ITEM RESPONSE THEORY: THE APPLICATION OF BOTH DICHOTOMOUS AND POLYTOMOUS ITEM RESPONSE MODELS TO A PROVINCIAL EXAM DATA SET

by

Keith Andrew Boughton

B.Sc., University of Victoria, 1996

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

MASTER OF ARTS
in the Department of Psychological Foundations

We accept this thesis as conforming
to the required standard

[redacted]

Dr. J. O. Anderson, Supervisor (Department of Psychological Foundations)

[redacted]

Dr. W. Muir, Department Member (Department of Psychological Foundations)

[redacted]

Dr. M. Hunter, Outside Member (Department of Psychology)

[redacted]

Dr. L. Yore, External Examiner (Department of Social and Natural Sciences)

Supervisor: Dr. J. O. Anderson

ABSTRACT

This study was conducted to investigate the feasibility of using binary logistic models with the dichotomously scored items (multiple choice) and partial credit models with the polytomously scored items (open-ended) from the January administration of the 1996 British Columbia grade 12 Provincial Math Examination. The findings in regard to feasibility show a need for better ways of testing for unidimensionality. The item-fit statistics showed the 2-parameter (2PL) and 3-parameter logistic (3PL) models as better fitting models than the 1-parameter (1PL) model with the multiple choice section of the exam. The 1-parameter partial credit model (1PCM) from the RUMM software program (Sheridan, Andrich, & Luo, 1997) showed 1 of 8 items as fitting the model in the open-ended section of the exam. Invariance of parameter estimates was supported for the 1PL, 2PL, and 3PL models. The consistency of student classification results suggested that the 1PL/1PCM model combination had closer estimates to the Provincial Exam raw scores than the 2PL/2PCM model combination. The item characteristic response functions (ICRFs) from the 1PCM were found to be very meaningful for the interpretation of how the item categories were operating within the open-ended section of the exam. Lastly, it should be noted that the different software programs housing the IRT models did not produce equivalent estimates and therefore, the same software should be used in order to have comparable results. The above suggests that it may be feasible to use IRT models with the Provincial Math Examinations.

Examiners:

[redacted]

Dr. J. O. Anderson, Supervisor (Department of Psychological Foundations)

[redacted]

Dr. W. Muir, Department Member (Department of Psychological Foundations)

[redacted]

Dr. M. Hunter, Outside Member (Department of Psychology)

[redacted]

Dr. L. Yore, External Examiner (Department of Social and Natural Sciences)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

The completion of my thesis could not have been accomplished without the help and support of several people. First, I would like to acknowledge the positive improvements made to my thesis by my committee members, Dr. W. Muir, Dr. M. Hunter, and Dr. L. Yore. Their conscientious insights and expertise were invaluable to the progression of not only my thesis, but to my personal understanding of the process of the scientific inquiry.

I would also like to acknowledge the personnel from the Evaluation and Analysis Branch within the Ministry of Education for taking the time to cohesively impart their knowledge of how Provincial Examinations are developed, implemented, and analyzed. I would especially like to thank Ross Norrington for his unique ability to simplify the complexity of the Provincial Examination process and his perseverance in explaining all parts of the process to me. His enthusiasm for the field was most appreciated and contagious.

Most importantly, I would like to acknowledge the endless contributions made to my thesis and to my personal development in the field of measurement by my supervisor, Dr. John O. Anderson. He has given me insurmountable support, direction, and opportunity to advance in the field, and I could not thank him enough for the wisdom and professional integrity he has demonstrated in the last two years.

To my family and friends, I wish to thank them for their encouragement and support. A special thanks to Tamara for being so receptive to my incessant verbosity and for her many late nights of editing.

DEDICATION

To the memory of my mother Nancy Veronica Boughton

1943-1992

Chapter One

Introduction

The purpose of a school leaving exam is to obtain an estimation of student achievement as it relates to expected learning outcomes for a particular course and to determine graduation status. These examinations are often paper and pencil format and are composed of multiple choice and open-ended items. The examinations are usually created by a group of teachers or a committee consisting of individuals familiar with the particular subject area of interest. A table of specifications developed by the committee generally ensures that exam items are balanced for content and are related to instructional objectives. Traditionally, each exam created in this manner was intended to be used only once with an entire population and then to be released into the public domain (Anderson, Muir, Bateson, Blackmore, & Rogers, 1990).

The British Columbia Policy, Evaluation and Analysis Branch within the Ministry of Education currently produces Grade 12 Provincial Examinations in over 15 subject areas for 5 administrations each year. A large amount of time and money is used to create new exams for each administration. Because of this the Ministry has investigated a system of item storage and retrieval called item banking. With recent developments in item banking software, it is possible to create large item banks in which previously used exam items (pretested) can be stored, selected, and re-used to create new examinations. This process of item banking could reduce the time and money spent on new examinations and item creation. Another advantage of using pretested items is that the item response data can be analyzed with various procedures to detect poor items which may then be reviewed or removed from the item bank. The procedure currently used for

exam development involves committees of teachers and classical test theory (CTT) based analyses and is implemented after an exam has been administered. It should be noted that there is currently no pilot testing of the exam items. The last advantage of the item banking process is that the items in the item bank can have their characteristics such as item difficulty and discrimination attached to them within the item bank which allows for more efficient item selection and exam equating. Under the current system of exam analysis (CTT) the item difficulty is an index of how easy or difficult an item is and is based on the proportion of examinees who answer the item correctly. Item discrimination is an index of how well an item separates out low and high ability students. This index is based on a correlation between examinee performance on a test item compared to the total test score. Items with high discriminations are desired by test builders in order to better estimate an examinee's ability (Crocker & Algina, 1986). In regards to the CTT based analysis, Hambleton, Swaminathan, and Rogers (1991) state that,

