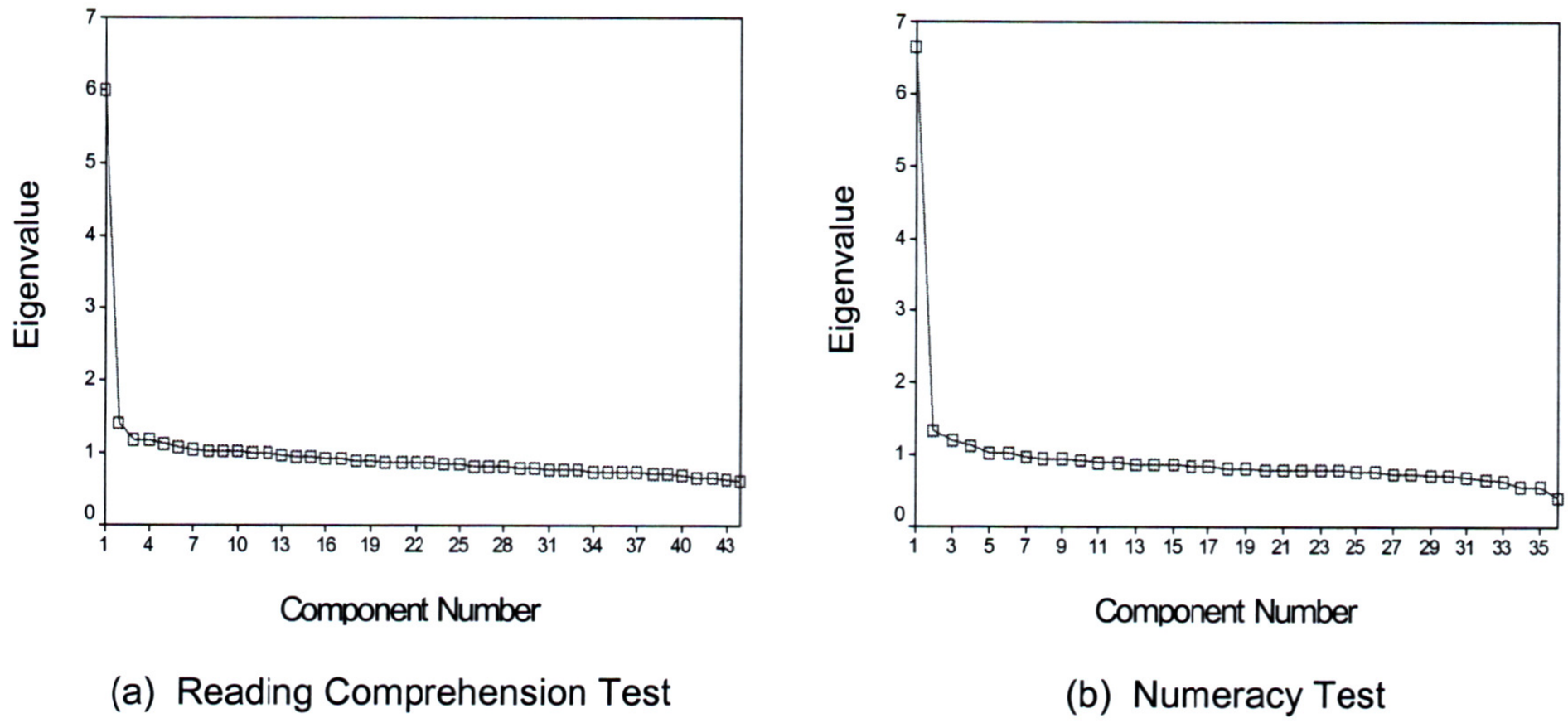
# CHAPTER 4

# RESULTS

The results of the study to assess the feasibility of using IRT to analyse FSA 2000 data are organized under the following headings: unidimensionality, local independence, nonspeededness, equal item discrimination, lower asymptote, item and ability parameter invariance, and model-data fit.

## Unidimensionality

Results of the unidimensionality tests using exploratory principal components factor analysis are presented in Figure 6 and Table 1. Figure 6 displays the scree plots of eigenvalues for all components of the reading comprehension (a) and numeracy (b) tests. Table 1 lists the eigenvalues and percentage of variance accounted for by the first five components.




(a) Reading Comprehension Test (b) Numeracy Test

Figure 6. Scree plots of eigenvalues for all components of the reading comprehension and numeracy tests.

Table 1

Percentage of Variance Accounted for the 5 Largest Components of the Reading Comprehension and Numeracy Data Sets

| | Component | Eigenvalue | % of Variance |
|---|---|---|---|
| Reading Comprehension | | | |
| | 1 | 6.00 | 13.64 |
| | 2 | 1.43 | 3.24 |
| | 3 | 1.19 | 2.71 |
| | 4 | 1.17 | 2.66 |
| | 5 | 1.14 | 2.60 |
| Numeracy | | | |
| | 1 | 6.65 | 18.46 |
| | 2 | 1.34 | 3.73 |
| | 3 | 1.22 | 3.38 |
| | 4 | 1.14 | 3.18 |
| | 5 | 1.03 | 2.85 |

As shown in Table 1, the first component of the reading comprehension test accounted for only 13.64% of the total test variance, well below the 20% required for essential unidimensionality. The first component of the numeracy test accounted for a

higher percentage of total test variance (18.46%) but still did not meet the 20% criterion for essential unidimensionality.

A review of eigenvalues plotted in Figure 6 and listed in Table 1 shows a high value for the first component and relatively equal values for the remaining components. For the reading comprehension test, the eigenvalue for the largest component was 4.2 times larger than that for the second largest component, slightly under the criterion for establishing clear dominance of the first component. For numeracy, the eigenvalue for the largest component was 5.0 times larger than the eigenvalue for the second largest component, indicating clear dominance of the first component and providing evidence for unidimensionality.

To gain further insight into the dimensionality of the data sets, a statistical test obtained through Poly-DIMTEST (a software package specifically designed to test for unidimensionality on dichotomous and polytomous data sets) was employed. Poly-DIMTEST makes use of the statistical index *T* developed by Stout (1987, 1990) to assess departure from unidimensionality. For purposes of this study, a *T* value with $p<.05$ was sufficient to reject the null hypothesis of unidimensionality (i.e., suggesting multidimensionality).

Poly-DIMTEST requires a pre-determined subset of relatively homogeneous items (called AT1) to be drawn from the entire set of items. As noted by Nandakumar et al. (1998), "this can be achieved either through expert opinion or exploratory factor analysis" (p. 101). A criticism of a version of Poly-DIMTEST designed for dichotomous items was that the items selected for AT1 may influence findings regarding dimensionality (Linacre, 1994; Wang, 1994). To help address any possible limitations in Poly-DIMTEST,

analyses using different sets of items for AT1 were run on each of the reading comprehension and numeracy data sets. Two of these analyses were based on items identified through expert opinion (denoted Expert 1 and 2) and two of these analyses were based on exploratory principal axis factor analysis (denoted FA1 and FA2). The results of these analyses are provided in Table 2, along with a general description of each item set.

As shown in Table 2, the results obtained through Poly-DIMTEST are mixed. At $\alpha = 0.05$, the null hypothesis of unidimensionality was not rejected in three out of four cases for reading comprehension. For numeracy, both AT1 subsets identified through expert opinion resulted in the rejection of the null hypothesis for unidimensionality; the two AT1 subsets identified through factor analysis, however, provided evidence for unidimensionality.

## Local Independence

To investigate the assumption of local independence, a principal axis factor analysis was performed on the reading comprehension and numeracy data sets and a residual correlation matrix with the first factor removed was produced. The residual correlation matrix for reading comprehension is provided in Appendix A and a comparable matrix for numeracy is provided in Appendix B. On both matrices, item pairs with absolute residual values of 0.10 or higher are denoted with "LR" (for large residuals). Item pairs with absolute residual values between 0.05 to 0.10 are denoted with "MR" (for moderate residuals).

Table 2

Poly-DIMTEST Results Testing Departure From Unidimensionality for Reading Comprehension and Numeracy Data Sets

| | AT1 Subset | # of Items | *T* | *p* |
|---|---|---|---|---|
| Reading Comprehension | | | | |
| | Expert 1-rc | 8 | -2.96 | >.998 |
| | Expert 2-rc | 8 | -2.81 | >.997 |
| | FA1-rc | 8 | 1.17 | >.121 |
| | FA2-rc | 8 | 6.29 | <0.001* |
| Numeracy | | | | |
| | Expert 1-n | 8 | 8.09 | <0.001* |
| | Expert 2-n | 6 | 3.26 | <0.001* |
| | FA1-n | 8 | -2.88 | >.998 |
| | FA2-n | 6 | 1.17 | >.120 |

Note. Content of each AT1 subset is described below.

Expert 1-rc: Reading comprehension items identified through expert opinion; items assessed critical analysis skills by asking students to identify the best of several options provided.

Expert 2-rc: Reading comprehension items dealing with the identification of a main idea from text; identified through expert opinion.

FA1-rc: Reading comprehension items with the eight largest positive loadings on the second factor as determined by factor analysis without rotation.

FA2-rc: Reading comprehension items with the eight largest negative loadings on the second factor as determined by factor analysis without rotation.

Expert 1-n: Numeracy items dealing with statistics and probability; identified through expert opinion.

Expert 2-n: Numeracy items dealing with patterns and relations; identified through expert opinion.

FA1-n: Numeracy items with the eight largest positive loadings on the second factor as determined by factor analysis with varimax rotation.

FA2-n: Numeracy items with the six largest positive loadings on the second factor as determined by factor analysis without rotation.

*$p<.05$, rejecting the null hypothesis of essential unidimensionality

## Local Independence of Reading Comprehension Items

The residual correlation matrix for reading comprehension (Appendix A) shows that 7 out of 946 (0.7%) unique item pairings had large residuals with absolute values of 0.10 or more. Twenty-six (2.7%) unique item pairs had moderate residuals with absolute values between 0.05 and 0.10. The largest residual obtained was 0.18.

A more detailed review of the seven pairs of items with absolute residual values of 0.10 or more showed that nine items demonstrated local dependence with at least one other item, and in all cases these residuals were positive (indicating a positive relationship remained after the first factor was removed). A review of the actual test items offered some practical insight into possible reasons for the large residuals calculated for the seven pairs of items. Adjacent multiple-choice items 17 and 18 both asked students to describe the feelings of a character in a given poem. Multiple-choice item 28 (which required students to identify the character of a story as a snake) shared large residuals with multiple-choice items 29, 31, and 32 (which related to characteristics of a snake). Multiple-choice questions 31 and 32, which dealt with the same characteristic of a snake (fangs that secrete poison), also shared a large residual. Open-ended question 5 shared large residuals with open-ended question 4 and 6. All three of these questions were based on the same passage (a narrative story on Canada's *Bluenose*).

In addition to the large residuals, a review of the moderate residuals showed all open-ended items demonstrated at a moderate relationship with at least one other open-ended question and no moderate relationship with any multiple-choice question.

## Local Independence of Numeracy Items

The residual correlation matrix for the numeracy data set (Appendix B) shows that 8 out of 630 (1.3%) unique item pairings had large residuals with absolute values of 0.10 or more. Eleven (1.7%) unique item pairs had moderate residuals with absolute values between 0.05 and 0.10. The largest residual obtained was 0.36.

A more detailed review of the eight pairs of items with absolute residual values of 0.10 or more showed that fourteen items demonstrated some local dependence with at least one other item, and in all cases these residuals were positive (indicating a positive relationship when the first factor was removed). Multiple-choice items 1 and 4, both of which involved the summing of products, shared a large residual. Multiple-choice item 2 shared large residuals with items 7 and 8; all of these questions involved the calculation of area. Multiple-choice questions 12 and 13, a pair of items with an especially large residual (0.36), were very similar questions asking students to identify possible combinations from three common sets. Multiple-choice items 14 and 15, questions that asked how much change should be received from purchases from two distinct food stands, obtained a residual greater than 0.10. A large residual was also obtained from multiple-choice items 25 and 26, both of which were probability questions based on the same prompt. Open-ended item 17 shared a large residual with the adjacent open-ended item 18; these items did not test the same content but shared commonalities in their format (both were open-ended) and their location in the test (at the end of Part 1 of the numeracy test booklet). A large residual was also obtained for open-ended items 18 and 36, both of which asked students to find all combinations of purchases based on a fixed amount of currency.

In addition to the above, one further pattern that the residual correlation matrix for numeracy shows is that three out of the four open-ended questions share at least one large residual with another open-ended question but none share a large residual with any multiple-choice questions.

## Nonspeededness

To determine whether speed was a factor in completing the reading comprehension or numeracy tests, the percentage of examinee who did not complete each of the last five items on Parts 1 and 2 of both tests was examined.

### Reading Comprehension

Ninety-eight to 99 percent of students responded to the last 5 questions in Part 1 of the reading comprehension test. A higher percentage of students, ranging from 2% to 10%, did not respond to the last 5 open-ended questions included in Part 2 of the test. Overall, the 90% response-rate criterion for nonspeededness was met by all reading comprehension items.

### Numeracy

Ninety-eight to 99 percent of students responded to the last three multiple-choice questions in Part 1 of the numeracy test. The response rate dropped to 94% for the next open-ended question and 89% for the last open-ended question on this part of the test.

As with the multiple-choice questions in Part 1, 98% to 99% of students responded to the last three multiple-choice questions in Part 2 of the numeracy test. The

response rate dropped to 93% for the last two open-ended questions in this component of the test.

To determine if the 89% response rate for the last question in Part 1 of the numeracy test was due to time constraints or item difficulty, the response rates of all remaining numeracy items were reviewed. This review showed that 93% to 100% of students completed all other items, providing evidence that time limitations may have played some role in student performance on the final item in Part 1 of the numeracy test.

## Equal Item Discrimination

The equal item discrimination assumption held by one parameter IRT models was tested by calculating the point-biserial and point-polyserial item-test score correlations for each item and determining what percentage of the correlations fall outside an acceptable range (operationalized as the average item-test score correlation ± 0.15). As presented in Table 3, the results of these analyses show 37.5% of the dichotomous reading comprehension items and only 3.1% of the dichotomous numeracy items fall outside the average item-test score correlation ± 0.15. No polytomous item fell outside the specified range in either data set.

## Lower Asymptote

To determine whether each data set possessed a zero or non-zero lower asymptote, the performance of examinees scoring in the lowest decile of each sample on the five most difficulty multiple-choice items was examined. The proportions (*p*-values) of these

Table 3

Summary Point-Biserial and Point-Polyserial Correlation Statistics

| Data Set | Point-Biserials for Dichotomous Items | Point-Polyserials for Polytomous Items |
|---|---|---|
| | Summary Descriptive Statistics | |
| Reading Comprehension | | |
| Mean | 0.52 | 0.71 |
| SD | 0.15 | 0.05 |
| Range | 0.63 | 0.11 |
| Numeracy | | |
| Mean | 0.55 | 0.74 |
| SD | 0.07 | 0.06 |
| Range | 0.29 | 0.14 |
| | Number and Percentage of Items Outside Mean ± .15 | |
| Reading Comprehension | | |
| No. of items | 15 | 0 |
| % of items | 37.5% | 0.0% |
| Numeracy | | |
| # of items | 1 | 0 |
| % of items | 3.1% | 0.0% |

students who answered the five most difficult reading comprehension and numeracy questions correctly are listed in Table 4. Overall *p*-values are also provided for each item. As shown in Table 4, the proportion of students in the lowest decile who correctly answered the five most difficult reading comprehension and numeracy question ranged from 0.10 to 0.32. Most of these proportions fell within the more limited range of 0.20 to 0.30.

Table 4

Proportion of Students in the Lowest Decile Who Answered the Five Most Difficult Multiple-Choice Questions Correctly

| | | | | | |
|---|---|---|---|---|---|
| Reading Comprehension | | | | | |
| Five Most Difficult MC Questions | MC 37 | MC 35 | MC 24 | MC 10 | MC 27 |
| Overall *p*-value | 0.39 | 0.44 | 0.51 | 0.52 | 0.62 |
| *P*-value of Students in Lowest Decile | 0.32 | 0.20 | 0.28 | 0.30 | 0.28 |
| Numeracy | | | | | |
| Five Most Difficult MC Questions | MC 3 | MC 21 | MC 25 | MC 5 | MC 7 |
| Overall *p*-value | 0.33 | 0.45 | 0.49 | 0.50 | 0.50 |
| *p*-value of Students in Lowest Decile | 0.10 | 0.18 | 0.20 | 0.20 | 0.28 |

## Item and Ability Parameter Invariance

A number of difficulties with estimating item and ability parameters using PARSCALE arose during the investigation of the invariance property. As they influenced the final results, these difficulties and their solutions are outlined here, followed by the final results of the invariance investigation.

### Difficulties Addressed Prior to Invariance Investigation

#### Problems With Conducting One Parameter Analyses With PARSCALE

The most current version of PARSCALE (3.2) was used in preliminary analyses for this study. A surprising and troublesome finding from these preliminary analyses was that, despite what its manual states, PARSCALE 3.2 would not analyse data using 1PL or partial credit models (it did not recognize the CSLOPE command required for the calculation of common slopes). This previously undocumented finding was reviewed and verified by the technical support person responsible for PARSCALE at the company that distributes the software. This difficulty was not found in other versions of PARSCALE and was resolved by using an advanced copy of the soon-to-be-released PARSCALE 3.5 (Muraki & Bock, in press).

#### Sampling Precautions and Estimation Difficulties

The investigation of item and ability parameter invariance required both random and extreme groupings of students (low scoring and high scoring) and items (easy and hard). Even with the newest version of PARSCALE, initial analyses of these groups

proved difficult, and in some cases impossible (e.g., some data sets would not meet the 0.01 convergence criterion after more than 1000 estimation cycles). To minimize problems with estimating item and ability parameters with these extreme groups, a number of steps were undertaken, as outlined below.

First, to reduce possible problems relating to sample size, all parameter invariance analyses were based on a random selection of 3,000 students (versus a potential 5,000 for random and low/high ability groupings and 10,000 for random or easy/hard subsets of items). This also speeded up the estimation procedure (which, in some cases, took over 40 minutes to complete for a data set comprising 10,000 students) and so had some practical benefit as well.

Second, to remove potential problems in parameter estimation due to extreme scores, students who scored zero or perfect on the random and easy/hard sets of items were removed prior to randomly selecting 3,000 students from the original sample.

Third, successful parameter estimation on a number of data sets depended on how PARSCALE command files were structured. Some data sets, specifically the Low and High Scoring Groups in reading comprehension, would not converge using the 1PL-PCM if the command file was designed to estimate dichotomous and polytomous items separately and combine their relevant parameter estimates at the end of the estimation process. In the two data sets mentioned, estimates could not be obtained for the polytomous items when analysed separately from the dichotomous items. However, if the PARSCALE command file was structured so that dichotomous and polytomous items were jointly estimated as one test (i.e., without the need to combine estimates at the end of the estimation process), the 1PL-PCM converged for these data sets. Although

increasing the amount of analyses required for purposes of the study, this finding was useful because it helped identify starting values for certain parameters (e.g., the difficulty parameter) that could be inputted into command files where joint estimation was not possible (e.g., 2PL-PCM analyses).

To complicate things, a finding reverse to the one described above was obtained for two parameter analyses of other data sets (High Scoring Group in reading and High and Low Scoring Groups in numeracy). In all three of these cases, convergence was obtained for the 2PL-GPCM only if the parameters were estimated separately and then combined at the end.

The finding that successful parameter estimation sometimes depended on how the PARSCALE command file was structured complicated the parameter invariance investigation. To enable equivalent comparisons to be made across item or ability groupings, data sets that were directly compared as part of the parameter invariance investigation were analysed using the same estimation procedure.

Even with the precautionary steps and unexpected findings listed above, one last difficulty had to be addressed during the parameter invariance investigation. A number of data sets (Even and Hard Items and Random Groups 1 and 2 for reading comprehension; Odd and Even Items and Low and High Scoring Examinee Groups for numeracy) did not meet the convergence criterion of 0.01 for each estimation cycle in PARSCALE using a 3PL model. To help determine if the estimation problem was due to limitations in PARSCALE, multiple-choice items were analysed through BILOG 3.11 (Mislevy & Bock, 1998). In each case, parameter estimates were quickly obtained using a 3PL model in BILOG, suggesting a possible weakness of PARSCALE in estimating dichotomous

data using the 3PL model. To enable an investigation of parameter invariance for the complete data sets, steps were taken to incorporate BILOG *c*-parameter values into the PARSCALE analyses (based on the hypothesis that PARSCALE was having difficulty estimating the *c*-parameter in these extreme data sets). Confirmed by PARSCALE technical support personnel as being an acceptable procedure, BILOG estimates were calculated for each data set using the same scale as was used for PARSCALE (a logistic function with a scaling factor of 1.7). The *c*-parameter values obtained through BILOG were then entered into the relevant PARSCALE command files and each data set was re-analysed. These alternative procedures resulted in convergence for all but one of these problematic data sets (the low ability data set for numeracy still did not converge). To avoid potential confounds resulting from these alternative procedures and facilitate equivalent comparisons, joint PARSCALE-BILOG analyses were carried out for all data sets that were to be directly compared to the problematic data sets.

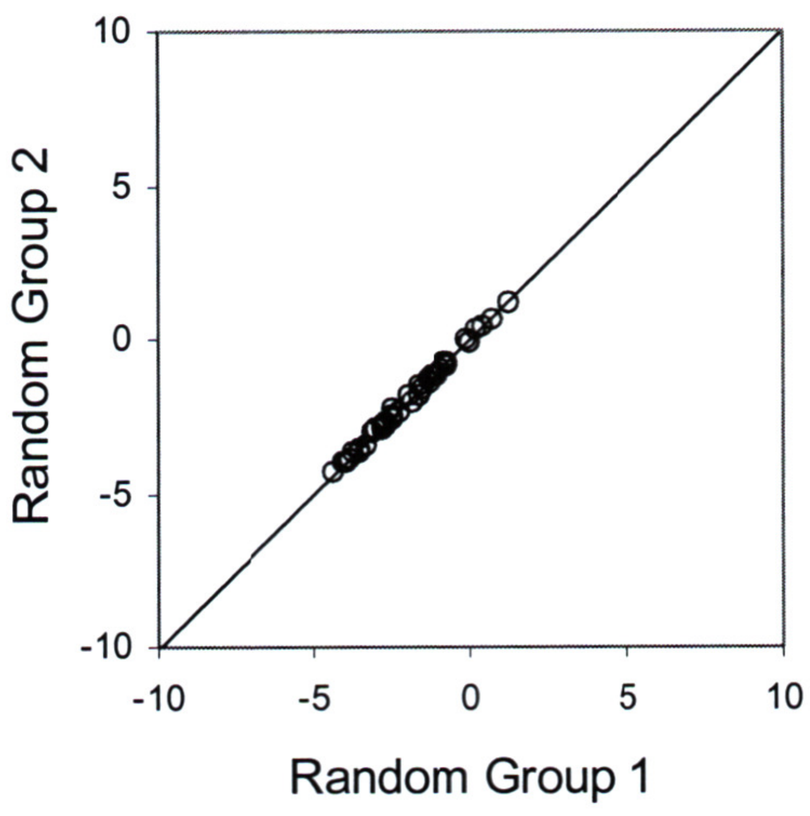
## Results From the Item and Ability Parameter Invariance Investigation

The final results of the investigation to determine if item and ability parameter invariance was obtained for the relevant IRT models are reported here.

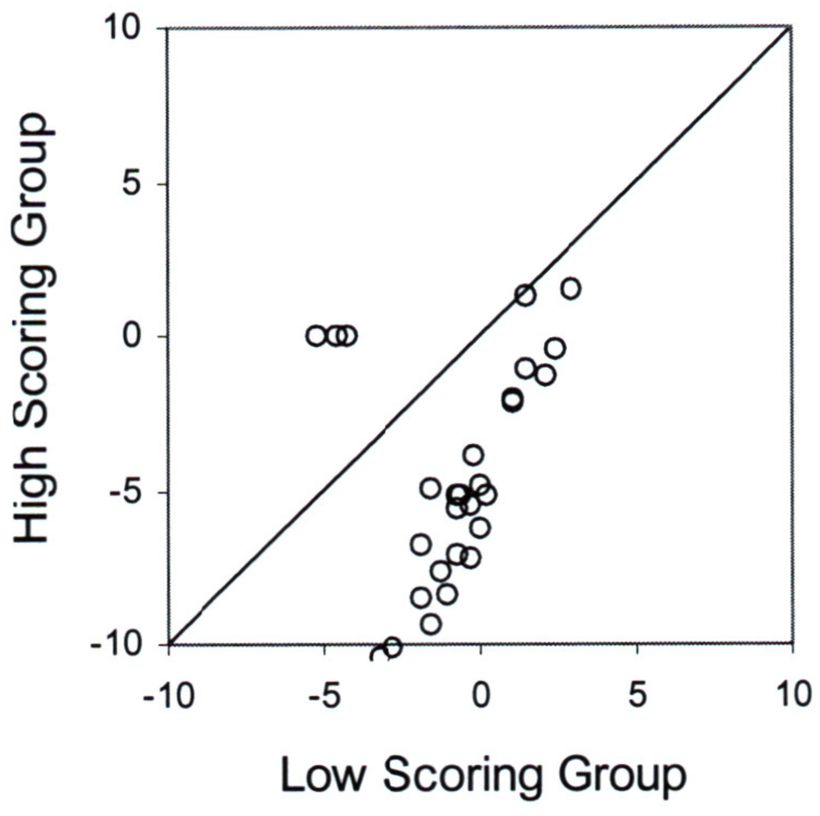
### Item Parameter Invariance

Item parameter invariance for each IRT model was assessed by correlating and plotting the *b*-parameter (item difficulty) estimates from two random samples of students (Random Groups 1 and 2) and comparing these results to the correlations and plots of low and high performing students (Low and High Scoring Groups). The scatterplots of the *b*-

parameter estimates for the full reading comprehension test (comprising all dichotomous and polytomous items) across combinations of relevant IRT models (i.e., 1PL-PCM, 2PL-PCM, 2PL-GPCM, 3PL-PCM, and 3PL-GPCM) are provided in Figures 7a to 7e. Related Pearson product-moment correlations ($r$) for the full test and dichotomous and polytomous items on the test are also reported. Comparable scatterplots and correlations for the numeracy test are provided in Figures 8a to 8e. To assist in the interpretation of plotted results, a diagonal line with a slope of 1 is included in each graph. Plotted values scattered closely around this line would provide clear evidence for parameter invariance.