> whether an item is hard or easy depends on the ability of the examinees being measured, and the ability of the examinees depends on whether the test items are hard or easy! Item discrimination and test score reliability and validity are also defined in terms of a particular group of examinees. Test and item characteristics change as the test context changes. Hence, 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 examines (p. 3).

Another procedure that could be used in the place of CTT is item response theory (IRT), which has some advantages over CTT and is a focus for this study. The first advantage of

IRT is that it would allow one to compare the performance of examinees who have taken different tests on the same score scale and would also allow the performance of different groups to be compared (Crocker & Algina, 1986). Mislevy and Bock (1990) state that,

> unlike classical test theory, IRT does not in general base the estimate of the respondent's ability (or other attribute) on the number-correct score. The only exception is the one-parameter logistic model, in which the estimate is a non-linear function of that score. To distinguish IRT scores from their classical counterparts, we refer to them as "scale" scores. The main advantages of scale scores are that they (1) remain comparable when items are added to or deleted from the tests, (2) weight the individual items optimally according to their discriminating powers, (3) have more accurate standard errors, (4) provide more flexible and robust adjustments for guessing than the classical corrections, and (5) are on the same continuum as the item locations (p. 12).

These advantages of using scale scores will be discussed in detail in subsequent chapters. Once IRT has been applied to the test data, parameter and ability estimates can be obtained and statistical equating methods can be used to produce comparable scores across different groups and tests in an item banking situation. Secondly, if the Provincial exams moved into the area of computerized adaptive testing[1], IRT is the only method that could be employed because the IRT estimates of ability do not require the same items to be presented to every examinee. This means that IRT can be used to rank all examinees

[1] Computer administered exams in which items are administered based on examinee performance to previous items.

on the same score continuum even if different items are presented to different examinees (Wainer, 1990). However, before this can be done, scaling methods are needed in order to link the score scale of IRT to the already established Ministry's letter grade scale.

Currently, the Ministry uses the raw scores (number correct score) from every examinee to calculate final grades. The number correct score is modified by standard setting committees which set cut-points in order to transform the Ministry-approved letter grade scale (A = 86-100%, B = 73-85%, C+ = 67-72%, C = 60-66%, C- = 50-59%, and F = 0-49%) (Province of British Columbia, 1994). It should be noted that in contrast to the Ministry scale, IRT uses a theta scale for examinee calibration. The theta scale is generally set at a mean of 0, a standard deviation of 1, and usually ranges from -3 to 3. Therefore, in order to use an IRT model with the current Provincial Exam system, the scores should be re-scaled to the already established and familiar Ministry letter grade scale. This topic of scaling will be elaborated at a later stage.

This study investigated the use of an IRT based analysis with a Grade 12 Provincial Math response data set which consists of both dichotomous (multiple choice) and polytomous (open-ended) items. This was an exploratory study that made comparisons between IRT/CTT parameter estimates, IRT/IRT parameter estimates (from different software programs), dichotomous/polytomous item response model information and the different software programs that were capable of analyzing the item response data. Most IRT applications have been developed for use with dichotomous data and applications of polytomous IRT models have not been widespread (Muraki, 1993). It is thus important to investigate to what extent the results of an IRT analysis would be interpretable when a polytomous model such as the partial credit model (Masters, 1990)

is applied to the polytomous items. If these IRT applications are found to produce meaningful results, they may be used to facilitate the application of item banks to develop and equate tests that use both multiple choice and open ended items.

Purpose

The purpose of this study was to investigate the feasibility of using binary logistic models with the dichotomously scored items and partial credit models with the polytomously scored items from the January administration of the 1996 British Columbia grade 12 Provincial Mathematics Examination. The investigation of feasibility consisted of goodness of fit, parameter invariance, and consistency of classification studies. Added to the study were comparisons of parameter estimates from the software programs PARSCALE (Muraki & Bock, 1996), BILOG (Mislevy & Bock, 1990), and RUMM (Sheridan, Andrich, & Luo, 1997); comparisons across different models of IRT, and the types of information coming from the open-ended items (item characteristic response functions). These software programs were chosen for this study based on their accessibility. In this study feasibility will be regarded as the practicality of using IRT in an item banking situation to replace the current CTT method and standards setting. The consistency of classification study compares IRT examinee estimates with the current CTT examinee estimates used by the Ministry. If the estimates differ to a large degree then it would not be feasible to use IRT methods in place of CTT methods because the measurement tools would then not be consistent. With an exam system already in place, the achievement estimates of future examinees through IRT must be comparable with past estimates. The ideal situation would then be if the IRT estimates were found to be comparable to the current measurement system while maintaining parameter