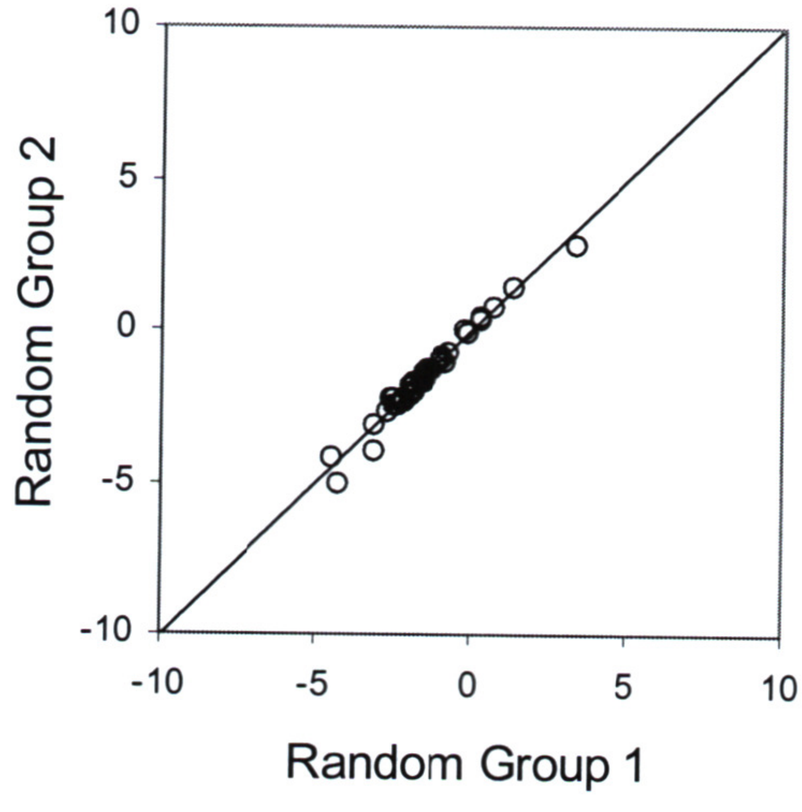



Full Test $r$ = 1.00
Dichotomous Items $r$ = 1.00
Polytomous Items $r$ = 1.00




Full Test $r$ = 0.64
Dichotomous Items $r$ = 0.56
Polytomous Items $r$ = 0.97

Figure 7a. Scatterplots and related correlations of reading comprehension item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 1PL-PCM.



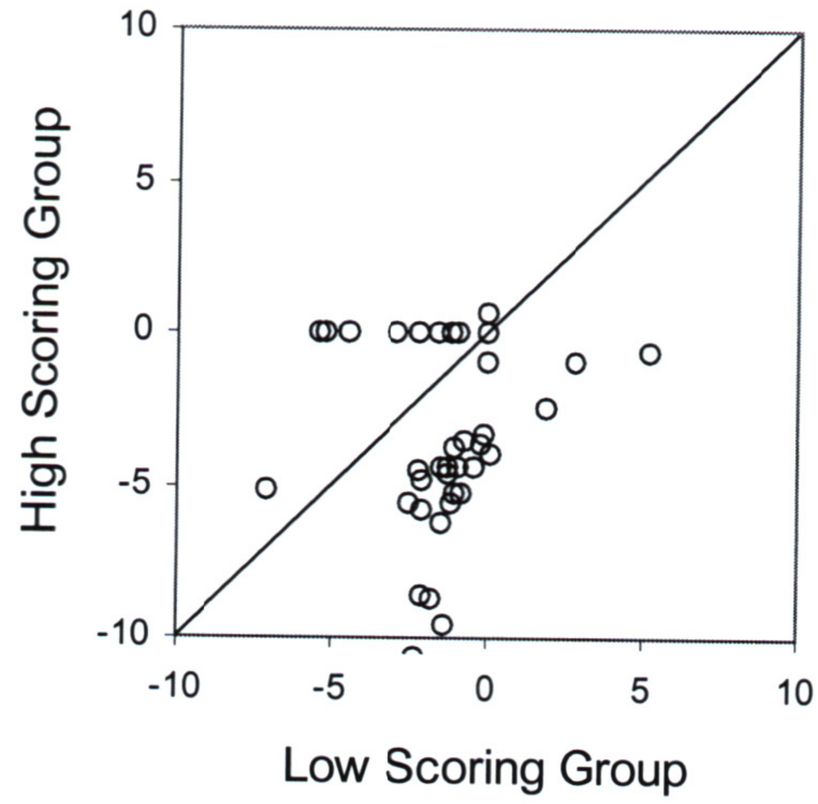
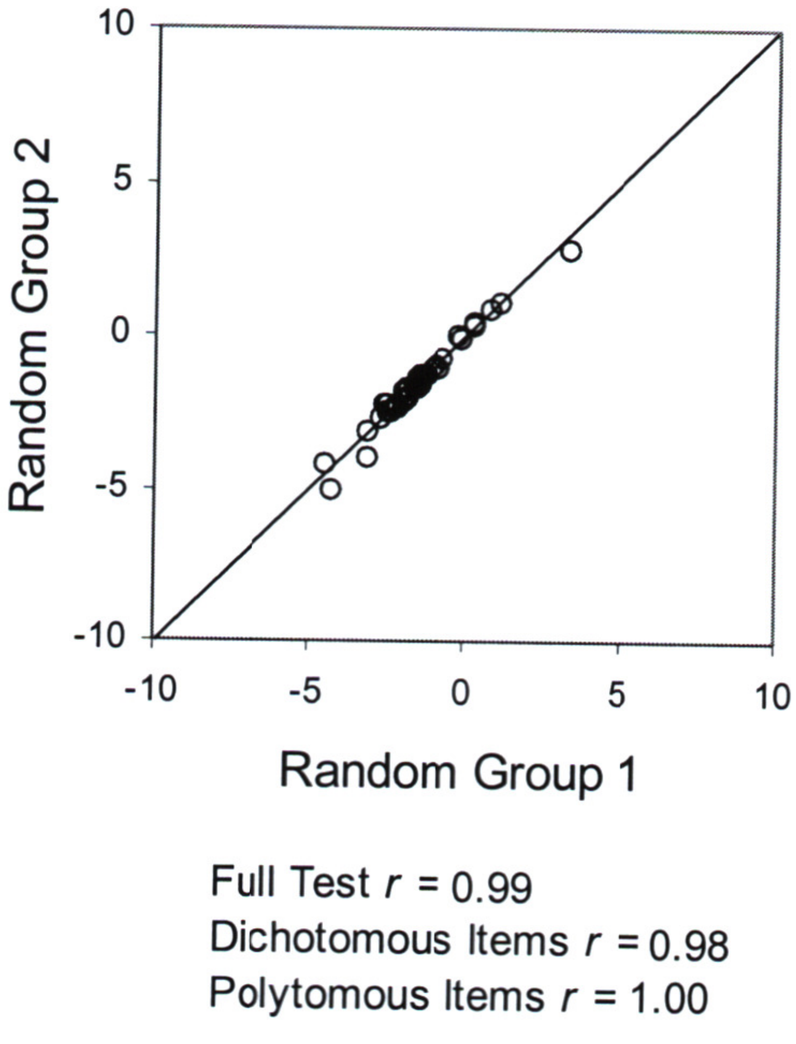




Figure 7b. Scatterplots and related correlations of reading comprehension item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 2PL-PCM.



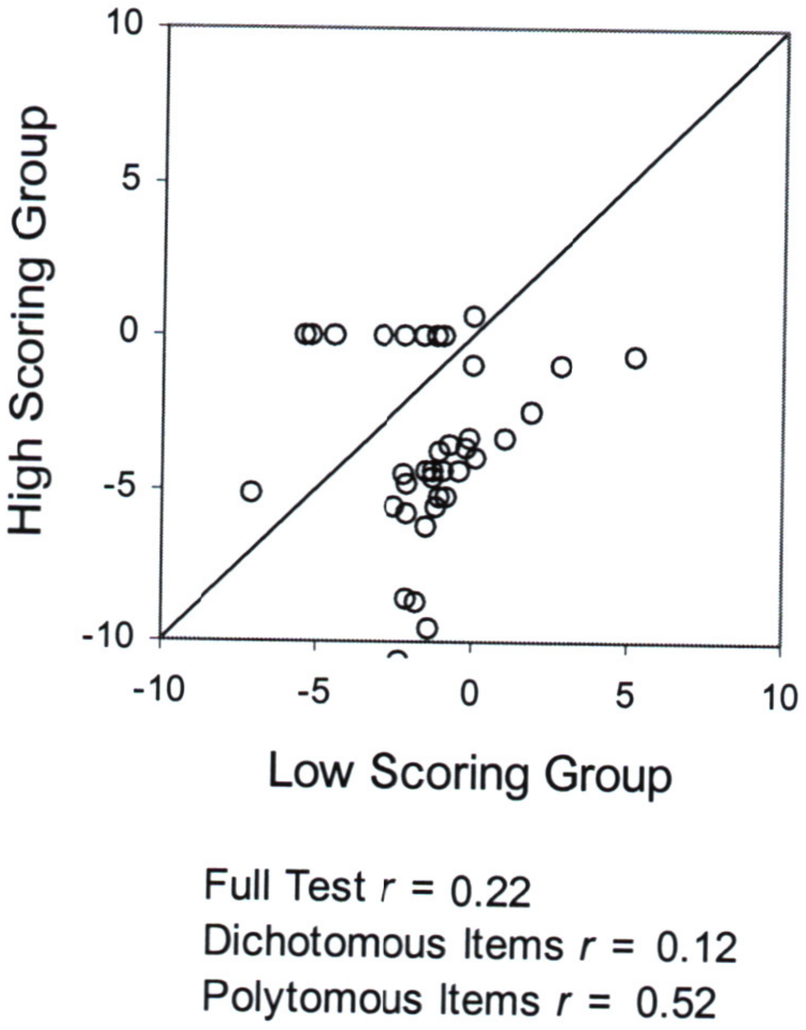
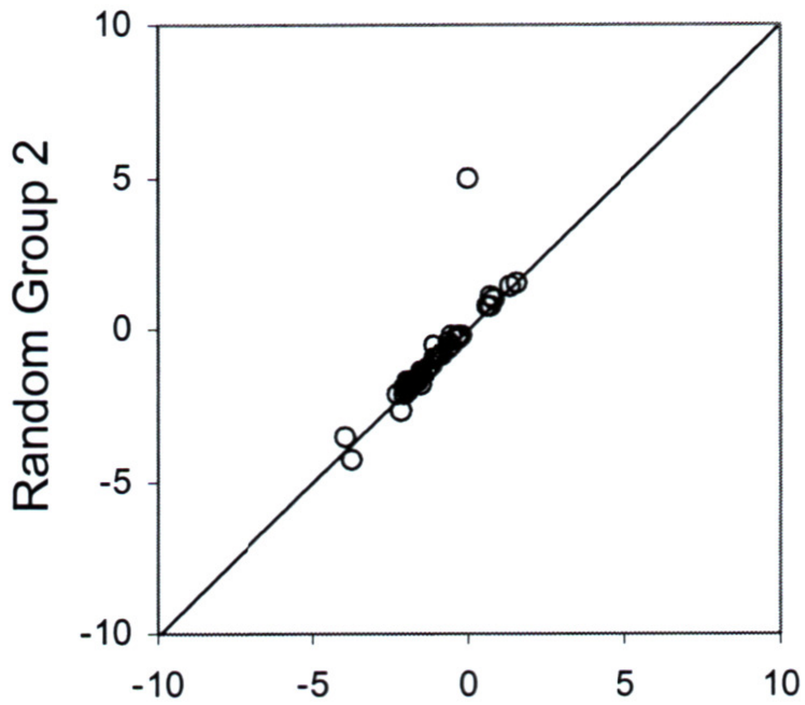
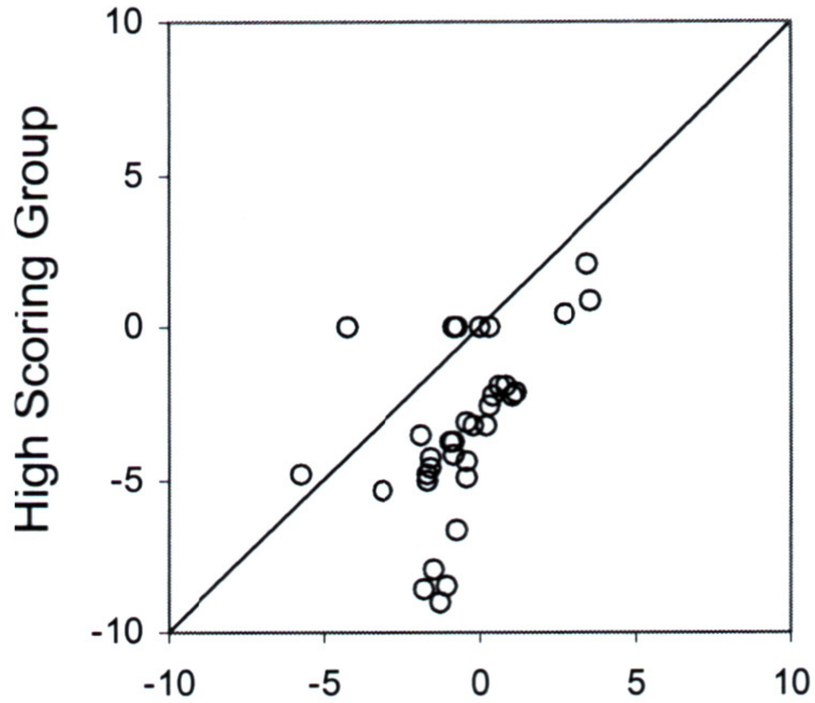




Figure 7c. Scatterplots and related correlations of reading comprehension item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 2PL-GPCM.

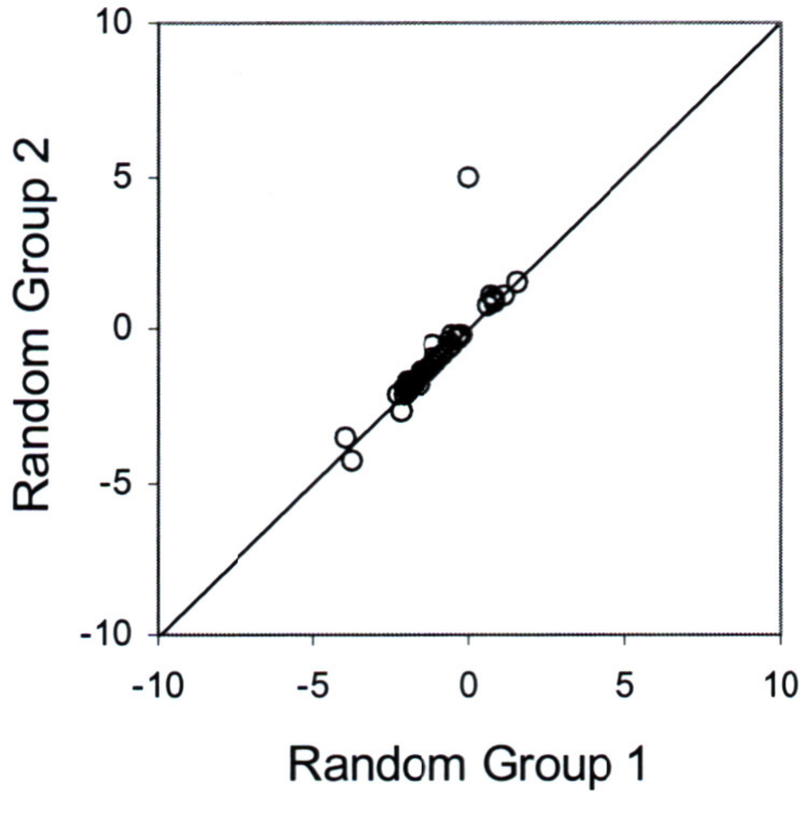



Full Test $r$ = 0.87
Dichotomous Items $r$ = 0.86
Polytomous Items $r$ = 1.00

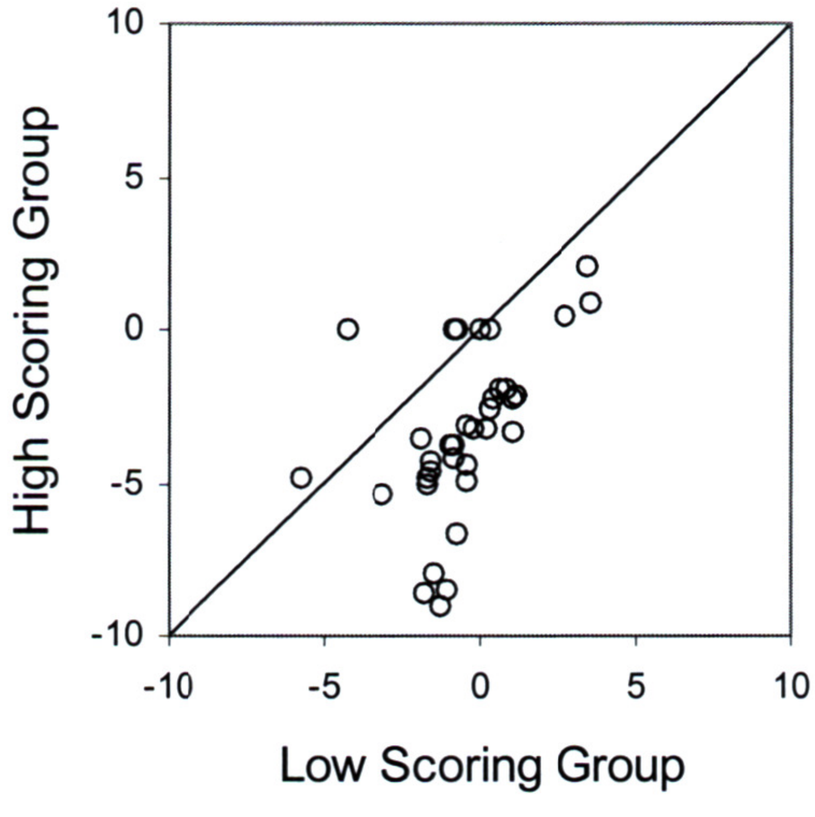



Full Test $r$ = 0.13
Dichotomous Items $r$ = 0.47
Polytomous Items $r$ = 0.67

Figure 7d. Scatterplots and related correlations of reading comprehension item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 3PL-PCM.

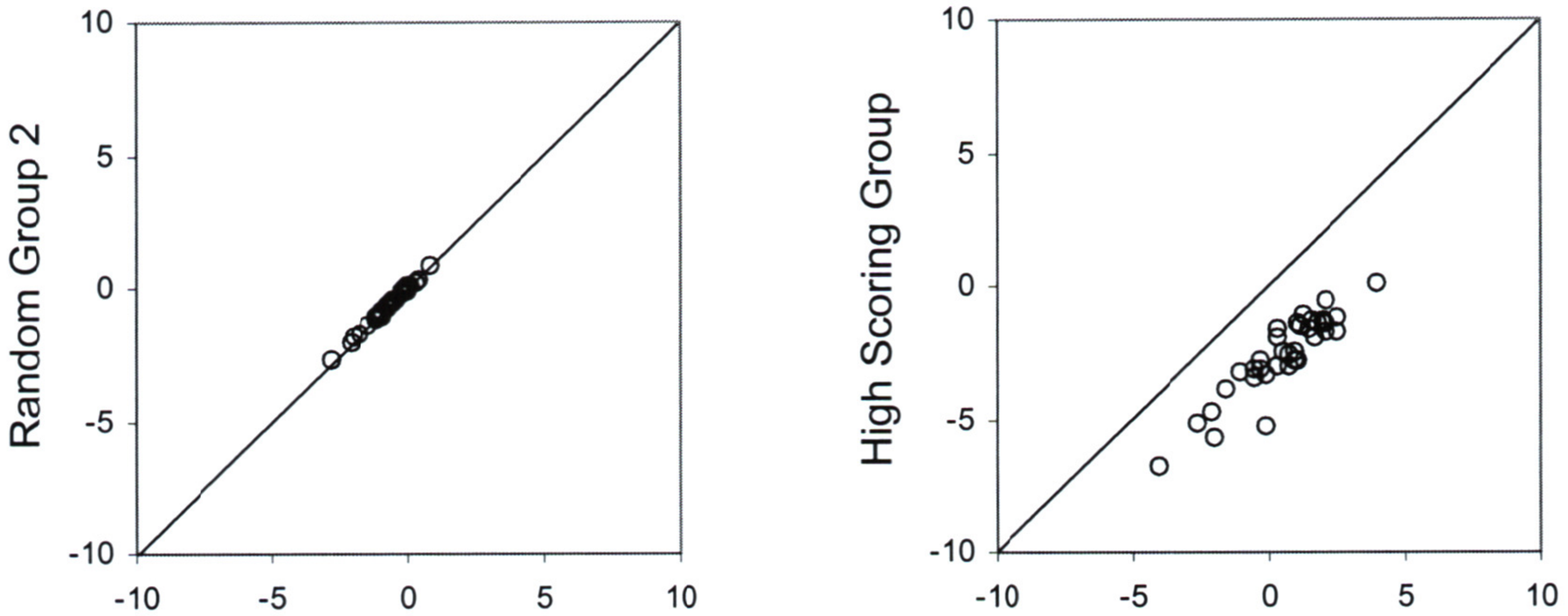



Full Test $r$ = 0.86
Dichotomous Items $r$ = 0.86
Polytomous Items $r$ = 1.00




Full Test $r$ = 0.24
Dichotomous Items $r$ = 0.47
Polytomous Items $r$ = 0.52

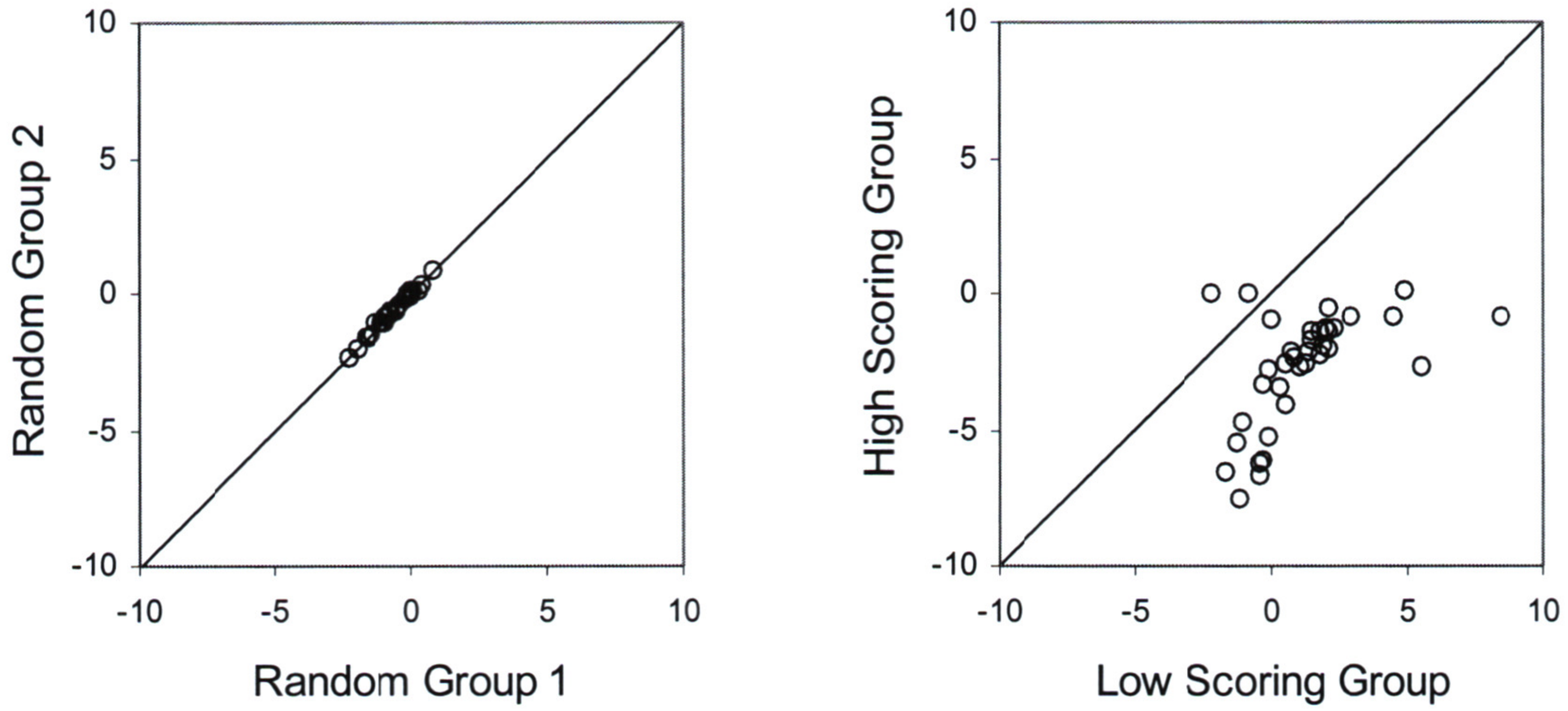
Figure 7e. Scatterplots and related correlations of reading comprehension item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 3PL-GPCM.




Full Test $r$ = 1.00
Dichotomous Items $r$ = 1.00
Polytomous Items $r$ = 1.00

Full Test $r$ = 0.91
Dichotomous Items $r$ = 0.93
Polytomous Items $r$ = 0.99

Figure 8a. Scatterplots and related correlations of numeracy item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 1PL-PCM.




Full Test $r$ = 0.99
Dichotomous Items $r$ = 0.99
Polytomous Items $r$ = 1.00

Full Test $r$ = 0.52
Dichotomous Items $r$ = 0.50
Polytomous Items $r$ = 0.99

Figure 8b. Scatterplots and related correlations of numeracy item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 2PL-PCM.

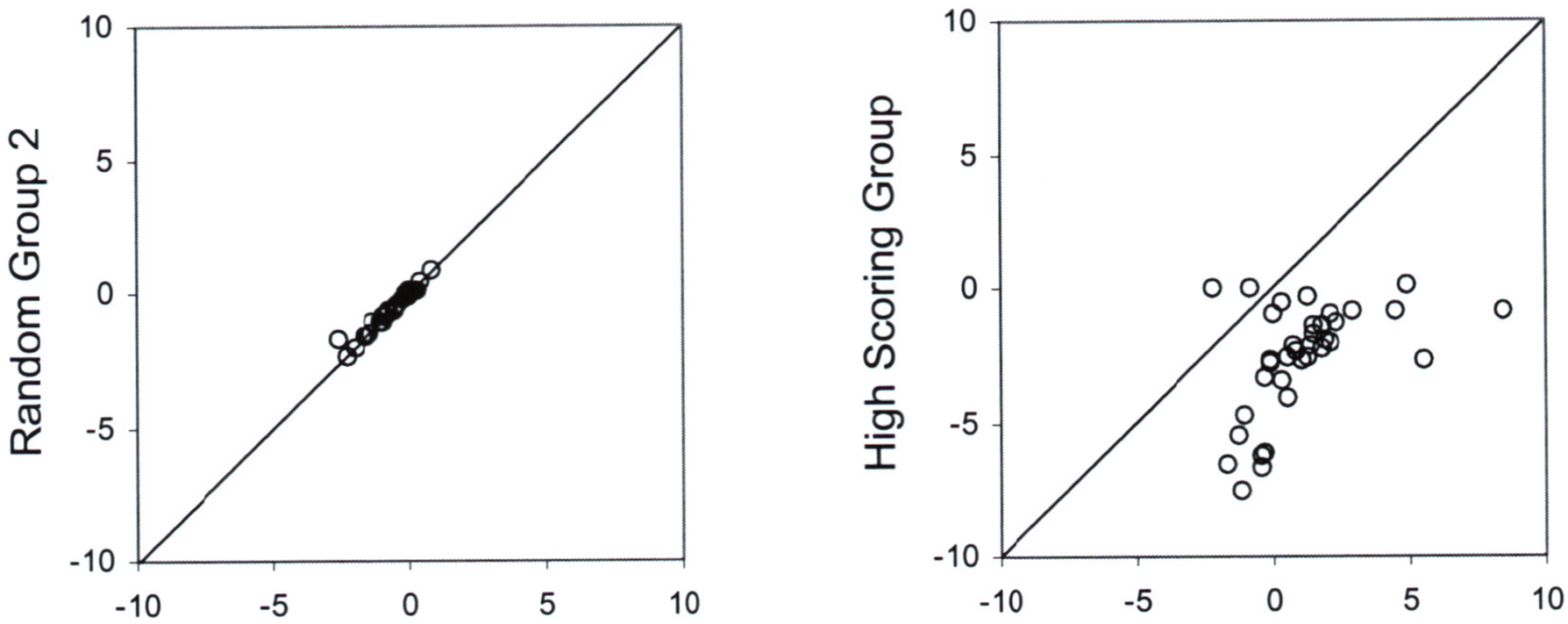



Figure 8c. Scatterplots and related correlations of numeracy item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 2PL-GPCM.

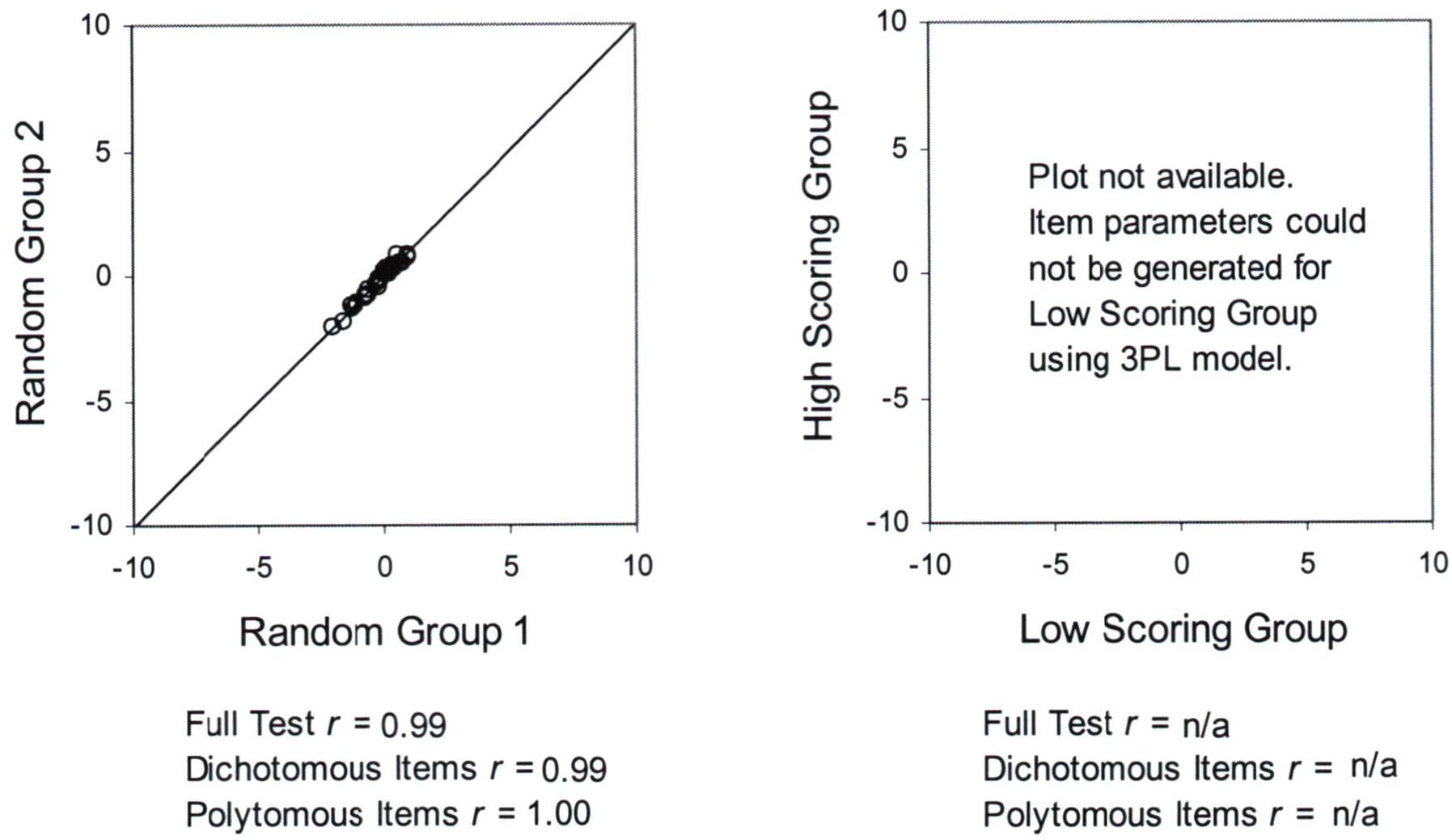



Figure 8d. Scatterplots and related correlations of numeracy item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 3PL-PCM. Item parameters could not be estimated for Low Scoring Group using 3PL model (would not converge).

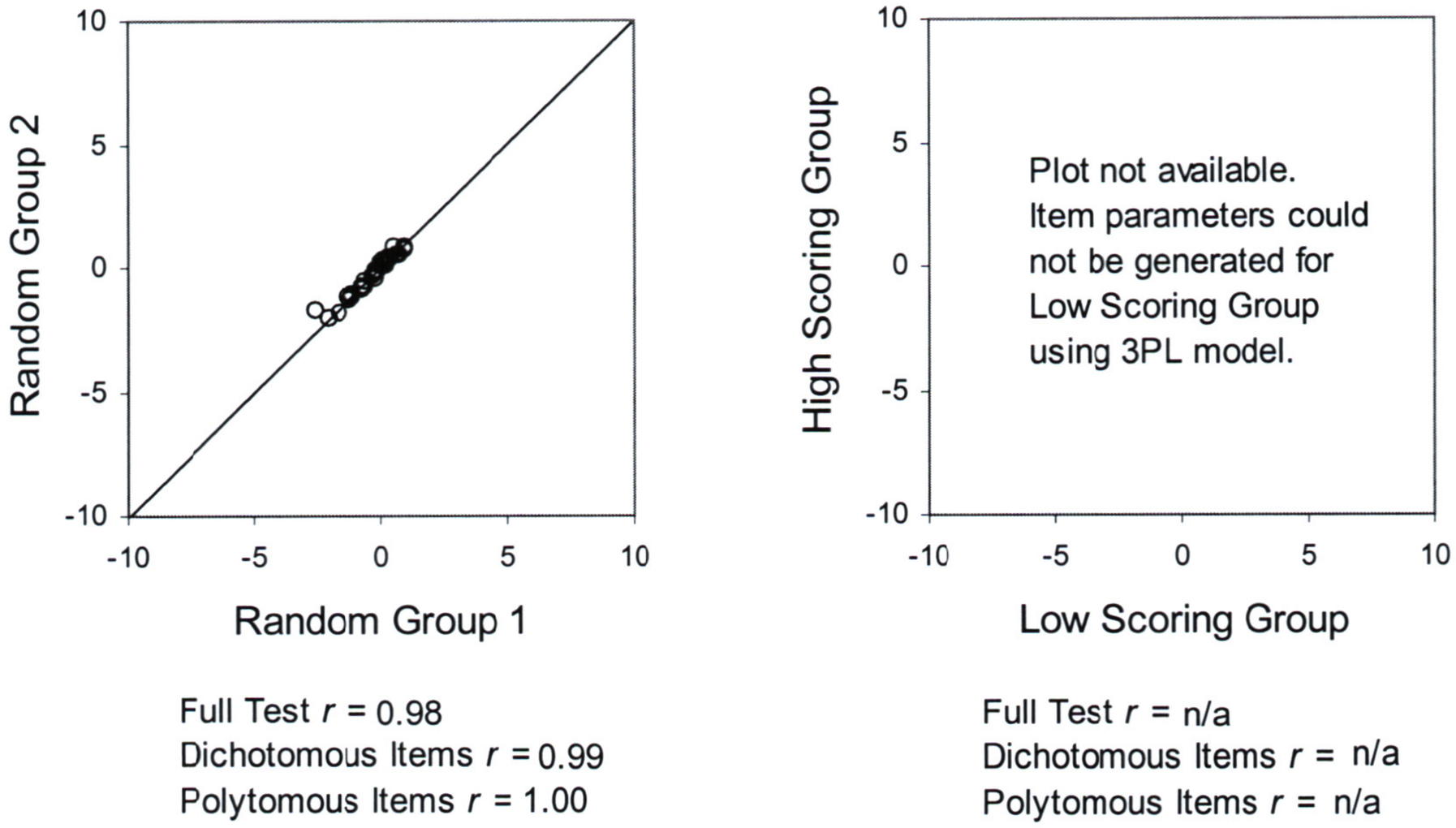



Figure 8e. Scatterplots and related correlations of numeracy item difficulty estimates for Random Groups 1/2 and Low/High Scoring Groups using 3PL-GPCM. Item parameters could not be estimated for Low Scoring Group using 3PL model (would not converge).

Ability Parameter Invariance

Ability parameter invariance for each IRT model was assessed by correlating and plotting the ability parameter of students on two randomly equivalent test forms (Odd and Even Items) and comparing the results to the correlations and plots obtained from test forms comprising easy or hard items. The scatterplots of ability estimates for reading comprehension obtained from all (i.e., dichotomous and polytomous) odd, even, easy, and hard items across combinations of relevant IRT models (i.e., 1PL-PCM, 2PL-PCM, 2PL-GPCM, 3PL-PCM, and 3PL-GPCM) are provided in Figures 9a to 9e. Related Pearson product-moment correlations ($r$) for the data depicted in each scatterplot (denoted MC+OE Items) and dichotomous and polytomous items from each subset of the test are also reported, as well as the Spearman-Brown reliability estimate based on the randomly

equivalent test forms (providing a corrected measure of reliability for the full length test). Comparable scatterplots and correlations for the numeracy test are provided in Figures 10a to 10e. As was done for the item parameter invariance investigation, a diagonal line with a slope of 1 is included in each graph. Plotted values scattered closely around this line would provide clear evidence for ability invariance.

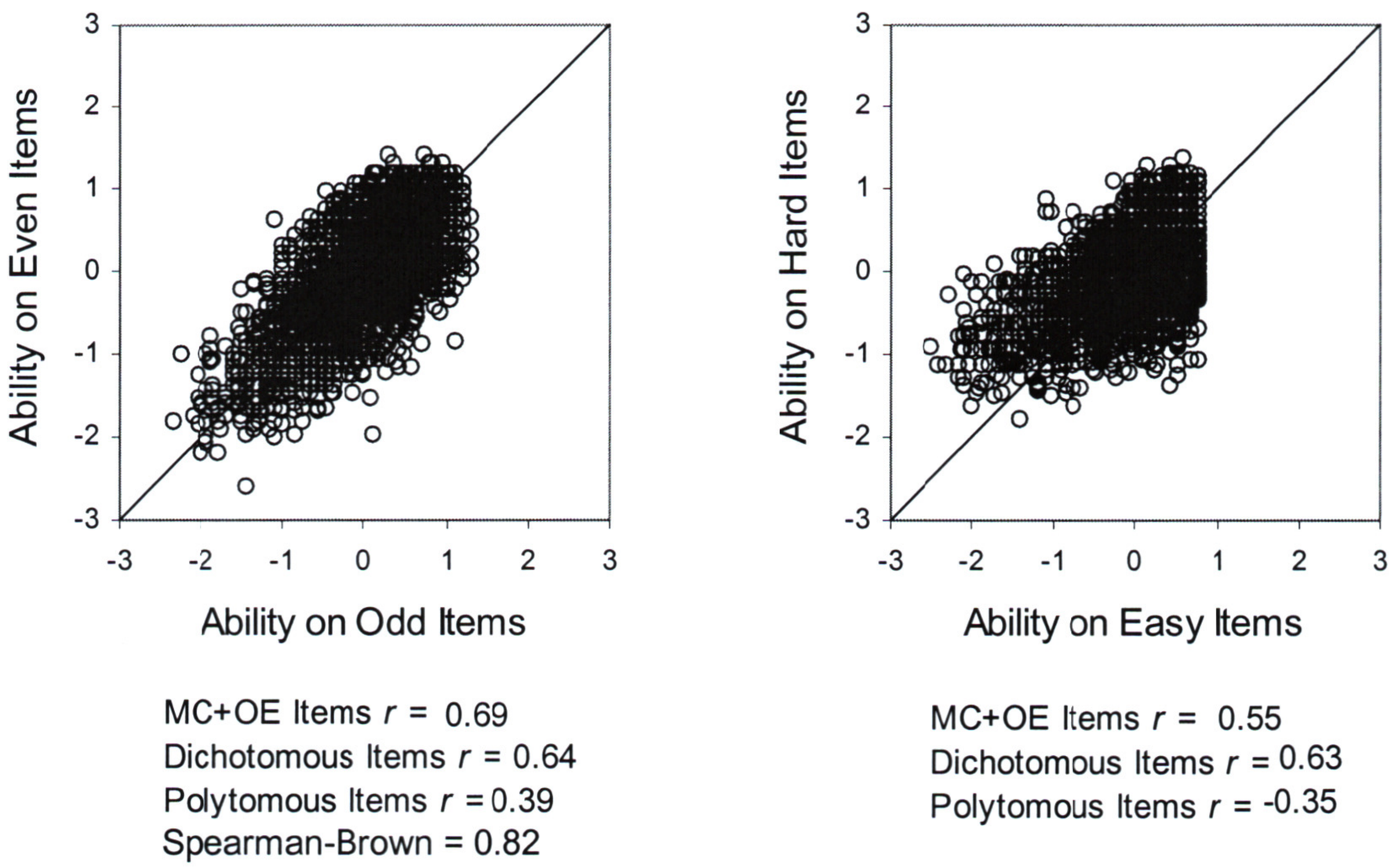



Figure 9a. Scatterplots and related correlations of reading comprehension ability estimates for Odd/Even and Easy/Hard Items using 1PL-PCM.

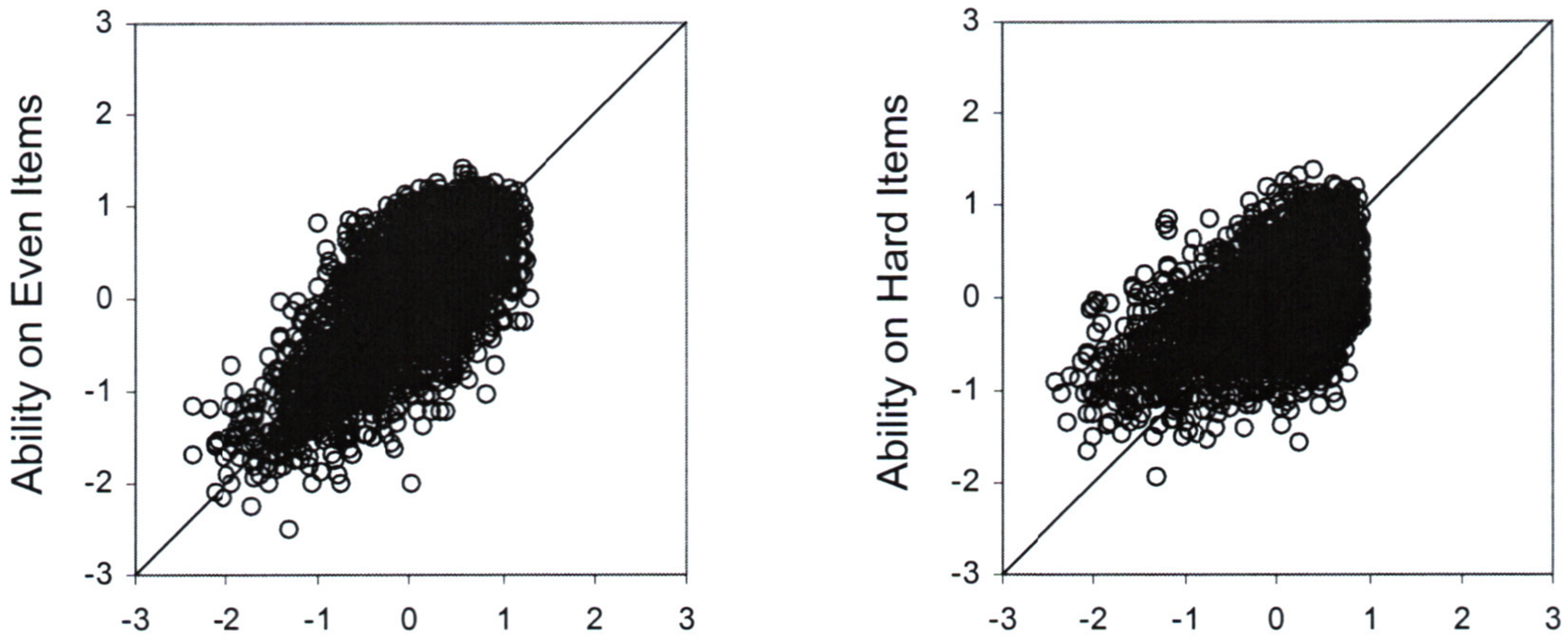



Figure 9b. Scatterplots and related correlations of reading comprehension ability estimates for Odd/Even and Easy/Hard Items using 2PL-PCM.

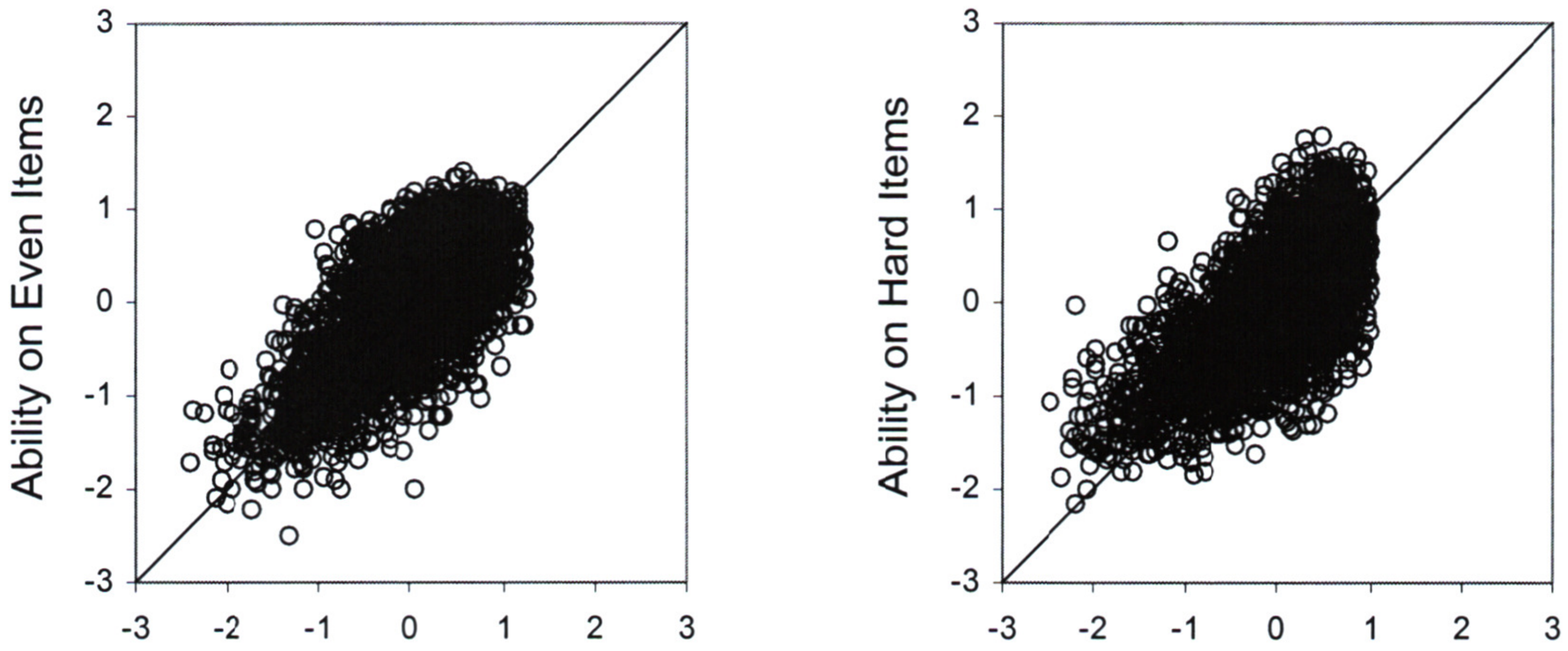



Figure 9c. Scatterplots and related correlations of reading comprehension ability estimates for Odd/Even and Easy/Hard Items using 2PL-GPCM.