invariance. This consistency is necessary because the ability estimates from CTT methods and standards setting are the only estimates of ability with which to compare the IRT estimates. In order to obtain all the benefits from the IRT models[2], the goodness of fit and the parameter invariance assumption underlying the IRT models must be met. Therefore, an evaluation of goodness of fit and parameter invariance will be employed in this study. The advantages of IRT can usually only be obtained with a model that fits the data set of interest. If the model fits the data set, then item and ability invariance should be maintained (Hambleton & Swaminathan, 1985). All of these assumptions have methods of detection and will be employed in this investigation. The feasibility issue then depends on whether or not consistency of classification, goodness of fit, and parameter invariance have been met. This is important in order to be able to use item banks and IRT models to create new examinations for each administration, and to accurately estimate examinee and item parameters. If accurate estimates are obtained and the IRT model assumptions are met, then it would be possible to compare and equate the exam results from one administration to another in an item banking situation using IRT.

[2] Mathematical expression giving the probability of a correct response to a test item as a function of the ability of the person responding.

Chapter Two

Literature Review

The scoring, analysis and reporting of B.C. Provincial Examinations currently uses raw scores and classical test theory-based analyses. Another method that could be employed is item response theory which would allow one to compare the performance of examinees who have taken different tests on the same score scale and would facilitate the performance of different groups to be compared (Crocker & Algina, 1986). Item banks, especially those using item response theory then have considerable potential over classical test theory. This chapter reviews both the current system of standard setting practices using CTT along with IRT and its assumptions.

Classical Test Theory

"In classical test theory, a test is regarded as a sample of items from a domain defined by generating rules or by content, process, and format specifications. If the items are a random sample of the domain, then the percent-correct score on the test estimates the domain score, that is, the expected percent correct for all items in the domain" (Bock, Thissen, & Zimowski, 1997, p.197). Within this framework, testing tries to measure a single underlying trait called achievement in regard to the Provincial examination process. Each individual possesses some degree of this underlying trait and the objective of measurement is to estimate the amount. The true score is the actual amount of the trait the examinee possesses. If an attempt is made to estimate the true score by administering several tests to an examinee, then several different observed scores will be obtained. The spread of observed scores is centered around the true score (mean of observed scores) and

is dependent on the error associated with that measure. Error then is the difference between the observed score and the true score.

True score = observed score ± error

Estimation of the amount of error is critical in order to be able to estimate the true score (true underlying ability) of any examinee (Lord, 1980).

Reliability refers to the consistency of scores (assuming a stable trait) from one measurement to another and within a measure itself. Extraneous factors such as attention, memory, effort, fatigue, emotional strain, and guessing introduce a certain amount of random measurement error into all assessment results (Linn & Gronlund, 1995). Measurement error includes effects on observed scores that occurred because the testing procedure depended on a particular set of questions, the time at which the test was given, or the person who scored the test. Reliability represents estimates of the error present in the observed scores. The greater the error, the more unreliable the observed scores become. Error estimates can be reported in terms of a "standard error of measurement" and can be computed from the standard deviation and the reliability coefficient. The equation for the standard error (*SEM*) is,

$$SEM = SD\sqrt{1-r} \qquad (1)$$

where *SD* is the standard deviation for a group's observed test scores and r is the reliability coefficient. Reliability coefficients are determined by several different methods such as, split-half, test re-test, equivalent forms, Kuder-Richardson, and inter-rater reliability methods (Livingston, 1988). It should be noted that inter-rater reliability cannot be used with equation 1.

Item Response Theory

Item response theory postulates that with any examinee/test item interaction there is an underlying ability or proficiency level that influences performance on that item as well as the item characteristics. The examinee performance then is a function of both item and person characteristics. An estimation of this ability level will allow for prediction of subsequent performance. IRT models the relationship between an examinee's score and the unobservable underlying ability, and is described in terms of a mathematical function. The relationship between ability and item performance can be described by a monotonically increasing function called an item characteristic function (ICF) illustrated below,

$$P_i(\Theta) = \frac{1}{1 + e^{-1.7(\Theta - b_i)}} \qquad (2)$$

where $P_i(\Theta)$ is the probability that an examinee with a certain ability or theta (denoted as Θ) will respond correctly to an dichotomously scored item i with a difficulty of b_i and a scaling factor of 1.7. This function specifies that an examinee who has a high ability level will also have a greater probability of answering that item i correctly (if b_i increases then $P_i(\Theta)$ decreases). If there is only one common ability being considered on a particular test (unidimensional), then the ICF is called an item characteristic curve (ICC) and is shown in Figure 1. This curve provides the probability of getting an item correct at each ability level (Hambleton & Swaminathan, 1985).