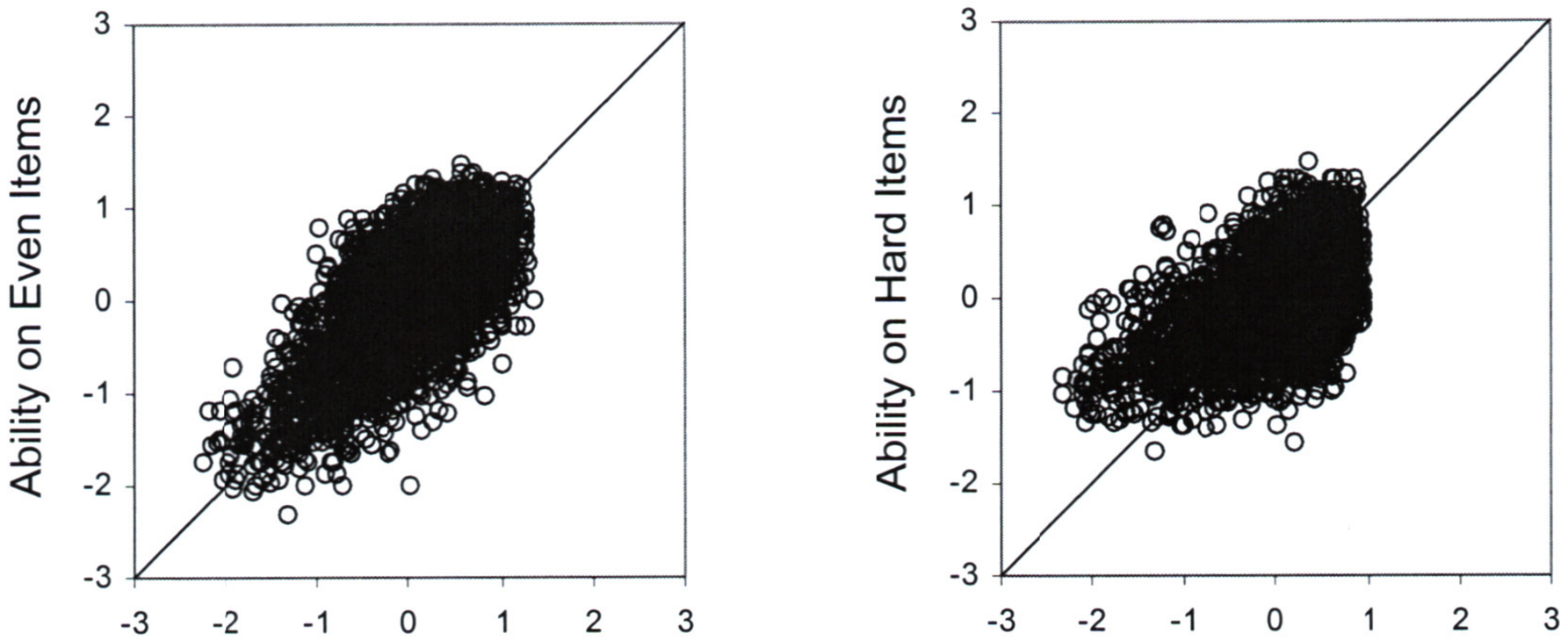



Figure 9d. Scatterplots and related correlations of reading comprehension ability estimates for Odd/Even and Easy/Hard Items using 3PL-PCM.

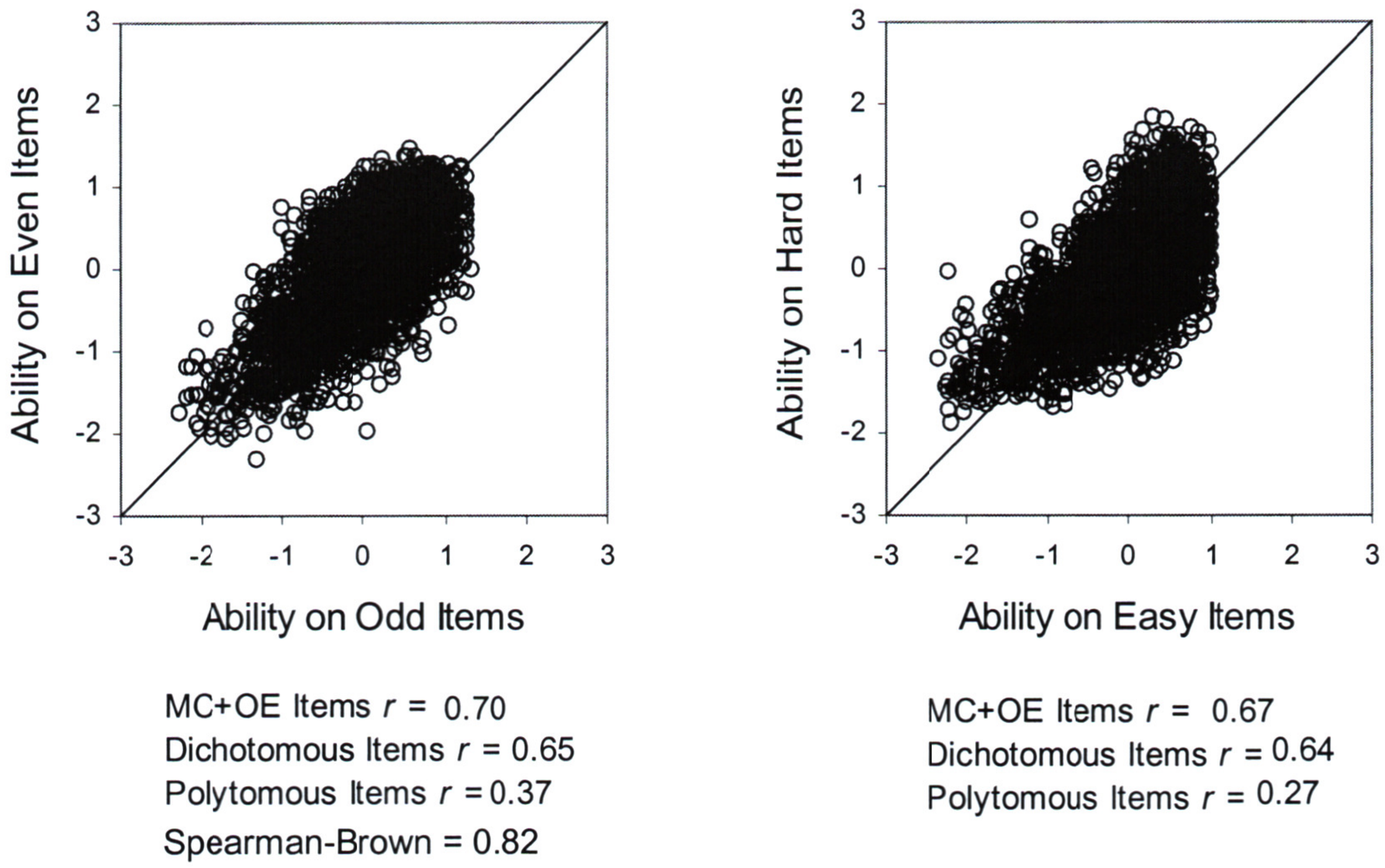
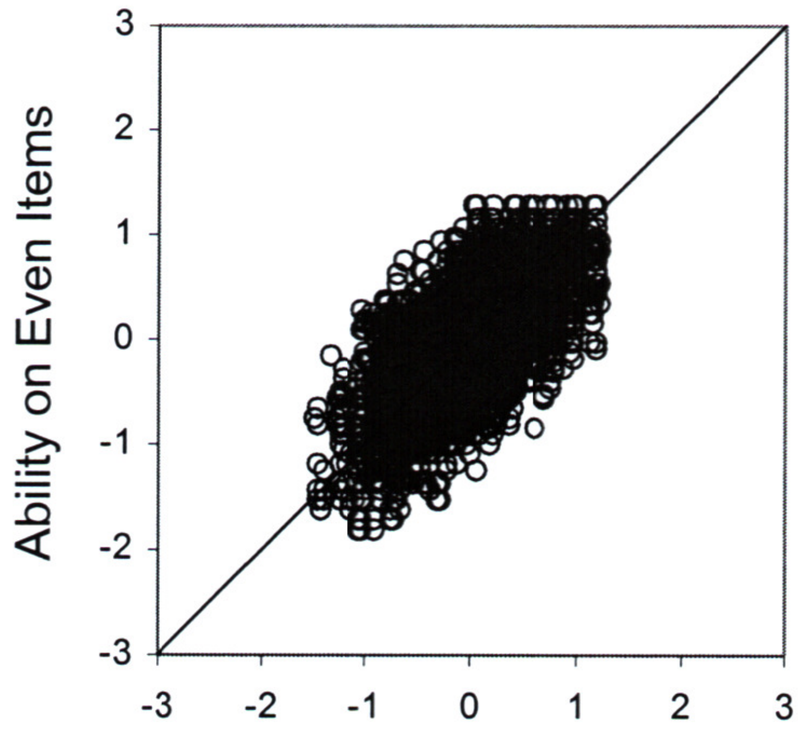
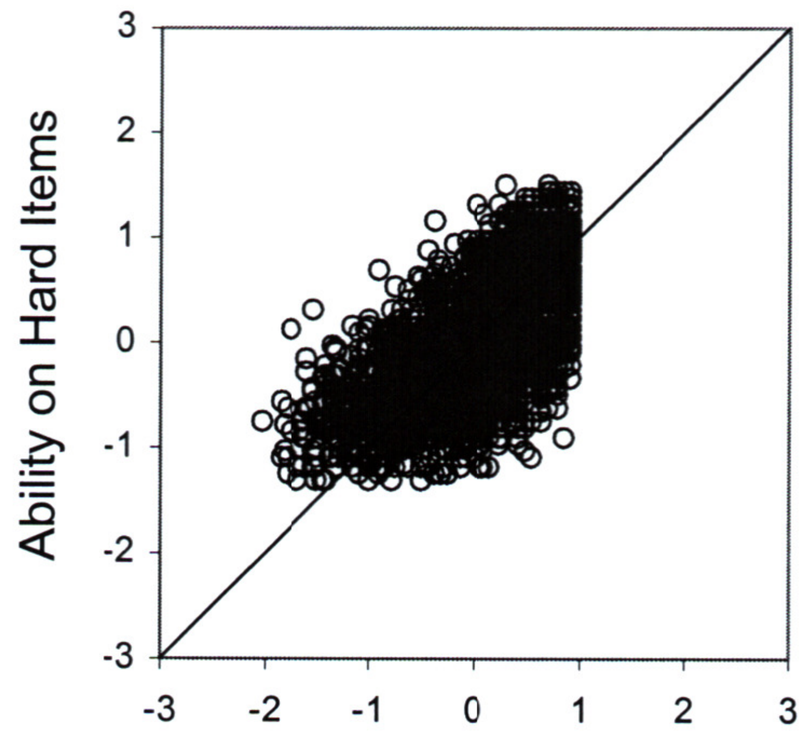



Figure 9e. Scatterplots and related correlations of reading comprehension ability estimates for Odd/Even and Easy/Hard Items using 3PL-GPCM.

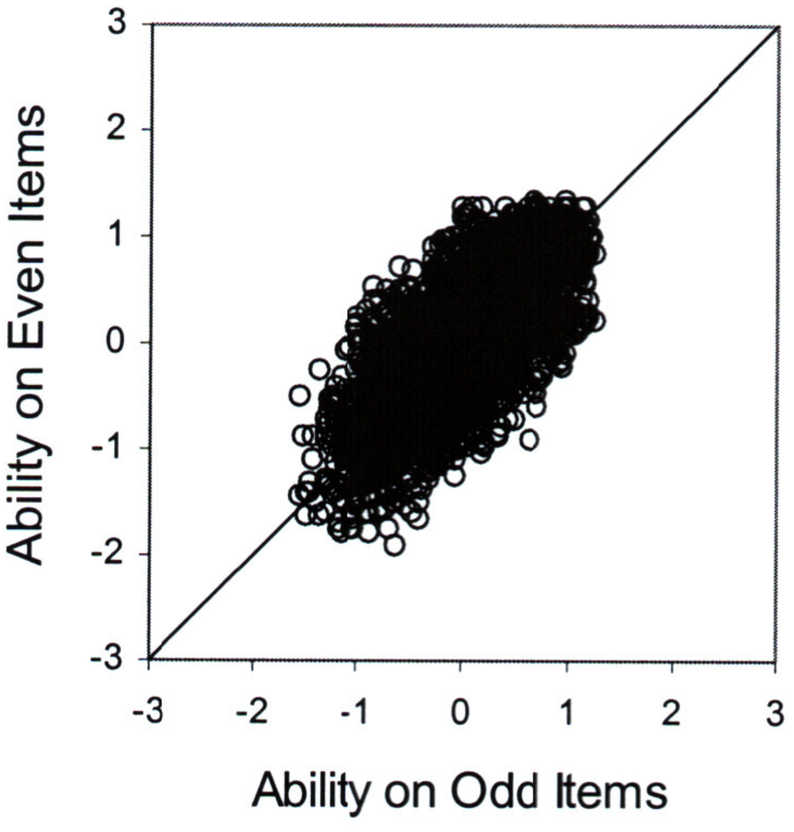



MC+OE Items $r$ = 0.71
Dichotomous Items $r$ = 0.69
Polytomous Items $r$ = 0.30
Spearman-Brown = 0.83

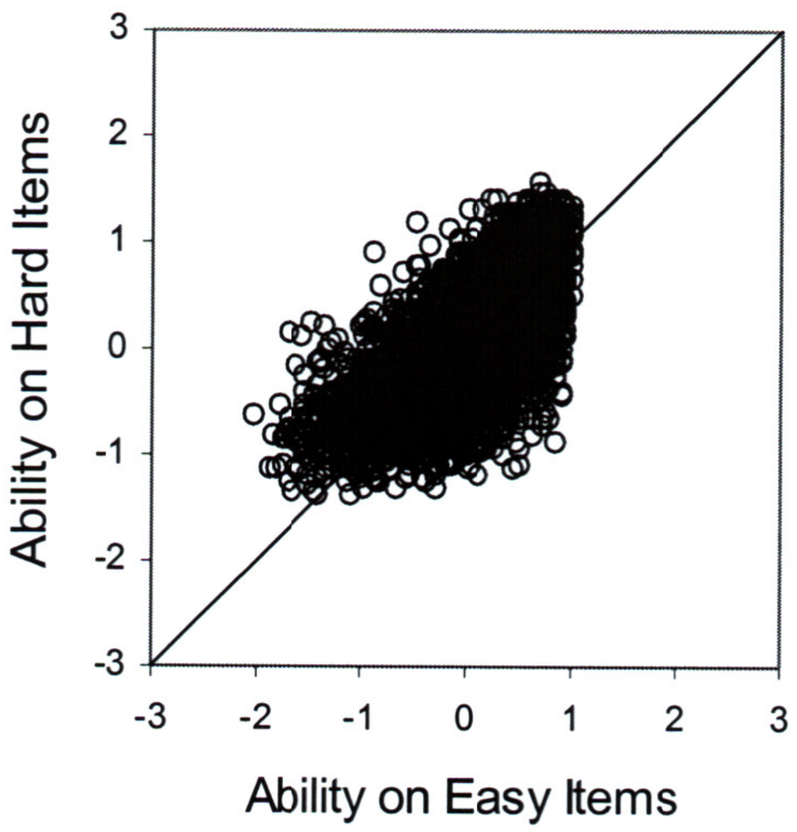



MC+OE Items $r$ = 0.66
Dichotomous Items $r$ = 0.61
Polytomous Items $r$ = 0.31

Figure 10a. Scatterplots and related correlations of numeracy ability estimates for Odd/Even and Easy/Hard Items using 1PL-PCM.




MC+OE Items $r$ = 0.71
Dichotomous Items $r$ = 0.69
Polytomous Items $r$ = 0.30
Spearman-Brown = 0.83




MC+OE Items $r$ = 0.65
Dichotomous Items $r$ = 0.60
Polytomous Items $r$ = 0.31

Figure 10b. Scatterplots and related correlations of numeracy ability estimates for Odd/Even and Easy/Hard Items using 2PL-PCM.

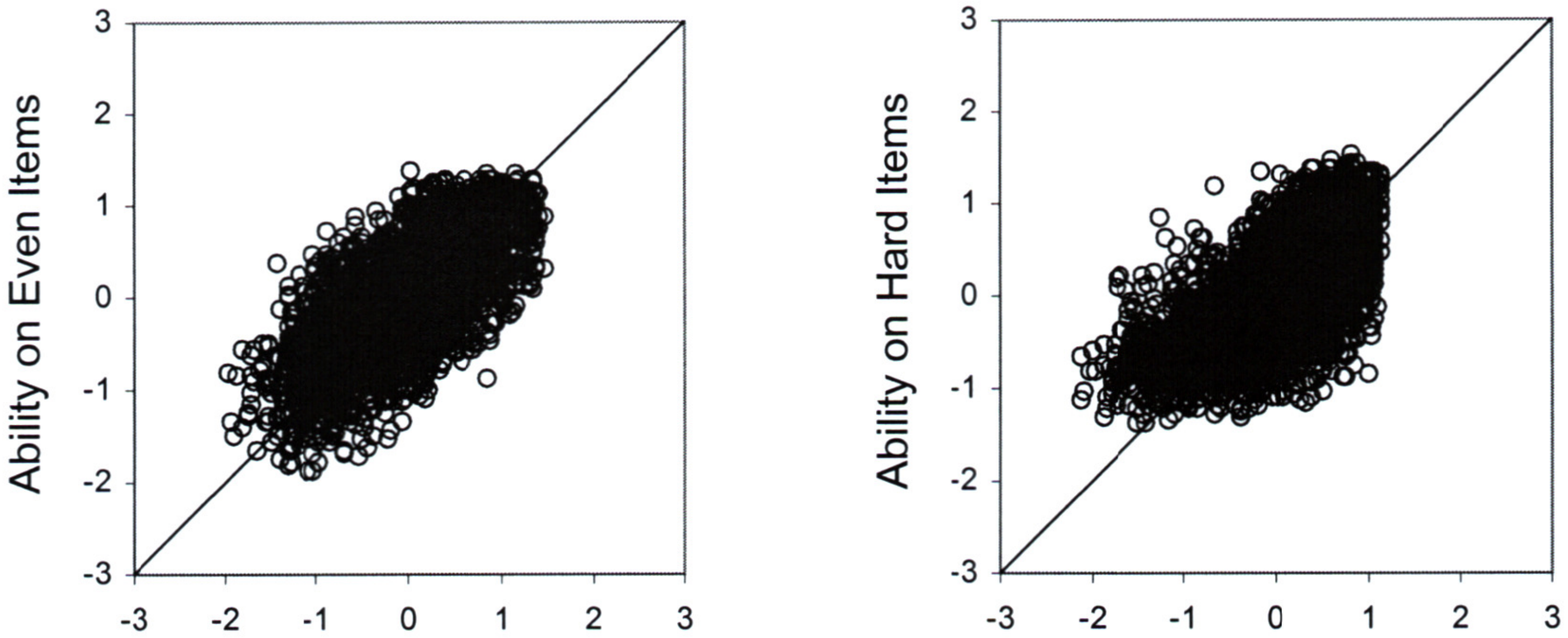



Figure 10c. Scatterplots and related correlations of numeracy ability estimates for Odd/Even and Easy/Hard Items using 2PL-GPCM.

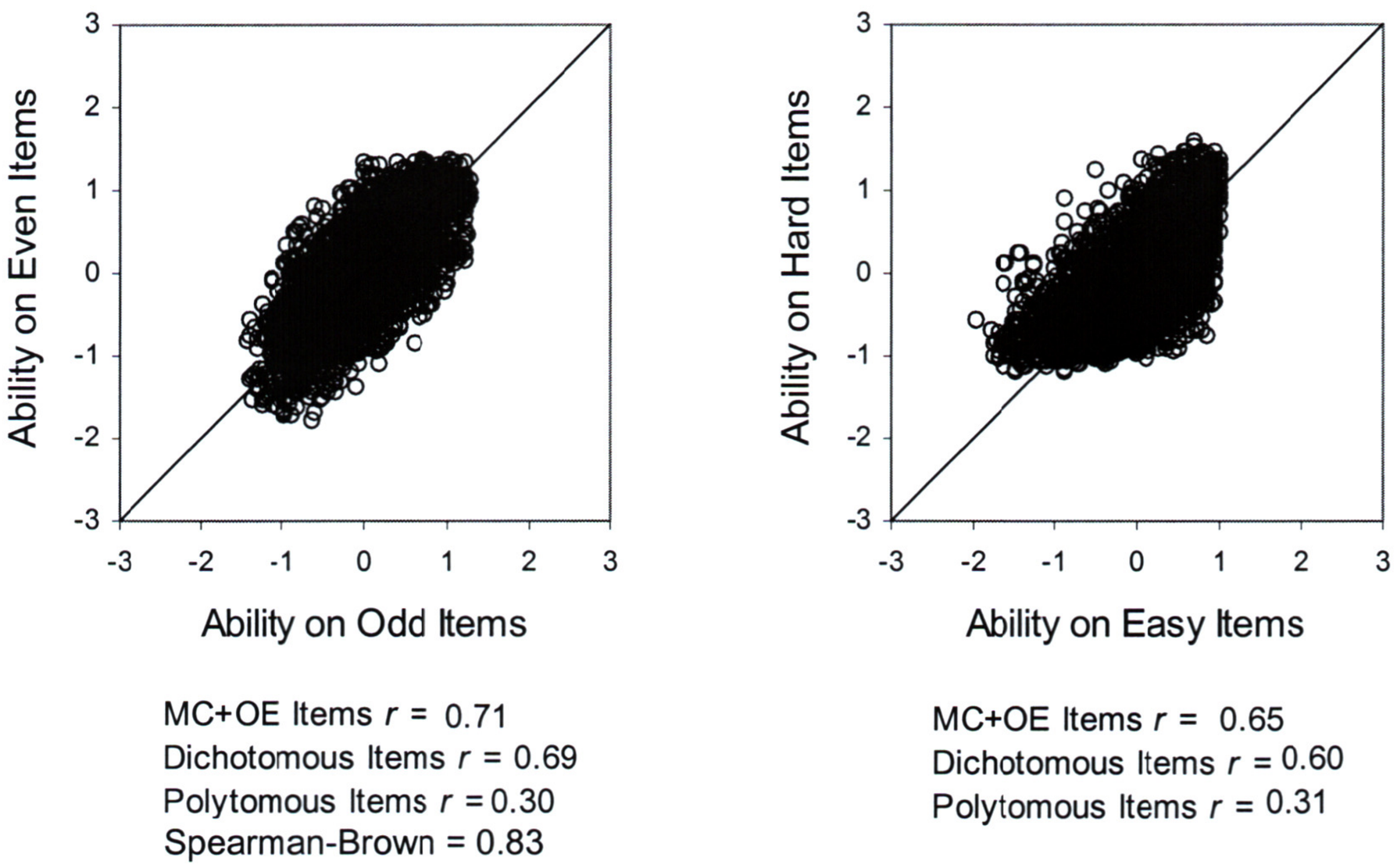



Figure 10d. Scatterplots and related correlations of numeracy ability estimates for Odd/Even and Easy/Hard Items using 3PL-PCM.

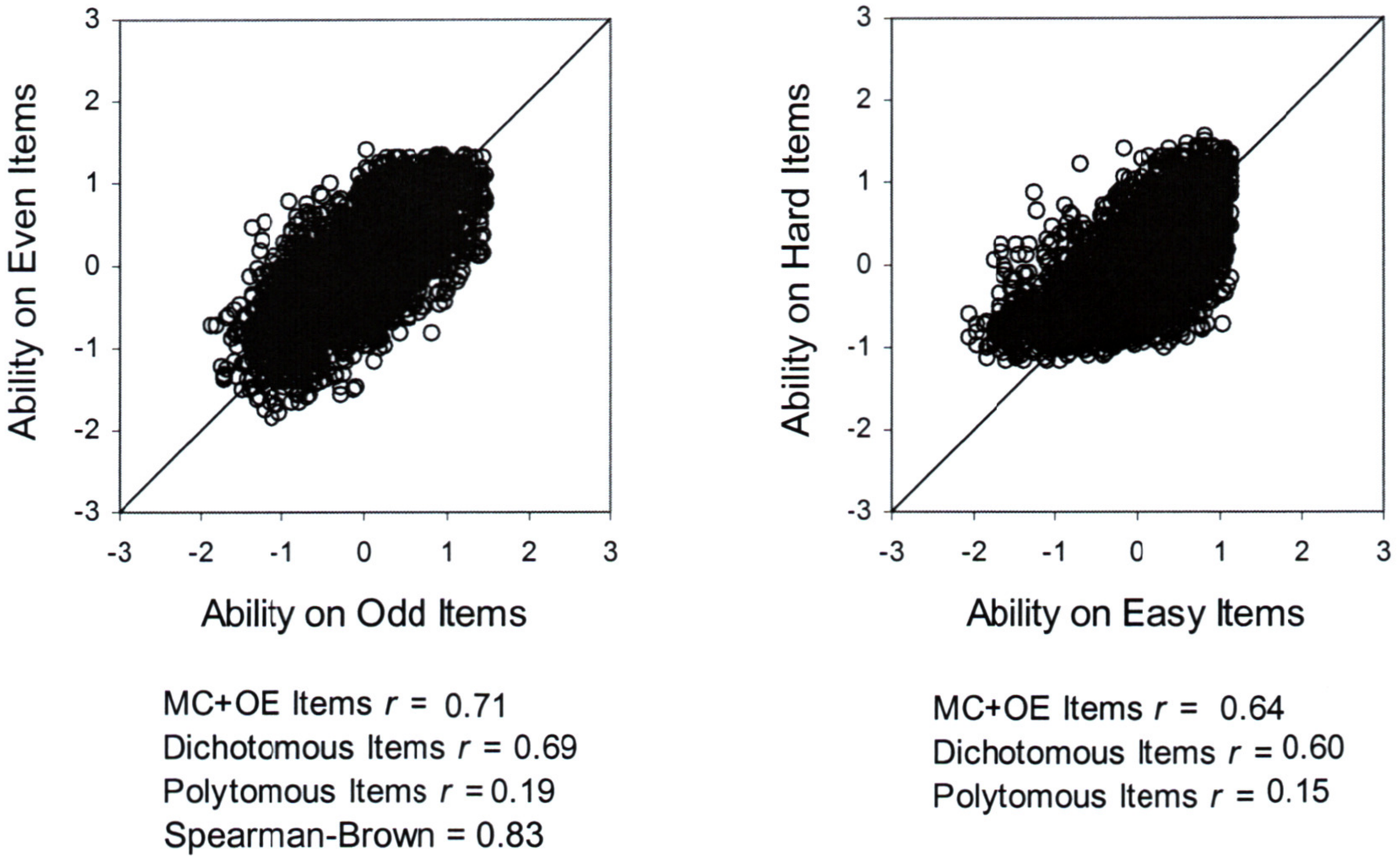



Figure 10e. Scatterplots and related correlations of numeracy ability estimates for Odd/Even and Easy/Hard Items using 3PL-GPCM.

To assist in the interpretation of IRT ability parameter invariance, classical reliability measures were calculated. As reported in Tables 5 and 6, with the exception of the Pearson correlation obtained for easy/hard reading comprehension items ($r = 0.67$), the reliability measures obtained for the total tests (i.e. comprising both dichotomous and polytomous items) were reasonably high (i.e., greater than 0.70). In general, there was a slight decrease in reliability when only dichotomous items were analysed and a significant decrease when only polytomous items were analysed. In all but one case, numeracy obtained higher reliability estimates than reading comprehension.

Table 5

Reliability Measures for Total Test and Dichotomous and Polytomous Items (Reading Comprehension)

| | Total (44 items) | Dichotomous (40 items) | Polytomous (4 items) |
|---|---|---|---|
| Full Data Set (N = 10,000) | | | |
| Cronbach's alpha | 0.81 | 0.80 | 0.57 |
| Split-half Spearman-Brown | 0.79 | 0.76 | 0.53 |
| Sample Data Sets (N = 3,000) | | | |
| Pearson *r* for odd/even items | 0.72 | 0.69 | 0.40 |
| Pearson *r* for easy/hard items | 0.67 | 0.66 | 0.36 |

Table 6

Reliability Measures for Total Test and Dichotomous and Polytomous Items (Numeracy)

| | Total (36 items) | Dichotomous (32 items) | Polytomous (4 items) |
|---|---|---|---|
| Full Data Set (N = 10,000) | | | |
| Cronbach's alpha | 0.85 | 0.85 | 0.64 |
| Split-half Spearman-Brown | 0.83 | 0.80 | 0.59 |
| Sample Data Sets (N = 3,000) | | | |
| Pearson *r* for odd/even items | 0.78 | 0.75 | 0.49 |
| Pearson *r* for easy/hard items | 0.75 | 0.66 | 0.51 |

## Model-Data Fit

The results of procedures to assess the extent to which various IRT models fit FSA 2000 reading comprehension and numeracy data are reported here. Table 7 lists the number reading comprehension items deemed to fit various IRT models across a range of sample sizes. Items were considered to have fit a model when they obtained a chi-square with $p \geq 0.01$ or a more stringent $p \geq 0.05$. Table 7 also reports overall test fit (based on all items tested under each model) at the same levels of significance. Comparable information for numeracy data is reported in Table 8.

The results of statistical tests to measure the relative fit of various IRT models are reported in Table 9 for the reading comprehension data set and Table 10 for the numeracy data set. The differences between the -2 log likelihood fit statistics obtained at the end of the parameter estimation cycle for each model are provided for different sample sizes, along with the level of statistical significance based on a chi-square distribution with the given degrees of freedom. A positive difference with $p < 0.05$ signifies significantly better model-data fit for the least restrictive model of the pair being compared. A negative difference indicates the more restrictive model fits the data better than the less restrictive model.

Table 7

Summary of Reading Comprehension Item and Test Fit by IRT Model and Sample Size

| | | Number of Fitting Items[a] | | Overall Test Fit | |
|---|---|---|---|---|---|
| IRT Model | Sample Size | $p \geq .05$ | $p \geq .01$ | $p \geq .05$ | $p \geq .01$ |
| 1PL | 500 | 28 | 36 | misfit | misfit |
| | 1500 | 15 | 17 | misfit | misfit |
| | 3000 | 8 | 11 | misfit | misfit |
| | 10000 | 2 | 2 | misfit | misfit |
| 2PL | 500 | 34 | 36 | fit | fit |
| | 1500 | 32 | 35 | misfit | misfit |
| | 3000 | 20 | 26 | misfit | misfit |
| | 10000 | 10 | 13 | misfit | misfit |
| 3PL | 500 | n/a[b] | n/a[b] | n/a[b] | n/a[b] |
| | 1500 | n/a[b] | n/a[b] | n/a[b] | n/a[b] |
| | 3000 | 23 | 28 | misfit | misfit |
| | 10000 | 0 | 0 | misfit | misfit |
| PCM | 500 | 0 | 0 | misfit | misfit |
| | 1500 | 0 | 0 | misfit | misfit |
| | 3000 | 0 | 0 | misfit | misfit |
| | 10000 | 0 | 0 | misfit | misfit |
| GPCM | 500 | 0 | 0 | misfit | misfit |
| | 1500 | 0 | 0 | misfit | misfit |
| | 3000 | 0 | 0 | misfit | misfit |
| | 10000 | 0 | 0 | misfit | misfit |

[a]The 1PL, 2PL, and 3PL analyses are based on 40 dichotomous items. The PCM and GPCM analyses are based on 4 polytomous items.

[b]Test data did not converge with 3PL model for sample sizes of 500 and 1500.

Table 8

Summary of Numeracy Item and Test Fit by IRT Model and Sample Size

| | | Number of Fitting Items[a] | | Overall Test Fit | |
|---|---|---|---|---|---|
| IRT Model | Sample Size | $p \geq .05$ | $p \geq .01$ | $p \geq .05$ | $p \geq .01$ |
| 1PL | 500 | 29 | 31 | misfit | fit |
| | 1500 | 19 | 22 | misfit | misfit |
| | 3000 | 13 | 19 | misfit | misfit |
| | 10000 | 3 | 5 | misfit | misfit |
| 2PL | 500 | 30 | 31 | fit | fit |
| | 1500 | 28 | 29 | misfit | misfit |
| | 3000 | 19 | 24 | misfit | misfit |
| | 10000 | 4 | 11 | misfit | misfit |
| 3PL | 500 | 29 | 31 | misfit | misfit |
| | 1500 | 26 | 32 | misfit | fit |
| | 3000 | 3 | 3 | misfit | misfit |
| | 10000 | 0 | 0 | misfit | misfit |
| PCM | 500 | 0 | 0 | misfit | misfit |
| | 1500 | 0 | 0 | misfit | misfit |
| | 3000 | 0 | 0 | misfit | misfit |
| | 10000 | 0 | 0 | misfit | misfit |
| GPCM | 500 | 0 | 0 | misfit | misfit |
| | 1500 | 0 | 0 | misfit | misfit |
| | 3000 | 0 | 0 | misfit | misfit |
| | 10000 | 0 | 0 | misfit | misfit |

[a]The 1PL, 2PL, and 3PL analyses are based on 32 dichotomous items. The PCM and GPCM analyses are based on 4 polytomous items.

Table 9

Differences in Fit Statistics Between Pairs of IRT Models for Reading Comprehension Data Sets of Various Sample Sizes

| Models Compared | N | d.f. | Difference | Significance |
|---|---|---|---|---|
| 1PL-2PL | 500 | 40 | 271.13 | $p < .001$ |
| | 1500 | 40 | 797.19 | $p < .001$ |
| | 3000 | 40 | 1484.89 | $p < .001$ |
| | 10000 | 40 | 4682.84 | $p < .001$ |
| 2PL-3PL | 500 | 40 | n/a[a] | n/a[a] |
| | 1500 | 40 | n/a[a] | n/a[a] |
| | 3000 | 40 | -614.18[b] | $p < .001$ |
| | 10000 | 40 | -33887.02[b] | $p < .001$ |
| PCM-GPCM | 500 | 4 | 11.12 | $p < .05$ |
| | 1500 | 4 | 18.55 | $p < .001$ |
| | 3000 | 4 | 73.65 | $p < .001$ |
| | 10000 | 4 | 253.36 | $p < .001$ |

[a]Test data did not converge with 3PL model for sample sizes of 500 and 1500.

[b]A negative difference indicates the -2 log likelihood fit statistic of the more restrictive model is smaller (meaning better fit) than that of the less restrictive model.

Table 10

Differences in Fit Statistics Between Pairs of IRT Models for Numeracy Data Sets of Various Sample Sizes

| Models Compared | N | d.f. | Difference | Significance |
|---|---|---|---|---|
| 1PL-2PL | 500 | 32 | 105.16 | $p < .001$ |
| | 1500 | 32 | 267.31 | $p < .001$ |
| | 3000 | 32 | 451.72 | $p < .001$ |
| | 10000 | 32 | 1503.73 | $p < .001$ |
| 2PL-3PL | 500 | 32 | -263.79[a] | $p < .001$ |
| | 1500 | 32 | 109.28 | $p < .001$ |
| | 3000 | 32 | -577.40[a] | $p < .001$ |
| | 10000 | 32 | -703.29[a] | $p < .001$ |
| PCM-GPCM | 500 | 4 | 19.28 | $p < .001$ |
| | 1500 | 4 | 28.54 | $p < .001$ |
| | 3000 | 4 | 114.87 | $p < .001$ |
| | 10000 | 4 | 383.53 | $p < .001$ |

[a]A negative difference indicates the -2 log likelihood fit statistic of the more restrictive model is smaller (meaning better fit) than that of the less restrictive model.

# CHAPTER 5

# SUMMARY, DISCUSSION, AND CONCLUSION

## Summary

As part of its provincial learning assessment program, the BC Ministry of Education has, since 1976, regularly assessed the extent to which students are attaining the knowledge and skills expected at various levels of the Kindergarten to Grade 12 education system. In 1999, the ministry introduced a new assessment called the Foundation Skills Assessment (FSA). This assessment differed from previous provincial assessments in several ways, including its content (the cross-curricular skills of reading comprehension, writing, and numeracy versus curriculum-specific subject areas), its administrative cycle (annual versus every 4 or 5 years), and its level and method of reporting (reporting provincial, district, school, and, starting in 2000, individual student results in relation to provincial standards of performance versus average scores reported at the provincial, district, and school levels only).

Given the unique features of the FSA, a number of significant practical and technical challenges were raised. As FSA tests are composed primarily of different questions from one year to the next, steps will need to be taken to ensure results on different FSA tests can be meaningfully compared over time. The importance of comparability is further increased with the reporting of results in relation to provincial standards, which convey a sense of consistency in measures. Finally, the provision of individual results, and the increased impact these results may have on individual students'

learning, underscores an increased responsibility to ensure that individual results are calculated in the most accurate manner possible.

As means to address the challenges brought forth by FSA 2000, this study explored the feasibility of using a modern measurement approach known as item response theory (IRT) to analyse FSA data. Specifically, the purposes of the study were twofold: first, to determine the extent to which data from the Grade 7 FSA 2000 tests meet the assumptions and properties of IRT and, second, to determine which of several popular item response models best fit Grade 7 FSA 2000 data. The assumptions and properties investigated included unidimensionality, local independence, nonspeededness, possible equal item discrimination indices, possible lower asymptote, item and ability parameter invariance, and model-data fit for the 1PL, 2PL, 3PL, PCM, and GPCM IRT models.

## Discussion

An overview and interpretation of the key findings are outlined below. To facilitate future studies on the feasibility of using IRT with FSA and other large-scale assessments, relevant limitations are also raised along with suggestions for improvement and future study.

### Unidimensionality

Investigations of unidimensionality, a common assumption of all five IRT models investigated in this study, offered mixed results both within and across procedures.

Procedures based on principal components analysis provided some evidence to suggest the numeracy data set was unidimensional (i.e., the calculation of one large

eigenvalue that was 5 times larger than the second largest eigenvalue), but also offered evidence suggesting the presence of more than one dominant component (the first component of the numeracy test only accounted for 18.4% of the total test variance). Comparable factor analytic studies for reading comprehension were more consistent, rejecting the presence of unidimensionality in both cases.

It was hoped a more sophisticated approach to assess for unidimensionality using the software package Poly-DIMTEST would offer further insight into the dimensionality of the data sets. Unfortunately, the results from Poly-DIMTEST were also mixed, and therefore the assumption of unidimensionality could not be unequivocally supported. Four different analyses were conducted in Poly-DIMTEST using different homogeneous item groups from each data set (described in Table 2). In marked contrast to the results obtained for factor analysis, three out of four Poly-DIMTEST analyses gave results that suggested the reading comprehension data set was essentially unidimensional. One Poly-DIMTEST analysis, however, suggested the data set was multidimensional. For numeracy, the Poly-DIMTEST results were evenly split, with two analyses resulting in a unidimensional finding and two a multidimensional finding. A number of researchers (e.g., Linacre, 1994; Wang, 1994) have reported that the results of a precursor to Poly-DIMTEST (DIMTEST) are dependent upon the subset of items on which the test for unidimensionality is based. The findings of this study support this position and generalize it to the more recent Poly-DIMTEST.

The lack of triangulation across, and indeed within, the various approaches used in this study to assess unidimensionality is something that has been reported in other studies (e.g., Boughton, 1998; Hattie et al. 1996; Linacre, 1994; Wang, 1994) and points

to limitations in one or all of the approaches used to investigate this key assumption. Other methods to test for unidimensionality, for example non-linear factor analysis and residual analyses, have shown promise (Hambleton & Rovinelli, 1986; Hattie, 1985) and may provide more consistent results. However, at this time no single index has been identified that will be generally accepted. Consequently, the conflicting results obtained in this study are likely something that all researchers must deal with until an accepted and proven index is established. The results should also raise caution to anyone wishing to make a strong assumption based on only one measure of unidimensionality.

## Local Independence

Another assumption of IRT assessed in this study was local independence, which means that, after the latent variable underlying performance is taken into account, examinee responses to items on a test are statistically independent (Kolen & Brennan, 1995, p. 158). The investigation into the assumption of local independence showed relatively few (0.7% for reading comprehension and 1.3% for numeracy) pairs of items demonstrated dependence (defined as a value of 0.10 or higher on a residual correlation matrix) once the first factor was removed. Overall these findings provide general support for local independence in both data sets.

In the limited instances where local dependence was observed between a pair of items, a review of the actual test questions provided useful insight into possible reasons for item dependence. Consistent with three causes outlined by Yen (1993), the few item pairs demonstrating dependence were generally linked to the same passage or setting, assessed a very specific skill within the larger domain area, or shared a common response

format (e.g., written response). Given IRT's reliance on local independence, the procedures used in this study could be of use to FSA test developers in the identification and modification of items demonstrating excessive local dependence during field testing, thus minimizing the chance that problematic items are included in the final test instrument.

## Nonspeededness

Another common assumption of IRT models investigated in this study was that students had sufficient time to respond to all items in the test. Assuming items at the end of each test would be most affected by any potential time constraints, the percentage of students responding to last 5 questions of each test were reviewed. This review showed that only one item was attempted by less than 90% of the examinees and that the response rate of the outlying item was a respectable 89%. Overall, these results suggest the reading comprehension and numeracy items comprising Grade 7 FSA 2000 tests were nonspeeded. They also provide evidence that the efforts made to address timing problems encountered with FSA 1999 were successful.

The approach taken to assess speededness in this study is not without faults. It is possible that students were pressured to answer all items on the test in the time allotted and simply guessed randomly on items for which they did not have time to meaningfully respond. To complement these results, additional post-administration analyses could be carried out on the data from FSA 2000 or future administrations of the assessment (e.g., although the results of this study emphasize the need for caution, Poly-DIMTEST could be used to assess whether the final items on a test are dimensionally distinct from the

remainder of the test [Stout, Douglas, Junker, & Roussos, 1993, p. 2]). However, judging by the results obtained, it would be reasonable to assume procedures carried out by the test developers either before the administration of FSA 2000 (i.e., the administration of timing pilots to a random group of students prior to the administration of the assessment) or during it (i.e., the provision of additional time to students as required) are sufficient to minimize the impact speed may have on student performance.

## Equal Item Discrimination

A review of the point-biserial and point-polyserial item-test score correlations showed the assumption of equal item discrimination was tenable for the multiple-choice and open-ended components of the numeracy test (only 3.1% of the multiple-choice items and no open-ended item fell outside a reasonable range) and the open-ended component of the reading comprehension test (although, admittedly, the small number of open-ended items limit the power of this test). With 37.5% of its items falling outside the average item-test score correlation ± 0.15, the multiple-choice component of the reading comprehension test appears to not satisfy the 1PL model's assumption of equal item discrimination.

As will be discussed later, analyses evaluating the relative fit of the one parameter and two parameter item response models showed both data sets could be more effectively represented using two parameters (difficulty *and* item discrimination). Given the only difference between the one and two parameter models is the latter's acceptance of unequal item discrimination, the results of the more sophisticated IRT analyses do not support an assumption of equal item discrimination for either FSA data set. The disparate

results between CTT- and IRT-based analyses for testing the equal item discrimination assumption provide further caution against using a single measure to test IRT assumptions.

## Lower Asymptote

An investigation into the performance of the lowest decile of students on the five most difficult multiple-choice items from each test provided evidence for a non-zero lower asymptote. The proportion of low-scoring students who correctly answered these difficult items (ranging from 0.10 to 0.32, with most proportions ranging from 0.20 to 0.30) was in line with the initial criteria of 0.25 used to demonstrate the presence of guessing. The findings also supported Lord's (1980, p. 17) assertion that guessing will fluctuate around random probability given the presence of attractive multiple choice alternatives (which will reduce the probability of correct response below chance) and students' partial knowledge (which will increase the probability of correct response beyond chance).

The procedure used to assess the lower asymptote of each data set is not without faults. As Hambleton and Swaminathan (1985, p. 161) point out, the results could be misleading if the students in the lowest decile were only low in ability in relation to other examinees but not low ability in any absolute sense. Given that FSA assesses knowledge and skills that all students are expected to be taught as part of the prescribed BC curriculum, it is reasonable to expect that the students in the lowest decile are not void of any knowledge and skill in the domains tested.

To help address this likely fault in the lower asymptote investigation, the relative fit of two models that differ only in their assumption of guessing was reviewed. As will be described later, these results showed that the 2PL model (which assumes a lower asymptote of zero) obtained better relative model-data fit than the 3PL model which allows for a non-zero lower asymptote. This suggests that assuming a lower asymptote of zero may be reasonable when applying IRT to Grade 7 FSA 2000 data. Given the conflicting findings outlined earlier, however, further investigation is warranted to better determine the true nature of the lower asymptote on FSA data (e.g., better fit might be obtained if the lower asymptote was fixed at some reasonable non-zero value rather than varying across items).

## Item and Ability Parameter Invariance

A cornerstone of item response theory, and its major distinction from classical test theory, is the property of parameter invariance (Hambleton et al., 1991, p. 18). Through this property IRT models can produce item parameters that are independent of examinee populations and ability parameters that are independent of the set of items administered to examinees. This study investigated item invariance across various IRT models by correlating *b*-parameter estimates of test items obtained from different groups of students (i.e., two randomly selected groups and two groups of students with high or low scores on the test). Ability invariance was assessed by correlating ability estimates obtained from different groups of items (either a random sample of items or samples comprising easy or difficult items) under different IRT models.

A number of difficulties were experienced during the parameter invariance investigation. These included problems running analyses under a 1PL or PCM using a current version of PARSCALE (version 3.2), difficulties obtaining convergence with extreme, and sometimes random, groupings of examinees and items, and limitations in calculating the 3PL model in PARSCALE versus another IRT program (BILOG 3.11). A number of solutions to these difficulties were discussed.

## Item Parameter Invariance

The results of this study showed that the reading comprehension and numeracy data sets did not demonstrate item parameter invariance across all groups.

As should be expected given the equivalent nature of the two groups, a strong linear relationship of item *b*-parameters was noted across random groups of students for both data sets (with the exception of the *b*-parameter for one item in the numeracy data set estimated under a 3PL model). In all cases, near perfect Pearson product-moment correlations of 0.98 or higher were obtained, suggesting item parameter invariance.

Unfortunately, the positive results for parameter invariance across random groups did not materialize across low and high scoring groups of students. With the exception of the 1PL-PCM model in numeracy (whose scatterplot was approximately the same as the randomly equivalent groups, given a linear transformation), all *b*-parameter scatterplots of low/high scoring students showed moderate to low linear relation. The correlation calculated for numeracy item difficulty estimates of the 1PL-PCM was a very respectable 0.91 for low/high scoring students. Across other models, however, most item difficulty correlations for the full test were very low. Of these, the correlations obtained for reading

comprehension items under the 2PL-PCM ($r = 0.04$), 3PL-PCM ($r = 0.13$), 3PL-GPCM ($r = 0.24$) stand out as being particularly problematic, as does the lack of convergence for the low scoring group when the numeracy data was analysed using a 3PL model.

One unintended but interesting finding from the study can be gleaned from the correlations of the item difficulty parameters for the reading comprehension polytomous items in Figures 7a and 7b. Due to problems with the estimation of polytomous items using the PCM, it was very difficult to obtain model-data convergence for low and high scoring groups using a 1PL-PCM model on this data set (Figure 7a). Repeated attempts at estimating the dichotomous and polytomous items separately using PARSCALE and combining them at the end of the estimation process were fruitless. Convergence was obtained, however, if the all items were estimated concurrently as one test (rather than two subtests). This estimation process resulted in a very high correlation for the polytomous items ($r = 0.97$). To enable a successful run of the same groups of student using a 2PL-PCM, polytomous item threshold values obtained from the "one test" 1PL-PCM analyses were entered as starting threshold values for PCM component of the "two test" 2PL-PCM command file in PARSCALE. As reported in Figure 7b, the resulting correlation for the polytomous items decreased to 0.67. This difference provides evidence to suggest parameter invariance may be dependent upon how parameters are estimated (i.e., jointly with other types of items as part of "one test" or separately as a sub-test of a "two test" analysis). To facilitate parameter invariance studies and help deal with convergence problems in data sets, this unexpected finding would be a worthy topic for future research.

Ability Parameter Invariance

The property of ability parameter invariance was assessed by comparing the ability parameter estimates obtained across two randomly equivalent sets of items and two sets of items of different levels of difficulty. The results of this study provide strong evidence that ability parameter invariance was not demonstrated for the Grade 7 reading comprehension and numeracy data sets of FSA 2000. All scatterplots, regardless of which item groupings (odd/even or easy/hard) or IRT model they depicted, demonstrated large amounts of scatter and moderate to low linear relationships. The Pearson product-moment correlations of the IRT ability estimates obtained for the full set of items ranged from 0.55 to 0.70 for reading comprehension and 0.64 to 0.71 for numeracy. These were less than comparable correlations calculated using CTT-based test scores (ranging from 0.67 to 0.72 for reading comprehension and 0.75 to 0.78 for numeracy). This may suggest IRT ability estimates are not as consistent as CTT-based test scores across different subsets of items. However such a view may be countered with the Spearman-Brown estimates obtained for the full tests, which suggest IRT ability estimates demonstrate equal or slightly higher amounts of internal consistency than their CTT counterparts (0.82 for IRT ability estimates versus 0.79 for CTT test scores on the reading comprehension test and 0.83 for IRT and CTT measures on the numeracy test).

The presence of negative correlations across polytomous reading comprehension items ($r$ = -0.35) under the PCM and very low correlations across polytomous numeracy items ($r$ = 0.15) under the GPCM are particularly concerning, even with the limited number of polytomous items involved in the study. These findings suggest student ability

estimates would vary dramatically depending on which polytomous items were used to make the IRT ability estimations. Such extreme correlations were not obtained under the CTT reliability analyses, connoting problems when applying IRT to the polytomous Grade 7 FSA data. In addition to raising general caution for the application of IRT to FSA data, these findings raise an important practical consideration when attempts are made to equate different tests. Items demonstrating such low or negative correlations should not be used as anchor items to equate different FSA tests. Doing so would result in unstable and unreliable parameter estimates.