Assumptions of item response theory. There are a few assumptions that should be met when choosing an IRT model. The first is unidimensionality which assumes that there is only one underlying trait or ability that can account for the examinee's responses

to a test. Crocker and Algina (1986) state that, "...a test is unidimensional if its items are statistically dependent in the entire population, and a single latent trait exists such that the items are statistically independent in each sub-population of examinees whose members are homogeneous with respect to the latent trait" (p. 343). Most test builders desire a unidimensional item pool in order to have a clear understanding of what the test scores mean. Nandakumar (1994) compared three methodologies for assessing unidimensionality. These methodologies were DIMTEST (Stout, Douglas, Junker, & Roussos, 1993), Holland and Rosenbaum's approach, and nonlinear factor analyses. All three approaches were found to be equivalent in the detection of unidimensionality for the simulated data in the study. Nandakumar (1991) stated that DIMTEST uses Stout's index to test the hypothesis of one versus more than one "essential dimension," when minor dimensions may be present. Stout's "essential unidimensionality" is different from the notion of unidimensionality alone. Hattie, Krawowski, Rogers, and Swaminathan (1996) state that, "the Stout procedure is based on the weaker principle of local independence and aims not to identify whether a set of items is or is not unidimensional, but whether there is a sufficiently dominant dimension such that the test user can proceed to meaningfully interpret a single total score across a set of items" (p. 13). Hattie et al. (1996) stated that Stout's procedures (within DIMTEST) were found to be easy to use, dependable, and reasonably robust in the assessment of essential unidimensionalty. DIMTEST was chosen for this study based on accessibility and its power in detecting dimensionality when compared to Holland and Rosenbaum's procedures and nonlinear factor analysis. Although DIMTEST and factor analysis are both being used to conduct the analysis, multidimensional scaling (MDS) was also used for the assessment of

unidimensionality (see Kruskal & Wish, 1978). This analysis comes in part because of the Oltman, Stricker, and Barrows (1990) article that states, "multidimensional scaling [MDS] appears to be a useful method for item-level analyses of test structure" (p. 21). Also, MDS is a "type" of nonlinear factor analysis which is similar to IRT.

Local independence is the next assumption that is related to but must not be confused with unidimensionality. This assumption states that an examinee's responses to different items must be statistically independent except for, in this case, the underlying trait called achievement. Also, local independence involves the notion that items should not be used that require a correct response to a previous item. If this is met then the responses to each item are explained only by the underlying ability or achievement.

The final assumption of the IRT models is that of speededness and is related to unidimensionality. It is necessary that all examinees have enough time to try all the items so that their ability level affects their responses and not the failure to reach an item. It is implicit in the unidimensional assumption that only one ability is being measured and not a second ability of speed in answering a question (Hambleton & Swaminathan, 1985).

Item-fit statistics. Item-fit is the extent to which the actual student responses match the predicted or modeled responses from an IRT analysis. Chi-square statistics can be used to report the results. It is important to assess model-data fit in order to ensure that the most appropriate model is used to analyze the data (Lord, 1980). Also, an item characteristic curve of item-fit can be plotted in order to easily view how well the model fits the response data.

There are some important features of interest when an item response model fits the data set and the pool of items all measure one underlying ability. The first is that the

item parameter estimates obtained from the application of IRT models are independent of the sample of examinees to which a test is administered (item invariance). Also, any set of items can be chosen from an item pool and the examinee estimates obtained are independent of that item selection (ability invariance) as long as local independence holds (Wainer, 1990).

Item response models. The one-parameter logistic model (or Rasch model) is an IRT model which assumes that all items have equal discriminating power and that guessing is minimal. The single parameter that is estimated is then the difficulty or b-parameter of an item. The equation for the 1-parameter logistic model (1PL) is

$$P_i(\Theta) = \frac{1}{1 + e^{-1.7a(\Theta - b_i)}} \qquad (3)$$

where $P_i(\Theta)$ is the probability that an examinee with a certain ability or theta (denoted as

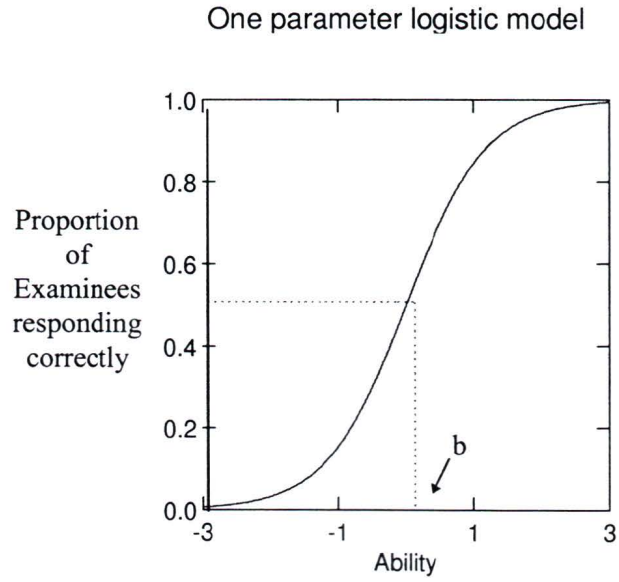


Figure 1. Item Characteristic Curve (1PL)