## Model-Data Fit

A critical part of any successful application of IRT is ensuring that an appropriate model is used to characterize the data. The various procedures already discussed provide useful insight into the feasibility of using various IRT models to represent Grade 7 FSA 2000 data. Two other methods were also used to help determine which of several IRT models best fit Grade 7 FSA 2000 data. Both of these methods made use of the chi-square statistic. To help control for the sensitivity of the chi-square statistic to large sample sizes, analyses were carried out on a range of sample sizes (500, 1,500, 3,000, and 10,000).

Consistent with the findings of Hambleton and Murray (1983), the number of FSA 2000 items judged to fit an IRT model at a given level of significance ($\alpha = 0.05$ or $\alpha = 0.01$) generally decreased as sample size increased (see Tables 4 and 5). The only exceptions to this general finding were for the PCM and GPCM (where no items fit the model at any sample size) and the 3PL analysis of the numeracy test with a sample of

1,500 (where the number of fitting items increased from the sample of 500). This general finding underscores the need to not rely on a single statistical index when judging the appropriateness of an IRT model to a data set (particularly when a large sample is involved). Sole reliance on chi-square, the default test statistic in PARSCALE, could result in the mistaken rejection of an IRT model simply due to the size of the sample being analysed.

## Relative Fit of Dichotomous IRT Models

Tables 4 and 5 also show that, in most cases, the 2PL model has a larger number of fitting items than the 1PL or 3PL models across different sample sizes, suggesting that the 2PL has a better relative fit than the other dichotomous IRT models. Additional support for the higher relative fit of the 2PL model to Grade 7 FSA 2000 data is provided by the finding that the 2PL was the only IRT model to achieve overall test fit at $p \geq 0.05$ for any of the four sample sizes and that it was also the only model that demonstrated at least one instance of overall test fit for both reading comprehension and numeracy.

A second, and less subjective procedure, used to judge the relative fit of different dichotomous IRT models involved the difference of the -2 log likelihood values obtained at the end of a successful estimation cycle. As displayed in Tables 6 and 7, in all but one case, the 2PL obtained significantly better fit than the 1PL and 3PL models across the two data sets and various sample sizes.

Given the 3PL imposes less restrictions than the 2PL, the results that suggested the 2PL fit Grade 7 FSA 2000 dichotomous data better than the 3PL were surprising. However, as Harris (1989) suggests, and was demonstrated in the difficulties experienced

during the invariance investigations of this study, difficulties in estimating the *c* parameter in real data sets may make the 2PL the preferred model for applications to the FSA. However, analyses of the results from the parameter invariance investigation does not endorse this view. With possible exception to the 1PL in numeracy, no IRT model stood out as fitting Grade 7 FSA 2000 data better than the rest. All demonstrated poor parameter invariance, which, as argued by Hambleton et al. (1991, p. 24), suggests poor model-data fit. These results may have some relation to the dimensionality findings outlined earlier, with one possible explanation being that the FSA data sets cannot be adequately represented as a whole using unidimensional IRT models.

Relative Fit of Polytomous IRT Models

As shown in Tables 4 and 5, the PCM and GPCM IRT models did not fit any polytomous items from the reading comprehension and numeracy Grade 7 FSA 2000 data sets. Not surprisingly given the item-level results, overall test fit was also not obtained for either model. The limited number of items involved in both cases (4), or limitations to the chi-square fit statistic provided in PARSCALE, may have contributed to these poor results. More sophisticated analyses, including the inspection of graphic displays of the response functions for each item (Muraki, 1993) and analysis of standardized residuals (Embretson & Reise, 2000, p. 237-238), may have shed better light on the suitability of polytomous IRT models to the Grade 7 FSA 2000 data. A review of the correlations calculated as part of the invariance investigations showed good item parameter invariance at times for the PCM but extremely poor ability parameter invariance (e.g., the negative correlation listed in Figure 9d). Here again, poor parameter estimates may be a result of

the very small number of polytomous items. Consequently, one step the Ministry of Education may wish to explore to improve the parameter estimation process on the polytomous data is adding additional items to the open-ended component of each test. Of course, any advantages this action might have should be weighed against practical disadvantages that would arise (e.g., the additional time required to administer and mark the tests).

The investigation of relative fit between the PCM and GPCM using the -2 log likelihood statistic provided statistical evidence to suggest the GPCM fit the polytomous reading comprehension and numeracy data sets better than the PCM. These results were statistically significant at $p < 0.001$ across all sample sizes but one (the sample of 500 students was significant at $p < 0.05$ for reading comprehension). However, given the problems with parameter invariance and low sample size, further research is warranted to determine the applicability of either polytomous IRT model to FSA data.

## Conclusion

The purposes of this study were to determine the extent to which data from the Grade 7 FSA 2000 tests meet the assumptions and properties relevant to IRT and which of several popular IRT models best fit Grade 7 FSA 2000 data. The investigations carried out scrutinized FSA data in a manner that has not been done before for any measurement model (e.g., CTT or IRT).

At the outset, it was determined that one of the skill areas tested by the FSA (writing) did not have a sufficient number of items to allow for IRT analyses. At the very minimum, if IRT is to be applied to all components of the FSA, the number of items in

each test must be greater than the number of traits assessed (Crocker & Algina, 1986, p. 339). The results of this study were mixed for the other two skill areas assessed by FSA (reading comprehension and numeracy). The investigations carried in this study showed a couple of important assumptions (local independence and nonspeededness) were met by the reading comprehension and numeracy data sets but other critical assumptions or expected properties (unidimensionality and parameter invariance) were not.

Investigations into relative fit suggested that the 2PL model fit the dichotomous data from the Grade 7 reading comprehension and numeracy tests of FSA 2000 better than the 1PL or 3PL. For polytomous data, the GPCM was judged to have better model-data fit than the PCM. However, the disappointing findings of the parameter invariance investigations undertaken as part of this study, as well as the limited number of polytomous items involved in each test, makes this relative fit of little consequence.

Given one of the purposes of FSA is to allow schools, districts, and the province to monitor change over time, the lack of parameter invariance demonstrated in this study stands out as being very problematic. To adequately compare the performance of different students on different tests, item and ability invariance must be demonstrated. The Grade 7 reading comprehension and numeracy data clearly did not demonstrate this condition.

While discouraging, the results of this study should not be viewed as damning for the use of IRT to analyse FSA data. The results should be considered to be a first critical look into the application of IRT to FSA data. Given this study focussed on only one of three grades tested in FSA 2000, its results must be considered to be limited in scope. Further investigations across all grades and administrations are warranted. Especially if

confirmed at the other grade levels, these results suggest steps should be taken to maximize model-data fit for each FSA test—preferably prior to administration. Hambleton et al. [1991, chap. 7] provide a good overview of how IRT can be used to improve model-data fit in the test construction process. Alternative procedures that could be explored to improve fit after administration include treating each test as being composed of a number of unidimensional subtests, rather than as one large unidimensional test or, if deemed appropriate, deleting specific items or examinees that do not fit the unidimensional model.

In theory, IRT shows significant promise to help the FSA achieve its goal of providing annual information on important foundation skills that can be used by BC's education system to evaluate existing programs and make plans for improvement. If an IRT model fits the data, it should provide an accurate way to estimate student ability. As the findings of this study indicate, important assumptions and properties underlying IRT may not hold in practice. Consequently, to help minimize any discrepancies between theory and practice, it is recommended that regular and varied investigations into the extent to which FSA data aligns with IRT should be carried out and incorporated into the test development and follow-through procedures.

# REFERENCES

Anderson, J. O. (1997). Comparison of various approaches to item analysis. Unpublished manuscript, University of Victoria, Victoria, BC.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. Journal of Educational Measurement, 28, 147-162.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 1-23). Vancouver, BC: Educational Research Institute of British Columbia.

Boughton, K. A. (1998). Item response theory: The application of both dichotomous and polytomous item response models to a provincial exam data set. Unpublished master's thesis. University of Victoria, Victoria, British Columbia, Canada.

Choi, I.C. & Bachman, L.F. (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. Language Testing, 9(1), 51-78.

Crawford, J. R., & Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. Journal of Clinical and Experimental Neuropsychology, 20(5), 755-762.

Crocker, L., & Algina. J. (1986). Introduction to classical and modern test theory. Orlando, FL: Harcourt Brace Jovanovich.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahway, NJ: Lawrence Erlbaum Associates.

EXCEL 2000 [Computer software]. (1999). Redmond, WA: Microsoft.

Fan, X., & Ping, Y. (1999, April). Assessing the Effect of Model-Data Misfit on the Invariance Property of IRT Parameter Estimates. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational measurement (3rd ed) (pp. 147-200). New York: Macmillan.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice,12(3), 38-47.

Hambleton, R. K., & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 71-94). Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10(3), 287-302.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Norwell, MA: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. Educational Measurement: Issues and Practice, 8, 35-41.

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

Hattie, J. (1985). Methodological review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9(2), 139-164.

Hattie, J., Krakowski, K., Rogers, H. J., Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. Applied Psychological Measurement, 20(1), 1-14.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Item response theory: Applications to Psychological measurement., Homewood: Ill.: Dorsey.

Keeves, J. P. (1994). National examinations: design, procedures and reporting. Paris: UNESCO.

Klinger, D. A. (2000). Recovery of decision consistency in shortened and modified scholarship examinations. Unpublished doctoral dissertation, University of Alberta.

Kolen, M. J., & Brennan, R. L. (1995). Test equating: Methods and practices. New York: Springer.

Lane, S., Stone, C. A., Ankenmann, R.D., & Liu, M. (1995). Examination of the assumptions and properties of the graded item response model: An example using a mathematics performance assessment. Applied Measurement In Education, 8(4), 313-340.

Li, H. H., & Stout, W. (1995, November). A version of DIMTEST to assess latent trait unidimensionality for mixed polytomous and dichotomous item response data. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Linacre, J. (1994, Autumn). A DIMTEST diminuendo. Rasch Measurement Transactions [On-line serial], 8(3). Available http://ericae.net/ft/rasch/R8313.HTM.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Masters, G. N. (1982) A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp.101-121). New York: Springer-Verlag Inc.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahway, New Jersey: Lawrence Erlbaum Associates.

Ministry of Education. (1999). Foundation Skills Assessment highlights 1999. Victoria, BC: Author.

Ministry of Education. (2000a). Interpreting and communicating British Columbia Foundation Skills Assessment results. Victoria, BC: Author.

Ministry of Education. (2000b). Foundations Skills Assessment 2000: Instructions for teachers/invigilators. Victoria, BC: Author.

Ministry of Education. (2000c). Foundation Skills Assessment 2000: Provincial results report. Victoria, BC: Author.

Mislevy, R. J., & Bock, R. D. (1998). BILOG 3.11 [Computer software]. Mooresville, IN: Scientific Software International.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. Applied Psychological Measurement, 17, 351-363.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 153-168). New York, NY: Springer.

Muraki, E., & Bock, R. D. (1997). PARSCALE (Version 3.2) [Computer software manual]. Chicago, IL: Scientific Software International, Inc.

Muraki, E., & Bock, R. D. (in press). PARSCALE (Version 3.5) [Computer software]. Chicago, IL: Scientific Software International, Inc.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses: Comparison of different approaches. Journal of Educational Measurement, 31(1), 17-35.

Nandakumar, R., Yu, F., Li, H. H., & Stout, W. (1998). Assessing unidimensionality of polytomous data. Applied Psychological Measurement, 22(2), 99-115.

Nunnally, J.C., & Bernstein, I.H. (1994). Psychometric theory. New York: McGraw-Hill.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. Journal of Educational Measurement; 31(3), 200-219.

Poly-DIMTEST [Computer software]. (1999). Champaign, IL: William Stout Institute for Measurement.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Rogers, H. J., & Hattie, J. A. (1987). A monte carlo evaluation of several person and item fit statistics in latent trait models. Applied Psychological Measurement, 11(1), 47-57.

S.P.S.S. for Windows 10.0 , Statistical package for the social sciences [Computer software]. (1999). Chicago, IL: SPSS Inc.

Stout, W. F. (1987). A nonparametric approach to assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Tabachnick, B. G., & Fidell, L. S. (1996). Using multivariate statistics (3rd ed.). New York: HarperCollins.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Van der Linden, W.J., & Hambleton, R.K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 1-28). New York, NY: Springer.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration and proficiency estimation. In H. Wainer (Ed.), Computer adaptive testing: A primer (pp. 65-102). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, Y. C. L. (1994, April). Robustness of unidimensional IRT calibration in the presence of essential unidimensionality. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wright, B. D. (1999). Fundamental measurement for psychology. In S.E. Embretson & S.L. Hershberger (Eds.), The new rules of measurement: What every psychologist and educator should know (pp. 65-104). Mahwah, NJ: Lawrence Erlbaum Associates.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30(3), 187-213.

# APPENDIX A

## Reading Comprehension Residual Correlation Matrix with First Factor Removed[a].

| | MC1 | MC2 | MC3 | MC4 | |
|---|---|---|---|---|---|
| MC1 | | 0.025 | 0.044 | 0.020 | |
| MC2 | 0.025 | | 0.027 | 0.010 | |
| MC3 | 0.044 | 0.027 | | 0.026 | |
| MC4 | 0.020 | 0.010 | 0.026 | | |
| MC5 | -0.015 | 0.022 | -0.017 | 0.079 | MR |
| MC6 | 0.006 | 0.003 | 0.004 | 0.008 | |
| MC7 | 0.013 | 0.020 | 0.001 | 0.004 | |
| MC8 | -0.014 | 0.000 | -0.004 | -0.006 | |
| MC9 | -0.019 | 0.015 | -0.007 | -0.004 | |
| MC10 | 0.007 | 0.013 | 0.003 | -0.011 | |
| MC11 | -0.003 | 0.001 | -0.006 | -0.011 | |
| MC12 | 0.013 | 0.026 | 0.006 | -0.015 | |
| MC13 | 0.017 | 0.007 | 0.006 | 0.006 | |
| MC14 | 0.020 | 0.028 | 0.018 | 0.006 | |
| MC15 | 0.007 | -0.002 | 0.002 | 0.010 | |
| MC16 | 0.023 | 0.021 | -0.001 | 0.017 | |
| MC17 | 0.031 | 0.007 | -0.004 | 0.002 | |
| MC18 | 0.002 | 0.004 | -0.007 | -0.007 | |
| MC19 | 0.006 | -0.001 | -0.011 | 0.018 | |
| MC20 | 0.003 | 0.004 | 0.004 | 0.001 | |
| MC21 | -0.021 | 0.015 | -0.019 | -0.005 | |
| MC22 | 0.019 | 0.002 | -0.006 | 0.008 | |
| MC23 | -0.015 | -0.008 | -0.001 | 0.020 | |
| MC24 | -0.007 | -0.005 | -0.014 | 0.016 | |
| MC25 | -0.007 | -0.008 | 0.008 | 0.006 | |
| MC26 | -0.007 | -0.008 | 0.006 | -0.007 | |
| MC27 | 0.002 | -0.007 | 0.007 | -0.011 | |
| MC28 | 0.009 | 0.002 | 0.009 | -0.007 | |
| MC29 | 0.008 | 0.010 | -0.008 | 0.004 | |
| MC30 | -0.019 | -0.020 | 0.013 | -0.027 | |
| MC31 | -0.034 | 0.000 | -0.010 | -0.013 | |
| MC32 | -0.005 | -0.003 | -0.007 | 0.000 | |
| MC33 | -0.008 | -0.003 | -0.014 | -0.002 | |
| MC34 | 0.003 | -0.026 | -0.006 | 0.000 | |
| MC35 | -0.008 | -0.006 | -0.007 | 0.000 | |
| MC36 | -0.014 | -0.024 | -0.010 | -0.015 | |
| MC37 | -0.018 | 0.003 | 0.022 | -0.001 | |
| MC38 | -0.020 | -0.034 | 0.007 | 0.001 | |
| OE1_1M | -0.006 | -0.002 | -0.004 | 0.002 | |
| OE4_1M | -0.004 | -0.020 | -0.005 | -0.006 | |
| OE2_4M | 0.002 | -0.007 | 0.022 | -0.003 | |
| OE3_4M | -0.014 | 0.000 | -0.014 | -0.021 | |
| OE5_4M | 0.004 | -0.005 | -0.011 | -0.013 | |
| OE6_4M | -0.014 | -0.020 | -0.037 | -0.011 | |
| Number of moderate residuals (MR) by item. | 0 | 0 | 0 | 1 | |
| Number of large residuals (LR) by item. | 0 | 0 | 0 | 0 | |

Note. Across 946 unique pairings, there are 26 (2.7%) moderate residuals with absolute values between .05 and .10 (denoted by "MR") and 7 (0.7%) large residuals with absolute values greater than or equal to .1 (denoted by "LR").

[a]Residuals represent the difference between the observed and reproduced correlational matrices.

| | MC5 | MC6 | MC7 | MC8 | MC9 | MC10 | MC11 |
|---|---|---|---|---|---|---|---|
| MC1 | -0.015 | 0.006 | 0.013 | -0.014 | -0.019 | 0.007 | -0.003 |
| MC2 | 0.022 | 0.003 | 0.020 | 0.000 | 0.015 | 0.013 | 0.001 |
| MC3 | -0.017 | 0.004 | 0.001 | -0.004 | -0.007 | 0.003 | -0.006 |
| MC4 | 0.079 MR | 0.008 | 0.004 | -0.006 | -0.004 | -0.011 | -0.011 |
| MC5 | | 0.004 | -0.004 | -0.008 | 0.023 | -0.018 | 0.014 |
| MC6 | 0.004 | | 0.047 | 0.018 | 0.037 | 0.003 | 0.004 |
| MC7 | -0.004 | 0.047 | | 0.041 | 0.023 | 0.007 | 0.002 |
| MC8 | -0.008 | 0.018 | 0.041 | | 0.039 | 0.028 | 0.009 |
| MC9 | 0.023 | 0.037 | 0.023 | 0.039 | | -0.028 | 0.019 |
| MC10 | -0.018 | 0.003 | 0.007 | 0.028 | -0.028 | | -0.016 |
| MC11 | 0.014 | 0.004 | 0.002 | 0.009 | 0.019 | -0.016 | |
| MC12 | 0.016 | 0.021 | 0.015 | -0.014 | -0.007 | -0.030 | 0.039 |
| MC13 | 0.007 | -0.002 | 0.008 | -0.001 | 0.007 | -0.028 | 0.064 MR |
| MC14 | 0.009 | 0.018 | 0.012 | -0.007 | 0.004 | -0.005 | 0.013 |
| MC15 | -0.026 | -0.015 | 0.003 | -0.012 | 0.019 | 0.021 | -0.012 |
| MC16 | 0.002 | 0.017 | 0.027 | -0.007 | 0.006 | 0.038 | -0.015 |
| MC17 | -0.011 | -0.013 | -0.018 | -0.005 | -0.003 | -0.012 | -0.001 |
| MC18 | -0.006 | 0.016 | 0.000 | -0.008 | -0.022 | 0.049 | 0.004 |
| MC19 | -0.003 | 0.006 | -0.013 | -0.006 | 0.006 | 0.028 | 0.003 |
| MC20 | -0.003 | 0.011 | 0.016 | 0.019 | 0.008 | -0.017 | -0.006 |
| MC21 | 0.000 | -0.011 | 0.018 | 0.020 | 0.015 | -0.021 | 0.000 |
| MC22 | -0.015 | -0.022 | -0.011 | 0.003 | -0.002 | 0.009 | -0.005 |
| MC23 | -0.023 | -0.005 | 0.001 | 0.000 | 0.008 | 0.005 | -0.010 |
| MC24 | 0.003 | 0.019 | -0.016 | 0.008 | 0.017 | 0.011 | -0.025 |
| MC25 | -0.006 | -0.007 | -0.016 | 0.000 | 0.023 | 0.000 | 0.011 |
| MC26 | -0.006 | 0.004 | -0.005 | 0.001 | 0.009 | -0.013 | -0.003 |
| MC27 | -0.009 | -0.014 | -0.007 | -0.006 | 0.014 | 0.019 | 0.002 |
| MC28 | -0.021 | -0.018 | -0.008 | -0.005 | -0.005 | -0.007 | -0.008 |
| MC29 | -0.001 | -0.033 | 0.000 | 0.003 | -0.027 | 0.010 | 0.000 |
| MC30 | 0.010 | 0.001 | -0.015 | 0.002 | 0.006 | 0.015 | -0.002 |
| MC31 | -0.005 | -0.027 | 0.012 | -0.012 | -0.015 | -0.020 | 0.009 |
| MC32 | -0.005 | -0.024 | -0.003 | -0.022 | -0.019 | -0.047 | -0.001 |
| MC33 | 0.012 | -0.016 | -0.014 | -0.005 | -0.028 | -0.035 | -0.009 |
| MC34 | -0.009 | -0.029 | -0.029 | -0.005 | -0.008 | -0.028 | -0.017 |
| MC35 | 0.008 | 0.004 | -0.012 | -0.009 | -0.001 | 0.048 | -0.014 |
| MC36 | 0.001 | -0.008 | -0.013 | 0.011 | -0.021 | -0.008 | -0.011 |
| MC37 | 0.002 | -0.003 | -0.015 | 0.002 | 0.005 | 0.031 | 0.009 |
| MC38 | -0.006 | 0.017 | -0.008 | 0.001 | -0.012 | 0.001 | -0.013 |
| OE1_1M | -0.010 | -0.004 | -0.001 | 0.011 | -0.012 | -0.012 | -0.007 |
| OE4_1M | -0.008 | -0.006 | -0.017 | -0.010 | -0.011 | 0.012 | 0.002 |
| OE2_4M | -0.006 | 0.000 | 0.003 | -0.015 | -0.009 | 0.002 | -0.010 |
| OE3_4M | 0.000 | -0.003 | -0.008 | -0.003 | -0.014 | 0.005 | -0.008 |
| OE5_4M | 0.013 | -0.003 | -0.028 | -0.009 | 0.001 | 0.016 | 0.006 |
| OE6_4M | 0.010 | -0.001 | -0.015 | -0.022 | -0.016 | 0.015 | 0.006 |
| # of MR | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| # of LR | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| | MC12 | MC13 | MC14 | MC15 | MC16 | MC17 | MC18 |
|---|---|---|---|---|---|---|---|
| MC1 | 0.013 | 0.017 | 0.020 | 0.007 | 0.023 | 0.031 | 0.002 |
| MC2 | 0.026 | 0.007 | 0.028 | -0.002 | 0.021 | 0.007 | 0.004 |
| MC3 | 0.006 | 0.006 | 0.018 | 0.002 | -0.001 | -0.004 | -0.007 |
| MC4 | -0.015 | 0.006 | 0.006 | 0.010 | 0.017 | 0.002 | -0.007 |
| MC5 | 0.016 | 0.007 | 0.009 | -0.026 | 0.002 | -0.011 | -0.006 |
| MC6 | 0.021 | -0.002 | 0.018 | -0.015 | 0.017 | -0.013 | 0.016 |
| MC7 | 0.015 | 0.008 | 0.012 | 0.003 | 0.027 | -0.018 | 0.000 |
| MC8 | -0.014 | -0.001 | -0.007 | -0.012 | -0.007 | -0.005 | -0.008 |
| MC9 | -0.007 | 0.007 | 0.004 | 0.019 | 0.006 | -0.003 | -0.022 |
| MC10 | -0.030 | -0.028 | -0.005 | 0.021 | 0.038 | -0.012 | 0.049 |
| MC11 | 0.039 | 0.064 MR | 0.013 | -0.012 | -0.015 | -0.001 | 0.004 |
| MC12 | | 0.081 MR | 0.089 MR | -0.017 | 0.016 | -0.005 | 0.006 |
| MC13 | 0.081 MR | | 0.058 MR | 0.011 | -0.011 | -0.003 | -0.018 |
| MC14 | 0.089 MR | 0.058 MR | | -0.016 | -0.013 | -0.024 | 0.011 |
| MC15 | -0.017 | 0.011 | -0.016 | | -0.003 | 0.005 | -0.006 |
| MC16 | 0.016 | -0.011 | -0.013 | -0.003 | | 0.029 | 0.019 |
| MC17 | -0.005 | -0.003 | -0.024 | 0.005 | 0.029 | | 0.119 LR |
| MC18 | 0.006 | -0.018 | 0.011 | -0.006 | 0.019 | 0.119 LR | |
| MC19 | -0.002 | -0.021 | -0.015 | 0.017 | 0.007 | 0.065 MR | 0.053 MR |
| MC20 | -0.023 | -0.028 | -0.017 | 0.008 | 0.000 | 0.003 | -0.010 |
| MC21 | -0.003 | 0.029 | -0.040 | -0.018 | -0.008 | 0.000 | -0.024 |
| MC22 | -0.011 | -0.001 | -0.019 | 0.034 | -0.003 | -0.005 | -0.015 |
| MC23 | -0.021 | -0.010 | -0.012 | 0.003 | -0.009 | 0.004 | -0.001 |
| MC24 | -0.037 | -0.027 | -0.018 | 0.014 | 0.017 | 0.017 | 0.013 |
| MC25 | 0.010 | -0.015 | -0.010 | 0.001 | -0.001 | 0.001 | -0.005 |
| MC26 | -0.008 | -0.025 | -0.014 | 0.007 | -0.023 | -0.007 | -0.019 |
| MC27 | -0.005 | -0.029 | -0.013 | 0.000 | 0.012 | -0.026 | -0.017 |
| MC28 | -0.008 | 0.004 | -0.016 | 0.009 | -0.014 | -0.027 | -0.025 |
| MC29 | -0.002 | -0.016 | -0.014 | 0.014 | -0.001 | -0.019 | -0.021 |
| MC30 | -0.026 | -0.011 | -0.001 | 0.011 | -0.003 | -0.014 | -0.003 |
| MC31 | -0.008 | 0.014 | 0.001 | -0.018 | -0.029 | -0.024 | -0.057 MR |
| MC32 | -0.039 | -0.004 | -0.021 | -0.029 | -0.025 | -0.010 | -0.030 |
| MC33 | -0.013 | 0.018 | -0.002 | -0.004 | -0.012 | -0.019 | -0.025 |
| MC34 | -0.015 | -0.002 | -0.034 | -0.015 | -0.020 | -0.006 | -0.032 |
| MC35 | 0.013 | -0.020 | 0.008 | 0.009 | 0.020 | 0.004 | 0.030 |
| MC36 | -0.010 | 0.003 | 0.001 | 0.006 | -0.005 | -0.003 | -0.001 |
| MC37 | -0.009 | -0.007 | 0.002 | 0.015 | -0.012 | -0.038 | -0.009 |
| MC38 | -0.014 | -0.007 | -0.016 | -0.004 | -0.017 | -0.001 | -0.007 |
| OE1_1M | 0.000 | -0.001 | 0.010 | 0.009 | -0.006 | 0.004 | 0.014 |
| OE4_1M | -0.003 | -0.007 | 0.024 | -0.024 | -0.015 | 0.006 | -0.001 |
| OE2_4M | -0.025 | -0.009 | 0.006 | 0.016 | -0.017 | -0.004 | 0.017 |
| OE3_4M | 0.008 | -0.005 | -0.014 | -0.002 | 0.010 | 0.005 | 0.018 |
| OE5_4M | 0.004 | -0.020 | 0.019 | -0.024 | 0.002 | 0.006 | 0.017 |
| OE6_4M | 0.002 | -0.002 | 0.011 | 0.003 | 0.007 | -0.004 | 0.019 |
| # of MR | 2 | 3 | 2 | 0 | 0 | 1 | 2 |
| # of LR | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