Θ) will respond correctly to a dichotomously scored item i (item scored either right or wrong), with a difficulty denoted as b_i and a scaling factor of 1.7. The ICC for this model is shown in Figure 1. The difficulty index (b_i) is the point on the ability scale where an examinee has a 50% probability of answering the item i correctly. The discrimination or a parameter is proportional to the slope of the ICC at index b_i on the theta scale. However, the discrimination is not free to vary for the 1PL model. The scaling factor of 1.7 was introduced in order for the logistic function to behave as close to the normal ogive function (based on the cumulative normal distribution) as possible. "Logistic functions have the important advantage of being more convenient to work with then the normal ogive functions" (Hambleton, Swaminathan, & Rogers, 1991, p.14). The curved line in Figure 1 represents, at each ability level, the probability of a correct response to a multiple choice item or any item that is scored either correct or incorrect. The point on the ability scale below the point of inflection at $P_i(\Theta)=0.5$ is referred to as item difficulty. It should be noted that the item difficulty is reported on the same scale as ability which allows for a direct comparison between person and item parameters. The 1PL model gives the probability of answering an item correctly in terms of the interaction between item difficulty (b_1) and examinee ability (Θ). If the item difficulty was increased in Figure 1, then the curve would shift to the right and the probability of a correct response would be reduced for a given Θ. To better model some types of response data, a second model was introduced called the 2-parameter logistic model which allows for differentially discriminating items (Hambleton & Swaminathan, 1985).

The two-parameter logistic model (2PL) is a more general model of the 1-parameter and has item characteristic curves that vary in both difficulty (b_i) and

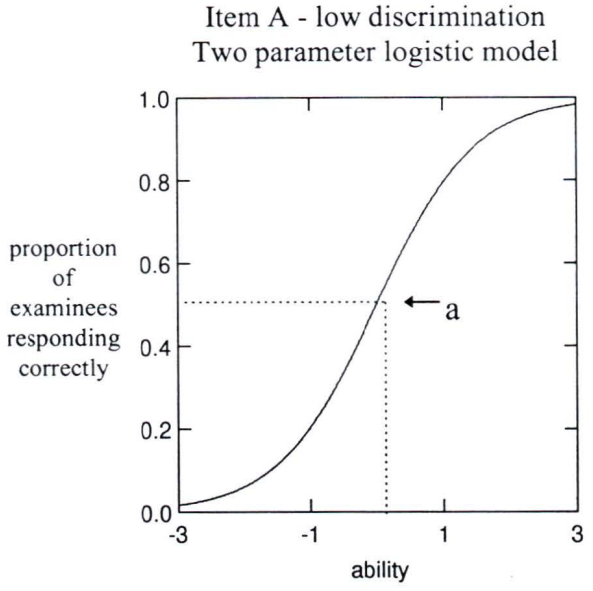


Figure 2. Item A Characteristic Curves (2PL)

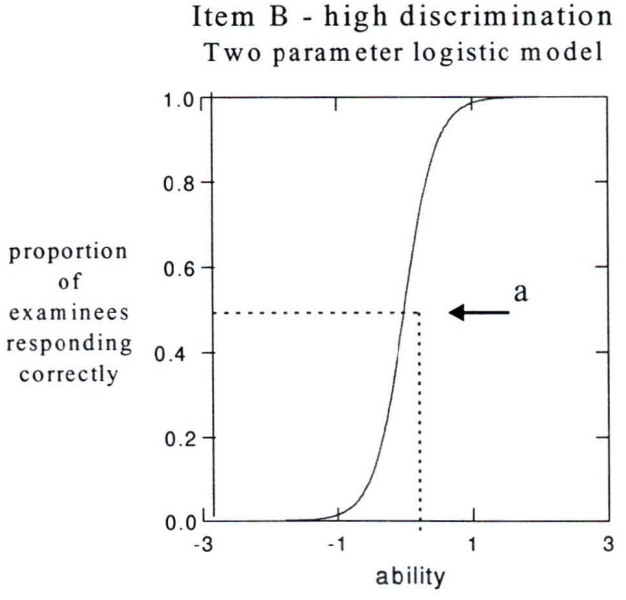


Figure 3. Item B Characteristic Curves (2PL)

discrimination (a_i) parameters. The discrimination is the slope of the item characteristic curve. The discrimination or a_i parameter is proportional to the slope of the ICC at index b_i on the theta scale. The mathematical model is

$$P_i(\Theta) = \frac{1}{1 + e^{-1.7a_i(\Theta - b_i)}} \qquad (3)$$

where $P_i(\Theta)$ is the probability that an examinee with a certain ability or theta (Θ) will respond correctly to an dichotomously scored item i with a difficulty (b_i), discrimination (denoted as a_i) and a scaling factor of 1.7. Figures 2 and 3 show ICCs for item A and item B for the 2PL model with the discrimination (a_i) index labeled. Item A has a discrimination index of 0.8 and Item B has a discrimination index of 1.0. The steeper slopes signify that the item is more highly discriminating and thus will allow for better separation of examinees with different ability levels. This in turn will allow for more precise measurement of an examinee's ability level. For example, item B is more highly discriminating (greater slope) then item A and therefore has more power to distinguish between a low ability examinee and a high ability examinee (Hambleton & Swaminathan, 1985).