| | MC19 | MC20 | MC21 | MC22 | MC23 | MC24 | MC25 |
|---|---|---|---|---|---|---|---|
| MC1 | 0.006 | 0.003 | -0.021 | 0.019 | -0.015 | -0.007 | -0.007 |
| MC2 | -0.001 | 0.004 | 0.015 | 0.002 | -0.008 | -0.005 | -0.008 |
| MC3 | -0.011 | 0.004 | -0.019 | -0.006 | -0.001 | -0.014 | 0.008 |
| MC4 | 0.018 | 0.001 | -0.005 | 0.008 | 0.020 | 0.016 | 0.006 |
| MC5 | -0.003 | -0.003 | 0.000 | -0.015 | -0.023 | 0.003 | -0.006 |
| MC6 | 0.006 | 0.011 | -0.011 | -0.022 | -0.005 | 0.019 | -0.007 |
| MC7 | -0.013 | 0.016 | 0.018 | -0.011 | 0.001 | -0.016 | -0.016 |
| MC8 | -0.006 | 0.019 | 0.020 | 0.003 | 0.000 | 0.008 | 0.000 |
| MC9 | 0.006 | 0.008 | 0.015 | -0.002 | 0.008 | 0.017 | 0.023 |
| MC10 | 0.028 | -0.017 | -0.021 | 0.009 | 0.005 | 0.011 | 0.000 |
| MC11 | 0.003 | -0.006 | 0.000 | -0.005 | -0.010 | -0.025 | 0.011 |
| MC12 | -0.002 | -0.023 | -0.003 | -0.011 | -0.021 | -0.037 | 0.010 |
| MC13 | -0.021 | -0.028 | 0.029 | -0.001 | -0.010 | -0.027 | -0.015 |
| MC14 | -0.015 | -0.017 | -0.040 | -0.019 | -0.012 | -0.018 | -0.010 |
| MC15 | 0.017 | 0.008 | -0.018 | 0.034 | 0.003 | 0.014 | 0.001 |
| MC16 | 0.007 | 0.000 | -0.008 | -0.003 | -0.009 | 0.017 | -0.001 |
| MC17 | 0.065 MR | 0.003 | 0.000 | -0.005 | 0.004 | 0.017 | 0.001 |
| MC18 | 0.053 MR | -0.010 | -0.024 | -0.015 | -0.001 | 0.013 | -0.005 |
| MC19 | | 0.021 | -0.013 | 0.012 | -0.014 | 0.025 | -0.007 |
| MC20 | 0.021 | | 0.092 MR | 0.013 | 0.062 MR | -0.008 | 0.008 |
| MC21 | -0.013 | 0.092 MR | | 0.039 | 0.056 MR | 0.002 | 0.017 |
| MC22 | 0.012 | 0.013 | 0.039 | | 0.066 MR | 0.030 | -0.020 |
| MC23 | -0.014 | 0.062 MR | 0.056 MR | 0.066 MR | | 0.005 | 0.034 |
| MC24 | 0.025 | -0.008 | 0.002 | 0.030 | 0.005 | | -0.032 |
| MC25 | -0.007 | 0.008 | 0.017 | -0.020 | 0.034 | -0.032 | |
| MC26 | -0.002 | 0.023 | 0.016 | -0.005 | 0.006 | -0.009 | 0.028 |
| MC27 | -0.010 | -0.007 | -0.015 | 0.019 | -0.016 | 0.011 | 0.029 |
| MC28 | -0.039 | -0.001 | -0.010 | 0.002 | 0.001 | -0.021 | 0.009 |
| MC29 | -0.001 | -0.020 | -0.028 | 0.015 | -0.016 | 0.001 | 0.001 |
| MC30 | -0.005 | -0.026 | -0.055 MR | -0.001 | -0.024 | 0.006 | -0.002 |
| MC31 | -0.018 | 0.009 | 0.010 | -0.016 | 0.017 | -0.026 | 0.001 |
| MC32 | -0.027 | 0.002 | 0.009 | -0.020 | 0.007 | -0.022 | 0.006 |
| MC33 | -0.014 | 0.008 | 0.022 | 0.005 | -0.020 | -0.015 | -0.008 |
| MC34 | -0.029 | -0.009 | 0.037 | -0.019 | -0.011 | -0.014 | 0.006 |
| MC35 | 0.009 | -0.026 | -0.046 | 0.002 | -0.030 | 0.036 | 0.006 |
| MC36 | 0.002 | -0.008 | -0.010 | -0.017 | -0.008 | 0.013 | 0.003 |
| MC37 | -0.012 | 0.015 | -0.003 | 0.001 | 0.021 | -0.015 | 0.000 |
| MC38 | 0.012 | -0.015 | 0.005 | -0.005 | -0.001 | -0.005 | 0.000 |
| OE1_1M | -0.013 | -0.008 | -0.010 | 0.002 | 0.006 | 0.006 | -0.015 |
| OE4_1M | -0.013 | -0.005 | -0.008 | -0.017 | -0.011 | 0.007 | -0.006 |
| OE2_4M | -0.011 | -0.016 | -0.007 | -0.001 | 0.007 | 0.012 | -0.012 |
| OE3_4M | -0.007 | -0.014 | -0.013 | -0.012 | -0.017 | -0.007 | -0.015 |
| OE5_4M | 0.010 | -0.017 | -0.021 | -0.012 | -0.010 | 0.020 | -0.028 |
| OE6_4M | 0.015 | -0.045 | -0.027 | -0.014 | -0.035 | 0.008 | -0.005 |
| # of MR | 2 | 2 | 3 | 1 | 3 | 0 | 0 |
| # of LR | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| | MC26 | MC27 | MC28 | MC29 | MC30 | MC31 | MC32 |
|---|---|---|---|---|---|---|---|
| MC1 | -0.007 | 0.002 | 0.009 | 0.008 | -0.019 | -0.034 | -0.005 |
| MC2 | -0.008 | -0.007 | 0.002 | 0.010 | -0.020 | 0.000 | -0.003 |
| MC3 | 0.006 | 0.007 | 0.009 | -0.008 | 0.013 | -0.010 | -0.007 |
| MC4 | -0.007 | -0.011 | -0.007 | 0.004 | -0.027 | -0.013 | 0.000 |
| MC5 | -0.006 | -0.009 | -0.021 | -0.001 | 0.010 | -0.005 | -0.005 |
| MC6 | 0.004 | -0.014 | -0.018 | -0.033 | 0.001 | -0.027 | -0.024 |
| MC7 | -0.005 | -0.007 | -0.008 | 0.000 | -0.015 | 0.012 | -0.003 |
| MC8 | 0.001 | -0.006 | -0.005 | 0.003 | 0.002 | -0.012 | -0.022 |
| MC9 | 0.009 | 0.014 | -0.005 | -0.027 | 0.006 | -0.015 | -0.019 |
| MC10 | -0.013 | 0.019 | -0.007 | 0.010 | 0.015 | -0.020 | -0.047 |
| MC11 | -0.003 | 0.002 | -0.008 | 0.000 | -0.002 | 0.009 | -0.001 |
| MC12 | -0.008 | -0.005 | -0.008 | -0.002 | -0.026 | -0.008 | -0.039 |
| MC13 | -0.025 | -0.029 | 0.004 | -0.016 | -0.011 | 0.014 | -0.004 |
| MC14 | -0.014 | -0.013 | -0.016 | -0.014 | -0.001 | 0.001 | -0.021 |
| MC15 | 0.007 | 0.000 | 0.009 | 0.014 | 0.011 | -0.018 | -0.029 |
| MC16 | -0.023 | 0.012 | -0.014 | -0.001 | -0.003 | -0.029 | -0.025 |
| MC17 | -0.007 | -0.026 | -0.027 | -0.019 | -0.014 | -0.024 | -0.010 |
| MC18 | -0.019 | -0.017 | -0.025 | -0.021 | -0.003 | -0.057 MR | -0.030 |
| MC19 | -0.002 | -0.010 | -0.039 | -0.001 | -0.005 | -0.018 | -0.027 |
| MC20 | 0.023 | -0.007 | -0.001 | -0.020 | -0.026 | 0.009 | 0.002 |
| MC21 | 0.016 | -0.015 | -0.010 | -0.028 | -0.055 MR | 0.010 | 0.009 |
| MC22 | -0.005 | 0.019 | 0.002 | 0.015 | -0.001 | -0.016 | -0.020 |
| MC23 | 0.006 | -0.016 | 0.001 | -0.016 | -0.024 | 0.017 | 0.007 |
| MC24 | -0.009 | 0.011 | -0.021 | 0.001 | 0.006 | -0.026 | -0.022 |
| MC25 | 0.028 | 0.029 | 0.009 | 0.001 | -0.002 | 0.001 | 0.006 |
| MC26 | | 0.027 | 0.037 | -0.001 | 0.017 | 0.021 | 0.012 |
| MC27 | 0.027 | | 0.036 | 0.080 MR | 0.035 | 0.002 | -0.002 |
| MC28 | 0.037 | 0.036 | | 0.108 LR | 0.058 MR | 0.115 LR | 0.104 LR |
| MC29 | -0.001 | 0.080 MR | 0.108 LR | | 0.035 | 0.025 | 0.020 |
| MC30 | 0.017 | 0.035 | 0.058 MR | 0.035 | | 0.052 MR | 0.053 MR |
| MC31 | 0.021 | 0.002 | 0.115 LR | 0.025 | 0.052 MR | | 0.163 LR |
| MC32 | 0.012 | -0.002 | 0.104 LR | 0.020 | 0.053 MR | 0.163 LR | |
| MC33 | -0.009 | -0.001 | -0.008 | -0.005 | -0.017 | 0.049 | 0.019 |
| MC34 | 0.027 | -0.001 | 0.011 | 0.002 | 0.001 | 0.042 | 0.034 |
| MC35 | -0.003 | 0.009 | -0.024 | 0.027 | 0.023 | -0.044 | -0.042 |
| MC36 | -0.014 | -0.008 | -0.017 | -0.013 | 0.009 | -0.009 | -0.004 |
| MC37 | -0.008 | 0.013 | 0.019 | 0.026 | 0.003 | -0.009 | 0.001 |
| MC38 | 0.009 | -0.014 | -0.025 | -0.013 | -0.012 | -0.003 | -0.006 |
| OE1_1M | -0.017 | 0.002 | -0.012 | -0.014 | -0.003 | -0.018 | -0.004 |
| OE4_1M | -0.006 | -0.025 | -0.030 | -0.029 | -0.008 | -0.018 | -0.016 |
| OE2_4M | -0.005 | 0.004 | -0.023 | -0.009 | 0.022 | -0.016 | -0.009 |
| OE3_4M | -0.016 | 0.001 | -0.027 | -0.027 | -0.010 | -0.040 | -0.038 |
| OE5_4M | -0.032 | -0.028 | -0.043 | -0.025 | -0.018 | -0.037 | -0.039 |
| OE6_4M | -0.016 | -0.006 | -0.039 | -0.025 | -0.017 | -0.043 | -0.030 |
| # of MR | 0 | 1 | 1 | 1 | 4 | 2 | 1 |
| # of LR | 0 | 0 | 3 | 1 | 0 | 2 | 2 |

| | MC33 | MC34 | MC35 | MC36 | MC37 | MC38 |
|---|---|---|---|---|---|---|
| MC1 | -0.008 | 0.003 | -0.008 | -0.014 | -0.018 | -0.020 |
| MC2 | -0.003 | -0.026 | -0.006 | -0.024 | 0.003 | -0.034 |
| MC3 | -0.014 | -0.006 | -0.007 | -0.010 | 0.022 | 0.007 |
| MC4 | -0.002 | 0.000 | 0.000 | -0.015 | -0.001 | 0.001 |
| MC5 | 0.012 | -0.009 | 0.008 | 0.001 | 0.002 | -0.006 |
| MC6 | -0.016 | -0.029 | 0.004 | -0.008 | -0.003 | 0.017 |
| MC7 | -0.014 | -0.029 | -0.012 | -0.013 | -0.015 | -0.008 |
| MC8 | -0.005 | -0.005 | -0.009 | 0.011 | 0.002 | 0.001 |
| MC9 | -0.028 | -0.008 | -0.001 | -0.021 | 0.005 | -0.012 |
| MC10 | -0.035 | -0.028 | 0.048 | -0.008 | 0.031 | 0.001 |
| MC11 | -0.009 | -0.017 | -0.014 | -0.011 | 0.009 | -0.013 |
| MC12 | -0.013 | -0.015 | 0.013 | -0.010 | -0.009 | -0.014 |
| MC13 | 0.018 | -0.002 | -0.020 | 0.003 | -0.007 | -0.007 |
| MC14 | -0.002 | -0.034 | 0.008 | 0.001 | 0.002 | -0.016 |
| MC15 | -0.004 | -0.015 | 0.009 | 0.006 | 0.015 | -0.004 |
| MC16 | -0.012 | -0.020 | 0.020 | -0.005 | -0.012 | -0.017 |
| MC17 | -0.019 | -0.006 | 0.004 | -0.003 | -0.038 | -0.001 |
| MC18 | -0.025 | -0.032 | 0.030 | -0.001 | -0.009 | -0.007 |
| MC19 | -0.014 | -0.029 | 0.009 | 0.002 | -0.012 | 0.012 |
| MC20 | 0.008 | -0.009 | -0.026 | -0.008 | 0.015 | -0.015 |
| MC21 | 0.022 | 0.037 | -0.046 | -0.010 | -0.003 | 0.005 |
| MC22 | 0.005 | -0.019 | 0.002 | -0.017 | 0.001 | -0.005 |
| MC23 | -0.020 | -0.011 | -0.030 | -0.008 | 0.021 | -0.001 |
| MC24 | -0.015 | -0.014 | 0.036 | 0.013 | -0.015 | -0.005 |
| MC25 | -0.008 | 0.006 | 0.006 | 0.003 | 0.000 | 0.000 |
| MC26 | -0.009 | 0.027 | -0.003 | -0.014 | -0.008 | 0.009 |
| MC27 | -0.001 | -0.001 | 0.009 | -0.008 | 0.013 | -0.014 |
| MC28 | -0.008 | 0.011 | -0.024 | -0.017 | 0.019 | -0.025 |
| MC29 | -0.005 | 0.002 | 0.027 | -0.013 | 0.026 | -0.013 |
| MC30 | -0.017 | 0.001 | 0.023 | 0.009 | 0.003 | -0.012 |
| MC31 | 0.049 | 0.042 | -0.044 | -0.009 | -0.009 | -0.003 |
| MC32 | 0.019 | 0.034 | -0.042 | -0.004 | 0.001 | -0.006 |
| MC33 | | 0.086 MR | -0.034 | 0.055 MR | 0.006 | 0.044 |
| MC34 | 0.086 MR | | 0.004 | 0.051 MR | -0.001 | 0.098 MR |
| MC35 | -0.034 | 0.004 | | 0.008 | 0.000 | 0.010 |
| MC36 | 0.055 MR | 0.051 MR | 0.008 | | 0.026 | 0.067 MR |
| MC37 | 0.006 | -0.001 | 0.000 | 0.026 | | -0.011 |
| MC38 | 0.044 | 0.098 MR | 0.010 | 0.067 MR | -0.011 | |
| OE1_1M | -0.017 | 0.001 | 0.006 | -0.009 | -0.002 | -0.026 |
| OE4_1M | -0.002 | -0.009 | 0.006 | 0.001 | -0.018 | -0.014 |
| OE2_4M | -0.018 | -0.038 | 0.011 | -0.009 | 0.004 | -0.006 |
| OE3_4M | -0.008 | -0.019 | 0.012 | -0.006 | -0.022 | -0.010 |
| OE5_4M | -0.019 | -0.024 | 0.015 | -0.008 | -0.010 | -0.026 |
| OE6_4M | -0.006 | -0.027 | 0.028 | 0.008 | -0.002 | -0.002 |
| # of MR | 2 | 3 | 0 | 3 | 0 | 2 |
| # of LR | 0 | 0 | 0 | 0 | 0 | 0 |

| | OE1_1M | OE4_1M | OE2_4M | OE3_4M | OE5_4M | OE6_4M |
|---|---|---|---|---|---|---|
| MC1 | -0.006 | -0.004 | 0.002 | -0.014 | 0.004 | -0.014 |
| MC2 | -0.002 | -0.020 | -0.007 | 0.000 | -0.005 | -0.020 |
| MC3 | -0.004 | -0.005 | 0.022 | -0.014 | -0.011 | -0.037 |
| MC4 | 0.002 | -0.006 | -0.003 | -0.021 | -0.013 | -0.011 |
| MC5 | -0.010 | -0.008 | -0.006 | 0.000 | 0.013 | 0.010 |
| MC6 | -0.004 | -0.006 | 0.000 | -0.003 | -0.003 | -0.001 |
| MC7 | -0.001 | -0.017 | 0.003 | -0.008 | -0.028 | -0.015 |
| MC8 | 0.011 | -0.010 | -0.015 | -0.003 | -0.009 | -0.022 |
| MC9 | -0.012 | -0.011 | -0.009 | -0.014 | 0.001 | -0.016 |
| MC10 | -0.012 | 0.012 | 0.002 | 0.005 | 0.016 | 0.015 |
| MC11 | -0.007 | 0.002 | -0.010 | -0.008 | 0.006 | 0.006 |
| MC12 | 0.000 | -0.003 | -0.025 | 0.008 | 0.004 | 0.002 |
| MC13 | -0.001 | -0.007 | -0.009 | -0.005 | -0.020 | -0.002 |
| MC14 | 0.010 | 0.024 | 0.006 | -0.014 | 0.019 | 0.011 |
| MC15 | 0.009 | -0.024 | 0.016 | -0.002 | -0.024 | 0.003 |
| MC16 | -0.006 | -0.015 | -0.017 | 0.010 | 0.002 | 0.007 |
| MC17 | 0.004 | 0.006 | -0.004 | 0.005 | 0.006 | -0.004 |
| MC18 | 0.014 | -0.001 | 0.017 | 0.018 | 0.017 | 0.019 |
| MC19 | -0.013 | -0.013 | -0.011 | -0.007 | 0.010 | 0.015 |
| MC20 | -0.008 | -0.005 | -0.016 | -0.014 | -0.017 | -0.045 |
| MC21 | -0.010 | -0.008 | -0.007 | -0.013 | -0.021 | -0.027 |
| MC22 | 0.002 | -0.017 | -0.001 | -0.012 | -0.012 | -0.014 |
| MC23 | 0.006 | -0.011 | 0.007 | -0.017 | -0.010 | -0.035 |
| MC24 | 0.006 | 0.007 | 0.012 | -0.007 | 0.020 | 0.008 |
| MC25 | -0.015 | -0.006 | -0.012 | -0.015 | -0.028 | -0.005 |
| MC26 | -0.017 | -0.006 | -0.005 | -0.016 | -0.032 | -0.016 |
| MC27 | 0.002 | -0.025 | 0.004 | 0.001 | -0.028 | -0.006 |
| MC28 | -0.012 | -0.030 | -0.023 | -0.027 | -0.043 | -0.039 |
| MC29 | -0.014 | -0.029 | -0.009 | -0.027 | -0.025 | -0.025 |
| MC30 | -0.003 | -0.008 | 0.022 | -0.010 | -0.018 | -0.017 |
| MC31 | -0.018 | -0.018 | -0.016 | -0.040 | -0.037 | -0.043 |
| MC32 | -0.004 | -0.016 | -0.009 | -0.038 | -0.039 | -0.030 |
| MC33 | -0.017 | -0.002 | -0.018 | -0.008 | -0.019 | -0.006 |
| MC34 | 0.001 | -0.009 | -0.038 | -0.019 | -0.024 | -0.027 |
| MC35 | 0.006 | 0.006 | 0.011 | 0.012 | 0.015 | 0.028 |
| MC36 | -0.009 | 0.001 | -0.009 | -0.006 | -0.008 | 0.008 |
| MC37 | -0.002 | -0.018 | 0.004 | -0.022 | -0.010 | -0.002 |
| MC38 | -0.026 | -0.014 | -0.006 | -0.010 | -0.026 | -0.002 |
| OE1_1M | | 0.005 | 0.041 | 0.058 MR | 0.014 | 0.027 |
| OE4_1M | 0.005 | | -0.009 | 0.039 | 0.183 LR | 0.041 |
| OE2_4M | 0.041 | -0.009 | | 0.072 MR | 0.028 | 0.034 |
| OE3_4M | 0.058 MR | 0.039 | 0.072 MR | | 0.075 MR | 0.087 MR |
| OE5_4M | 0.014 | 0.183 LR | 0.028 | 0.075 MR | | 0.156 LR |
| OE6_4M | 0.027 | 0.041 | 0.034 | 0.087 MR | 0.156 LR | |
| # of MR | 1 | 0 | 1 | 4 | 1 | 1 |
| # of LR | 0 | 1 | 0 | 0 | 2 | 1 |

# APPENDIX B

Numeracy Residual Correlation Matrix with First Factor Removed[a].

| | MC1 | | MC2 | | MC3 | | MC4 | |
|---|---|---|---|---|---|---|---|---|
| MC1 | | | -0.030 | | -0.028 | | 0.100 | LR |
| MC2 | -0.030 | | | | 0.077 | MR | -0.026 | |
| MC3 | -0.028 | | 0.077 | MR | | | -0.015 | |
| MC4 | 0.100 | LR | -0.026 | | -0.015 | | | |
| MC5 | -0.010 | | 0.022 | | 0.045 | | -0.029 | |
| MC6 | 0.006 | | 0.006 | | 0.011 | | -0.013 | |
| MC7 | -0.028 | | 0.112 | LR | 0.038 | | -0.033 | |
| MC8 | -0.038 | | 0.172 | LR | 0.028 | | -0.022 | |
| MC9 | -0.022 | | 0.032 | | 0.019 | | -0.017 | |
| MC10 | 0.023 | | -0.025 | | -0.007 | | 0.036 | |
| MC11 | 0.012 | | 0.017 | | 0.002 | | 0.002 | |
| MC12 | 0.004 | | -0.023 | | -0.021 | | -0.002 | |
| MC13 | -0.012 | | -0.021 | | 0.003 | | 0.001 | |
| MC14 | 0.064 | MR | -0.041 | | -0.033 | | 0.066 | MR |
| MC15 | 0.041 | | -0.042 | | -0.033 | | 0.034 | |
| MC16 | 0.002 | | 0.014 | | 0.024 | | 0.000 | |
| MC19 | 0.055 | MR | -0.005 | | 0.001 | | 0.011 | |
| MC20 | 0.006 | | -0.012 | | -0.008 | | 0.014 | |
| MC21 | -0.031 | | 0.014 | | 0.078 | MR | -0.044 | |
| MC22 | -0.010 | | 0.002 | | -0.001 | | -0.018 | |
| MC23 | -0.003 | | 0.002 | | 0.009 | | 0.012 | |
| MC24 | 0.002 | | 0.004 | | 0.013 | | -0.013 | |
| MC25 | -0.021 | | -0.022 | | -0.004 | | -0.033 | |
| MC26 | -0.001 | | -0.048 | | -0.033 | | 0.017 | |
| MC27 | -0.012 | | -0.025 | | 0.005 | | 0.001 | |
| MC28 | 0.012 | | -0.041 | | -0.033 | | 0.023 | |
| MC29 | -0.003 | | -0.043 | | -0.046 | | 0.036 | |
| MC30 | 0.001 | | -0.015 | | -0.017 | | -0.005 | |
| MC31 | 0.027 | | -0.022 | | -0.030 | | 0.023 | |
| MC32 | -0.002 | | 0.000 | | -0.003 | | -0.008 | |
| MC33 | -0.010 | | -0.013 | | 0.012 | | -0.010 | |
| MC34 | -0.035 | | -0.020 | | 0.007 | | -0.025 | |
| OE17 | 0.000 | | 0.047 | | -0.009 | | 0.000 | |
| OE18 | -0.005 | | -0.030 | | -0.018 | | -0.014 | |
| OE35 | -0.004 | | -0.013 | | -0.032 | | -0.006 | |
| OE36 | -0.014 | | -0.036 | | -0.025 | | -0.004 | |
| Number of moderate residuals (MR) by item. | 2 | | 1 | | 2 | | 1 | |
| Number of large residuals (LR) by item. | 1 | | 2 | | 0 | | 1 | |

Note. Across 630 unique pairings, there are 11 (1.7%) moderate residuals with absolute values between .05 and .10 (denoted by "MR") and 8 (1.3%) large residuals with absolute values equal to or greater than .1 (denoted by "LR").