The three-parameter logistic model (3PL) was introduced in order to account for the probability of a lower ability examinee obtaining the correct response to a hard item. If an examinee of low ability got a hard item correct there may be another dimension or skill that allowed the examinee to answer that item. This multidimensionality problem can be accounted for by removing the item or by including a guessing or lower asymptote parameter (denoted as c_i) into the mathematical model. The equation for the 3-parameter model is

$$P_i(\Theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\Theta - b_i)}} \qquad (4)$$

where $P_i(\Theta)$ is the probability that an examinee with a certain ability or theta (denoted as Θ) will respond correctly to a dichotomously scored item i with difficulty (b_i); discrimination (a_i); lower asymptote (denoted as c_i); and a scaling factor of 1.7. The ICC for the 3PL model takes this form in Figure 4. The lower asymptote (c_i) or pseudo guessing parameter is the binomial floor on the probability of getting an item correct and in this case the probability of correctly responding never falls below (0.2). The 1PL, 2PL, and 3PL models are all dichotomous item response models (Hambleton & Swaminathan, 1985).

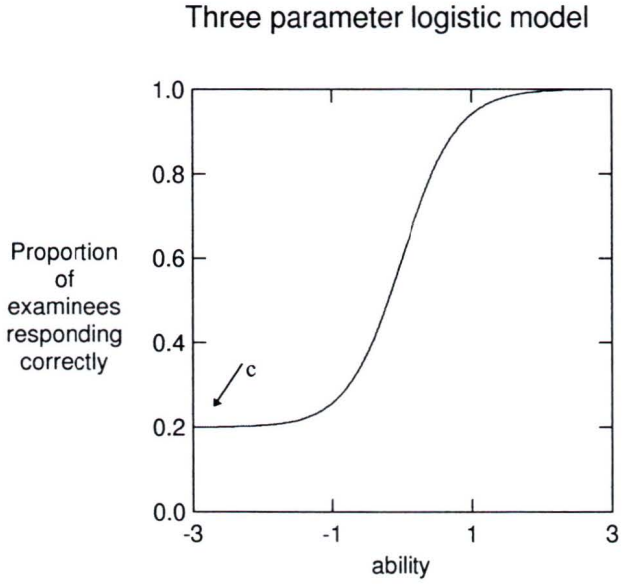


Figure 4. Item Characteristic Curve (3PL)

Partial credit models. Master's (1982) partial credit model (1PCM) was developed from the 1-parameter or Rasch model (1PL) but is generalized for use with unidimensional polytomous items (scored data with more than two categories). Dichotomous data are scored as right or wrong (0,1) and have two categories (e.g., multiple choice items) while polytomous items are scored with partial credit (0, 1, 2, etc.) given for partial success on a item (e.g., open-ended items). This partial credit model, like the 1PL model, has one item parameter which is the b-parameter(s), although in this case the b-parameters are called the item step parameters and are the points on an item category response function (ICRF) where the two categorical responses intersect (Figure 5). Both person and item step parameters are locations on the ability or theta scale. "In the [partial credit models] the [step parameter] b_{ih} may be additively decomposed as $b_{ih} = b_i - d_h$ in the same manner as in Andrich's (1978) rating scale model. The values of d_h are not necessarily ordered sequentially within an item. The parameter d_h is interpreted as the relative difficulty of category h in comparing other categories within an item or the deviate of each categorical threshold from the item location, b_i" (Muraki, 1997, p. 95) The subscript h signifies a category while the subscript i signifies the item. The item response is separated into categories which can be ordered from level 0 to level 3 if there is only four categories (Masters & Wright, 1997). In Figure 5 for example, the intersection between category P0 and P1 is b_{ih} (labeled b_1 in Figure 5)= -2.0 ; the intersection between category P1 and P2 is b_{ih} (labeled b_2 in Figure 5)= 0, and the intersection between category P2 and P3 is b_{ih} (labeled b_3 in Figure 5)= 2. Masters (1990) states that partial credit models can be applied to multi-step problems, "...which are common in subject areas like mathematics and the physical sciences where students must

first identify the problem type, select an appropriate solution strategy, and then apply this strategy which may itself involve a number of steps" (p. 391). These types of problems are designed to address a wide range of ability levels and to differentiate between the examinees responses to the open-ended items. The Provincial math exam uses this type of rating scale on the open-ended items and therefore was appropriately used with the Provincial exam data in this study.

Muraki's (1992) generalized partial credit model (2PCM) is also a unidimensional item response model for polytomous data. The difference between the 1PCM and the 2PCM is that the discriminating or slope parameter (a_j) is estimated. The equation for Muraki's 2PCM is

$$P_{jk}(\Theta) = \frac{\exp\left[\sum_{v=0}^{k} Da_j(\Theta - b_{jv})\right]}{\sum_{c=0}^{mj} \exp\left[\sum_{v=0}^{c} Da_j(\Theta - b_{jv})\right]} \qquad (5)$$

where $P_{jk}(\Theta)$ is the probability that an examinee of a given ability Θ will obtain a score of k on item j; b_{jv} is the step parameter (step v for item j) and marks the point on the ability scale for which two categories are equally likely (for example, b_{jv} in Figure 5 equals -2.0 and marks the intersection between P0 and P1); a_j is the discrimination or slope parameter, and D is a scaling factor of 1.7. With each step parameter being defined by two adjacent categories the interpretation becomes less straightforward in marking out the regions of an underlying variable compared to its dichotomous counterpart. Because of this, the partial credit analyses should be accompanied by an item category response function. The item category response function (ICRF) for a four category item is shown in Figure 5 and should help with the interpretation of examinees' responses. For

example, a math question on an exam might have four possible categorical responses that could each be awarded partial credit. The curve labeled P0 in Figure 5 represents the probability of obtaining a score of 0 for a given value of Θ and indicates that examinees