[a]Residuals represent the difference between the observed and reproduced correlational matrices.

| | MC5 | MC6 | MC7 | MC8 | MC9 | MC10 |
|---|---|---|---|---|---|---|
| MC1 | -0.010 | 0.006 | -0.028 | -0.038 | -0.022 | 0.023 |
| MC2 | 0.022 | 0.006 | 0.112 LR | 0.172 LR | 0.032 | -0.025 |
| MC3 | 0.045 | 0.011 | 0.038 | 0.028 | 0.019 | -0.007 |
| MC4 | -0.029 | -0.013 | -0.033 | -0.022 | -0.017 | 0.036 |
| MC5 | | 0.001 | 0.031 | 0.017 | 0.023 | 0.023 |
| MC6 | 0.001 | | 0.030 | -0.014 | 0.008 | 0.008 |
| MC7 | 0.031 | 0.030 | | 0.079 MR | 0.039 | -0.023 |
| MC8 | 0.017 | -0.014 | 0.079 MR | | 0.007 | -0.008 |
| MC9 | 0.023 | 0.008 | 0.039 | 0.007 | | 0.018 |
| MC10 | 0.023 | 0.008 | -0.023 | -0.008 | 0.018 | |
| MC11 | 0.002 | -0.002 | 0.007 | 0.022 | 0.020 | 0.031 |
| MC12 | 0.000 | -0.006 | -0.003 | -0.014 | 0.002 | 0.008 |
| MC13 | -0.003 | -0.005 | -0.009 | -0.015 | 0.003 | -0.003 |
| MC14 | -0.013 | -0.008 | -0.033 | -0.022 | -0.009 | 0.040 |
| MC15 | -0.003 | -0.001 | -0.028 | -0.019 | -0.018 | 0.011 |
| MC16 | 0.021 | -0.004 | 0.023 | 0.019 | 0.014 | 0.011 |
| MC19 | 0.016 | -0.002 | -0.013 | -0.016 | 0.000 | -0.001 |
| MC20 | -0.007 | -0.001 | -0.017 | 0.000 | -0.005 | -0.002 |
| MC21 | 0.011 | -0.010 | 0.012 | 0.009 | 0.013 | -0.011 |
| MC22 | -0.012 | 0.007 | -0.013 | -0.012 | 0.004 | 0.002 |
| MC23 | -0.010 | 0.021 | 0.008 | -0.008 | 0.012 | 0.007 |
| MC24 | -0.012 | -0.001 | 0.008 | -0.008 | -0.006 | -0.009 |
| MC25 | -0.001 | -0.003 | -0.005 | -0.014 | -0.009 | -0.012 |
| MC26 | -0.039 | 0.004 | -0.020 | -0.037 | -0.013 | -0.027 |
| MC27 | 0.012 | 0.009 | -0.018 | -0.023 | -0.004 | -0.009 |
| MC28 | 0.003 | 0.013 | -0.037 | -0.031 | -0.031 | -0.001 |
| MC29 | -0.009 | 0.002 | -0.026 | -0.037 | -0.006 | 0.011 |
| MC30 | -0.010 | -0.014 | -0.037 | -0.007 | 0.002 | -0.008 |
| MC31 | -0.009 | -0.006 | -0.031 | -0.007 | -0.026 | -0.016 |
| MC32 | -0.013 | 0.006 | -0.020 | -0.017 | -0.006 | -0.005 |
| MC33 | -0.003 | -0.006 | 0.011 | -0.013 | -0.003 | 0.008 |
| MC34 | 0.050 MR | -0.011 | -0.008 | -0.018 | 0.017 | -0.016 |
| OE17 | -0.030 | 0.002 | -0.017 | 0.059 MR | -0.003 | 0.003 |
| OE18 | -0.009 | -0.009 | -0.009 | -0.019 | -0.015 | -0.010 |
| OE35 | -0.013 | 0.000 | -0.018 | -0.029 | -0.014 | -0.010 |
| OE36 | -0.015 | -0.009 | -0.007 | -0.028 | -0.022 | -0.005 |
| # of MR | 1 | 0 | 1 | 2 | 0 | 0 |
| # of LR | 0 | 0 | 1 | 1 | 0 | 0 |

| | MC11 | MC12 | MC13 | MC14 | MC15 | MC16 |
|---|---|---|---|---|---|---|
| MC1 | 0.012 | 0.004 | -0.012 | 0.064 MR | 0.041 | 0.002 |
| MC2 | 0.017 | -0.023 | -0.021 | -0.041 | -0.042 | 0.014 |
| MC3 | 0.002 | -0.021 | 0.003 | -0.033 | -0.033 | 0.024 |
| MC4 | 0.002 | -0.002 | 0.001 | 0.066 MR | 0.034 | 0.000 |
| MC5 | 0.002 | 0.000 | -0.003 | -0.013 | -0.003 | 0.021 |
| MC6 | -0.002 | -0.006 | -0.005 | -0.008 | -0.001 | -0.004 |
| MC7 | 0.007 | -0.003 | -0.009 | -0.033 | -0.028 | 0.023 |
| MC8 | 0.022 | -0.014 | -0.015 | -0.022 | -0.019 | 0.019 |
| MC9 | 0.020 | 0.002 | 0.003 | -0.009 | -0.018 | 0.014 |
| MC10 | 0.031 | 0.008 | -0.003 | 0.040 | 0.011 | 0.011 |
| MC11 | | -0.002 | -0.014 | 0.038 | 0.028 | 0.025 |
| MC12 | -0.002 | | 0.361 LR | -0.008 | -0.005 | -0.006 |
| MC13 | -0.014 | 0.361 LR | | -0.005 | 0.006 | 0.007 |
| MC14 | 0.038 | -0.008 | -0.005 | | 0.142 LR | -0.010 |
| MC15 | 0.028 | -0.005 | 0.006 | 0.142 LR | | 0.018 |
| MC16 | 0.025 | -0.006 | 0.007 | -0.010 | 0.018 | |
| MC19 | -0.005 | -0.039 | -0.024 | 0.000 | -0.008 | -0.011 |
| MC20 | -0.008 | -0.021 | -0.027 | -0.007 | -0.019 | 0.017 |
| MC21 | 0.017 | -0.023 | -0.014 | -0.027 | -0.025 | 0.024 |
| MC22 | 0.000 | -0.023 | -0.005 | -0.018 | -0.004 | -0.015 |
| MC23 | -0.005 | -0.026 | -0.027 | -0.024 | -0.023 | 0.008 |
| MC24 | -0.012 | -0.011 | -0.017 | -0.026 | -0.024 | -0.018 |
| MC25 | -0.019 | -0.017 | -0.016 | -0.015 | -0.021 | -0.022 |
| MC26 | -0.018 | 0.011 | 0.002 | -0.019 | 0.014 | -0.049 |
| MC27 | -0.014 | -0.017 | -0.018 | -0.009 | -0.004 | -0.008 |
| MC28 | -0.026 | -0.021 | -0.023 | 0.015 | 0.018 | -0.015 |
| MC29 | -0.028 | -0.017 | -0.033 | 0.015 | 0.021 | -0.023 |
| MC30 | 0.000 | -0.017 | -0.012 | 0.016 | 0.003 | -0.025 |
| MC31 | -0.019 | -0.004 | -0.011 | 0.006 | 0.003 | 0.000 |
| MC32 | -0.013 | -0.028 | -0.025 | -0.007 | -0.002 | 0.001 |
| MC33 | -0.001 | -0.024 | -0.028 | -0.001 | 0.004 | -0.030 |
| MC34 | -0.011 | -0.039 | -0.019 | -0.009 | 0.003 | 0.016 |
| OE17 | 0.012 | -0.006 | -0.038 | 0.007 | 0.002 | 0.017 |
| OE18 | 0.009 | 0.007 | -0.007 | 0.005 | 0.003 | 0.010 |
| OE35 | -0.024 | -0.015 | 0.003 | -0.007 | 0.001 | -0.017 |
| OE36 | -0.013 | -0.028 | -0.021 | -0.014 | -0.009 | -0.012 |
| # of MR | 0 | 0 | 0 | 2 | 0 | 0 |
| # of LR | 0 | 1 | 1 | 1 | 1 | 0 |

| | MC19 | | MC20 | MC21 | | MC22 | MC23 | MC24 |
|---|---|---|---|---|---|---|---|---|
| MC1 | 0.055 | MR | 0.006 | -0.031 | | -0.010 | -0.003 | 0.002 |
| MC2 | -0.005 | | -0.012 | 0.014 | | 0.002 | 0.002 | 0.004 |
| MC3 | 0.001 | | -0.008 | 0.078 | MR | -0.001 | 0.009 | 0.013 |
| MC4 | 0.011 | | 0.014 | -0.044 | | -0.018 | 0.012 | -0.013 |
| MC5 | 0.016 | | -0.007 | 0.011 | | -0.012 | -0.010 | -0.012 |
| MC6 | -0.002 | | -0.001 | -0.010 | | 0.007 | 0.021 | -0.001 |
| MC7 | -0.013 | | -0.017 | 0.012 | | -0.013 | 0.008 | 0.008 |
| MC8 | -0.016 | | 0.000 | 0.009 | | -0.012 | -0.008 | -0.008 |
| MC9 | 0.000 | | -0.005 | 0.013 | | 0.004 | 0.012 | -0.006 |
| MC10 | -0.001 | | -0.002 | -0.011 | | 0.002 | 0.007 | -0.009 |
| MC11 | -0.005 | | -0.008 | 0.017 | | 0.000 | -0.005 | -0.012 |
| MC12 | -0.039 | | -0.021 | -0.023 | | -0.023 | -0.026 | -0.011 |
| MC13 | -0.024 | | -0.027 | -0.014 | | -0.005 | -0.027 | -0.017 |
| MC14 | 0.000 | | -0.007 | -0.027 | | -0.018 | -0.024 | -0.026 |
| MC15 | -0.008 | | -0.019 | -0.025 | | -0.004 | -0.023 | -0.024 |
| MC16 | -0.011 | | 0.017 | 0.024 | | -0.015 | 0.008 | -0.018 |
| MC19 | | | 0.016 | 0.019 | | 0.002 | 0.005 | 0.021 |
| MC20 | 0.016 | | | -0.033 | | 0.016 | -0.007 | 0.012 |
| MC21 | 0.019 | | -0.033 | | | 0.029 | 0.008 | 0.016 |
| MC22 | 0.002 | | 0.016 | 0.029 | | | 0.024 | 0.006 |
| MC23 | 0.005 | | -0.007 | 0.008 | | 0.024 | | 0.013 |
| MC24 | 0.021 | | 0.012 | 0.016 | | 0.006 | 0.013 | |
| MC25 | -0.007 | | -0.004 | 0.016 | | 0.001 | 0.020 | 0.017 |
| MC26 | -0.010 | | 0.022 | -0.027 | | -0.012 | 0.007 | 0.000 |
| MC27 | 0.007 | | 0.035 | 0.018 | | 0.011 | 0.008 | 0.001 |
| MC28 | -0.010 | | 0.047 | -0.001 | | 0.002 | -0.012 | -0.011 |
| MC29 | -0.007 | | 0.041 | -0.025 | | 0.013 | 0.017 | 0.001 |
| MC30 | 0.018 | | 0.011 | -0.001 | | 0.005 | 0.009 | 0.000 |
| MC31 | 0.032 | | 0.022 | 0.006 | | -0.001 | 0.014 | -0.012 |
| MC32 | 0.024 | | -0.007 | 0.012 | | -0.001 | 0.002 | 0.012 |
| MC33 | 0.007 | | -0.010 | 0.010 | | 0.016 | 0.003 | 0.022 |
| MC34 | 0.001 | | -0.006 | 0.006 | | -0.001 | 0.001 | -0.008 |
| OE17 | -0.012 | | 0.004 | -0.016 | | 0.006 | -0.015 | 0.005 |
| OE18 | -0.018 | | -0.013 | -0.016 | | -0.010 | -0.016 | -0.008 |
| OE35 | 0.005 | | -0.016 | -0.019 | | 0.006 | -0.008 | 0.024 |
| OE36 | -0.013 | | -0.004 | -0.002 | | -0.004 | -0.002 | 0.001 |
| # of MR | 1 | | 0 | 1 | | 0 | 0 | 0 |
| # of LR | 0 | | 0 | 0 | | 0 | 0 | 0 |

| | MC25 | MC26 | MC27 | MC28 | MC29 | MC30 |
|---|---|---|---|---|---|---|
| MC1 | -0.021 | -0.001 | -0.012 | 0.012 | -0.003 | 0.001 |
| MC2 | -0.022 | -0.048 | -0.025 | -0.041 | -0.043 | -0.015 |
| MC3 | -0.004 | -0.033 | 0.005 | -0.033 | -0.046 | -0.017 |
| MC4 | -0.033 | 0.017 | 0.001 | 0.023 | 0.036 | -0.005 |
| MC5 | -0.001 | -0.039 | 0.012 | 0.003 | -0.009 | -0.010 |
| MC6 | -0.003 | 0.004 | 0.009 | 0.013 | 0.002 | -0.014 |
| MC7 | -0.005 | -0.020 | -0.018 | -0.037 | -0.026 | -0.037 |
| MC8 | -0.014 | -0.037 | -0.023 | -0.031 | -0.037 | -0.007 |
| MC9 | -0.009 | -0.013 | -0.004 | -0.031 | -0.006 | 0.002 |
| MC10 | -0.012 | -0.027 | -0.009 | -0.001 | 0.011 | -0.008 |
| MC11 | -0.019 | -0.018 | -0.014 | -0.026 | -0.028 | 0.000 |
| MC12 | -0.017 | 0.011 | -0.017 | -0.021 | -0.017 | -0.017 |
| MC13 | -0.016 | 0.002 | -0.018 | -0.023 | -0.033 | -0.012 |
| MC14 | -0.015 | -0.019 | -0.009 | 0.015 | 0.015 | 0.016 |
| MC15 | -0.021 | 0.014 | -0.004 | 0.018 | 0.021 | 0.003 |
| MC16 | -0.022 | -0.049 | -0.008 | -0.015 | -0.023 | -0.025 |
| MC19 | -0.007 | -0.010 | 0.007 | -0.010 | -0.007 | 0.018 |
| MC20 | -0.004 | 0.022 | 0.035 | 0.047 | 0.041 | 0.011 |
| MC21 | 0.016 | -0.027 | 0.018 | -0.001 | -0.025 | -0.001 |
| MC22 | 0.001 | -0.012 | 0.011 | 0.002 | 0.013 | 0.005 |
| MC23 | 0.020 | 0.007 | 0.008 | -0.012 | 0.017 | 0.009 |
| MC24 | 0.017 | 0.000 | 0.001 | -0.011 | 0.001 | 0.000 |
| MC25 | | 0.166 LR | 0.011 | 0.010 | -0.006 | -0.007 |
| MC26 | 0.166 LR | | 0.019 | 0.036 | 0.046 | 0.015 |
| MC27 | 0.011 | 0.019 | | 0.063 MR | 0.047 | 0.022 |
| MC28 | 0.010 | 0.036 | 0.063 MR | | 0.092 MR | 0.035 |
| MC29 | -0.006 | 0.046 | 0.047 | 0.092 MR | | 0.004 |
| MC30 | -0.007 | 0.015 | 0.022 | 0.035 | 0.004 | |
| MC31 | 0.004 | 0.019 | -0.001 | 0.018 | 0.017 | -0.001 |
| MC32 | 0.009 | 0.003 | 0.003 | 0.001 | 0.021 | 0.015 |
| MC33 | 0.039 | 0.007 | -0.008 | -0.009 | 0.001 | 0.000 |
| MC34 | 0.002 | -0.003 | 0.031 | 0.018 | 0.013 | 0.003 |
| OE17 | -0.022 | -0.002 | -0.028 | -0.031 | -0.013 | -0.004 |
| OE18 | -0.002 | -0.014 | -0.034 | -0.013 | -0.025 | -0.014 |
| OE35 | -0.003 | 0.006 | -0.005 | 0.001 | -0.003 | 0.022 |
| OE36 | -0.005 | -0.006 | -0.012 | 0.011 | 0.015 | 0.018 |
| # of MR | 0 | 0 | 1 | 2 | 1 | 0 |
| # of LR | 1 | 1 | 0 | 0 | 0 | 0 |

| | MC31 | MC32 | MC33 | MC34 | |
|---|---|---|---|---|---|
| MC1 | 0.027 | -0.002 | -0.010 | -0.035 | |
| MC2 | -0.022 | 0.000 | -0.013 | -0.020 | |
| MC3 | -0.030 | -0.003 | 0.012 | 0.007 | |
| MC4 | 0.023 | -0.008 | -0.010 | -0.025 | |
| MC5 | -0.009 | -0.013 | -0.003 | 0.050 | MR |
| MC6 | -0.006 | 0.006 | -0.006 | -0.011 | |
| MC7 | -0.031 | -0.020 | 0.011 | -0.008 | |
| MC8 | -0.007 | -0.017 | -0.013 | -0.018 | |
| MC9 | -0.026 | -0.006 | -0.003 | 0.017 | |
| MC10 | -0.016 | -0.005 | 0.008 | -0.016 | |
| MC11 | -0.019 | -0.013 | -0.001 | -0.011 | |
| MC12 | -0.004 | -0.028 | -0.024 | -0.039 | |
| MC13 | -0.011 | -0.025 | -0.028 | -0.019 | |
| MC14 | 0.006 | -0.007 | -0.001 | -0.009 | |
| MC15 | 0.003 | -0.002 | 0.004 | 0.003 | |
| MC16 | 0.000 | 0.001 | -0.030 | 0.016 | |
| MC19 | 0.032 | 0.024 | 0.007 | 0.001 | |
| MC20 | 0.022 | -0.007 | -0.010 | -0.006 | |
| MC21 | 0.006 | 0.012 | 0.010 | 0.006 | |
| MC22 | -0.001 | -0.001 | 0.016 | -0.001 | |
| MC23 | 0.014 | 0.002 | 0.003 | 0.001 | |
| MC24 | -0.012 | 0.012 | 0.022 | -0.008 | |
| MC25 | 0.004 | 0.009 | 0.039 | 0.002 | |
| MC26 | 0.019 | 0.003 | 0.007 | -0.003 | |
| MC27 | -0.001 | 0.003 | -0.008 | 0.031 | |
| MC28 | 0.018 | 0.001 | -0.009 | 0.018 | |
| MC29 | 0.017 | 0.021 | 0.001 | 0.013 | |
| MC30 | -0.001 | 0.015 | 0.000 | 0.003 | |
| MC31 | | 0.036 | -0.003 | 0.023 | |
| MC32 | 0.036 | | 0.024 | 0.012 | |
| MC33 | -0.003 | 0.024 | | 0.012 | |
| MC34 | 0.023 | 0.012 | 0.012 | | |
| OE17 | -0.018 | -0.011 | 0.001 | -0.031 | |
| OE18 | -0.016 | -0.010 | -0.004 | 0.014 | |
| OE35 | 0.002 | 0.002 | -0.002 | 0.021 | |
| OE36 | 0.013 | 0.004 | 0.014 | 0.025 | |
| # of MR | 0 | 0 | 0 | 1 | |
| # of LR | 0 | 0 | 0 | 0 | |

| | OE17 | | OE18 | | OE35 | OE36 | |
|---|---|---|---|---|---|---|---|
| MC1 | 0.000 | | -0.005 | | -0.004 | -0.014 | |
| MC2 | 0.047 | | -0.030 | | -0.013 | -0.036 | |
| MC3 | -0.009 | | -0.018 | | -0.032 | -0.025 | |
| MC4 | 0.000 | | -0.014 | | -0.006 | -0.004 | |
| MC5 | -0.030 | | -0.009 | | -0.013 | -0.015 | |
| MC6 | 0.002 | | -0.009 | | 0.000 | -0.009 | |
| MC7 | -0.017 | | -0.009 | | -0.018 | -0.007 | |
| MC8 | 0.059 | MR | -0.019 | | -0.029 | -0.028 | |
| MC9 | -0.003 | | -0.015 | | -0.014 | -0.022 | |
| MC10 | 0.003 | | -0.010 | | -0.010 | -0.005 | |
| MC11 | 0.012 | | 0.009 | | -0.024 | -0.013 | |
| MC12 | -0.006 | | 0.007 | | -0.015 | -0.028 | |
| MC13 | -0.038 | | -0.007 | | 0.003 | -0.021 | |
| MC14 | 0.007 | | 0.005 | | -0.007 | -0.014 | |
| MC15 | 0.002 | | 0.003 | | 0.001 | -0.009 | |
| MC16 | 0.017 | | 0.010 | | -0.017 | -0.012 | |
| MC19 | -0.012 | | -0.018 | | 0.005 | -0.013 | |
| MC20 | 0.004 | | -0.013 | | -0.016 | -0.004 | |
| MC21 | -0.016 | | -0.016 | | -0.019 | -0.002 | |
| MC22 | 0.006 | | -0.010 | | 0.006 | -0.004 | |
| MC23 | -0.015 | | -0.016 | | -0.008 | -0.002 | |
| MC24 | 0.005 | | -0.008 | | 0.024 | 0.001 | |
| MC25 | -0.022 | | -0.002 | | -0.003 | -0.005 | |
| MC26 | -0.002 | | -0.014 | | 0.006 | -0.006 | |
| MC27 | -0.028 | | -0.034 | | -0.005 | -0.012 | |
| MC28 | -0.031 | | -0.013 | | 0.001 | 0.011 | |
| MC29 | -0.013 | | -0.025 | | -0.003 | 0.015 | |
| MC30 | -0.004 | | -0.014 | | 0.022 | 0.018 | |
| MC31 | -0.018 | | -0.016 | | 0.002 | 0.013 | |
| MC32 | -0.011 | | -0.010 | | 0.002 | 0.004 | |
| MC33 | 0.001 | | -0.004 | | -0.002 | 0.014 | |
| MC34 | -0.031 | | 0.014 | | 0.021 | 0.025 | |
| OE17 | | | 0.102 | LR | -0.005 | -0.006 | |
| OE18 | 0.102 | LR | | | 0.039 | 0.106 | LR |
| OE35 | -0.005 | | 0.039 | | | 0.092 | MR |
| OE36 | -0.006 | | 0.106 | LR | 0.092 | MR | |
| # of MR | 1 | | 0 | | 1 | 1 | |
| # of LR | 1 | | 2 | | 0 | 1 | |

# VITA

Surname: Goodman Given Names: Dean Palmi

Place of Birth: Edmonton, Alberta, Canada

Educational Institutions Attended:

| | |
|---|---|
| University of Victoria | 1996 to 2001 |
| Simon Fraser University | 1989 to 1995 |

Degrees Awarded:

| | | |
|---|---|---|
| B.A. | Simon Fraser University | 1995 |

UNIVERSITY OF VICTORIA PARTIAL COPYRIGHT LICENSE

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other university, or similar institution, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain by the University of Victoria shall not be allowed without my written permission.

Title of Thesis:

Assessing the Feasibility of Using Item Response Theory to Analyse British Columbia's Foundation Skills Assessment 2000

Author [redacted]

Dean Palmi Goodman

August 23, 2001