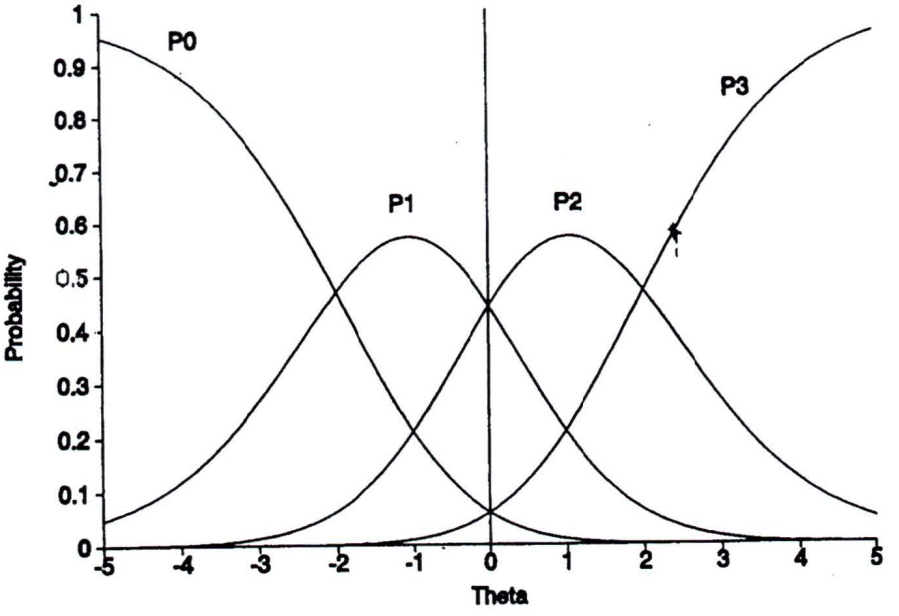


Figure 5. Item Category Response Functions (a=1.0, b_0=0.0, b_1=-2.0, b_2=0.0, b_3=2.0) (Donoghue, 1994, p. 298).

in this category did not obtain the correct answer to any part of the question (note that as the Θ increases, the probability of P0 decreases). P1 would indicate the probability of examinees at a given Θ answering the question correct at this level. For example, if a student had an ability level of -1, then they would have a 0.6 probability of correctly answering the first part of the open-ended item correctly (P1) or obtaining a score of 1, but would also have a 0.2 probability of obtaining a score of 0 or 2. It should be noted that the probability increases with Θ up to a point and then decreases as the probability of

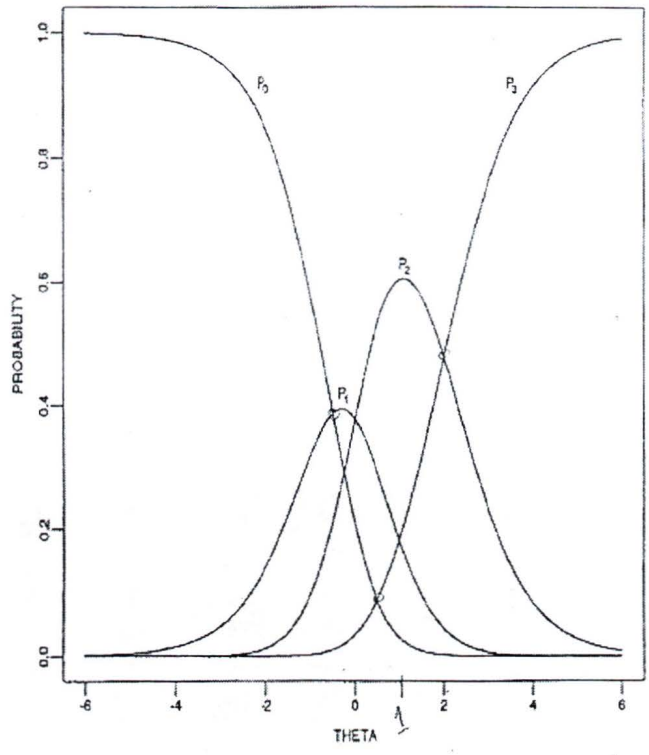


Figure 6. Partial credit model. a=0.7, b=(-0.5,0.0,2.0) (Muraki & Bock,1996, p.21).

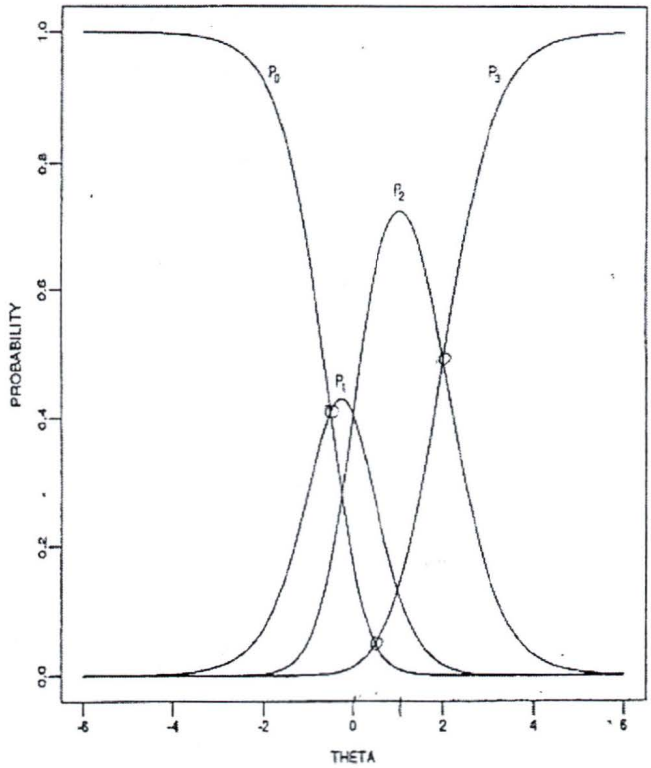


Figure 7. Partial credit model. a=1.0, b=(-0.5,0.0,2.0) (Muraki & Bock,1996, p.20).

a higher category increases. P2 indicates the probability of examinees at a given Θ obtaining the correct answer on the first and second part of the question. Lastly, P3 shows the probability of examinees at a given Θ who answered the question correctly in each category and obtained the correct final answer to the math problem (De Ayala, 1993; Donoghue, 1994; Masters, 1982; Muraki, 1993; Muraki, 1997). Therefore, if an examinee has an ability level of 4, they have a 0.8 probability of answering the question correctly at each step to obtain the correct final answer and a score of 3 (i.e. category P3). If an examinee has a lower ability of -2 (which happens to be the step or b_1 parameter), then they have an equal probability of 0.4 in obtaining either P0 or P1; a probability of 0.2 for either category P0 or P2, and a probability of near zero on category P3.

The step parameters b_0 - b_3 are indicators of the difficulty between steps and reports the ability level for which categories *k* and *k*-1 are equally likely. Thus, the intersection of P0 and P1 in Figure 5 is -2.0 and can be interpreted as a relatively easy step because of the low ability range. An examinee with an ability level just above -2.0 would have a greater probability of obtaining P1 than P0 or more simply would probably obtain the correct answer to the first part of the question. The 2PCM also estimates a slope parameter which indicates the degree to which categorical responses vary among items as the ability level changes (Muraki, 1993). For example, Figures 6 and 7 show two ICRFs with varying slope parameters. Figure 7 has a slope parameter of 1.0, while Figure 6 has a slope of 0.7. Items with steeper slopes (Figure 7) signify that the item is more highly discriminating and thus will allow for better separation of examinees with different levels of ability.

Differences Between CTT and IRT

The scoring, analysis, and reporting of the B.C. Provincial examinations currently uses raw scores and classical test theory-based analysis. The focus of the Provincial exams using CTT is on each test and on that group of students who wrote that test. This group and test dependency poses a measurement problem with the Provincial exams moving towards a system of item banking. Specifically, the item statistics used in CTT will change as the group who wrote the exam changes. There are several reasons why this is true and these will be discussed next.

In CTT, "unbiased assessment of item properties depends on representative samples from the target population" (p. 345). IRT on the other hand has unbiased estimates of item properties (item invariance) which can be obtained from non-representative samples (Embretson, 1996). Hambleton, Swaminathan, and Rogers (1991) state that,

> To some researchers, the property of item invariance may seem surprising. The property however, is a well-known feature of the linear regression model. In the linear regression model, the regression line for predicting a variable *Y* from a variable *X* is obtained as the line joining the means of the *Y* variable for each value of the *X* variable. When the regression model holds, the same regression line will be obtained within any restricted range of the *X* variable, that is, in any subpopulation on *X*, meaning that the slope and intercept of the line will be the same in any subpopulation on *X*. A derived index such as the correlation coefficient, which is not a parameter that characterizes the regression line, is not invariant across

> subpopulations. The difference between the slope parameter and the correlation coefficient is that the slope parameter does not depend on the characteristics of the subpopulation, such as its variability, whereas the correlation coefficient does (note, however, that the proper estimation of the line does require a heterogeneous sample). The same concepts also apply in item response models, which can be regarded as nonlinear regression models (p. 19).

IRT also has the property of ability invariance which means that the examinee ability estimates are not dependent on a particular set of test items provided local dependence holds. Therefore, CTT is test and group dependent while, in theory, IRT is not. The above holds true if the IRT model chosen fits the data exactly in the population (Hambleton et al., 1991). So, if the exam data fits the IRT model, then the resulting item characteristics can be considered invariant and can be used with confidence in an item banking situation.

Standard Setting

Standard setting is the process of setting cutoff scores for letter grade categories. This process is used in testing situations in order for a cutoff score to be set to allow for separation of ability levels into letter grade categories on a particular exam. Currently, the Ministry uses a standard setting committee to set cut-points on the raw score scale (number correct) and then mathematically adjusts each score onto to the Ministry-approved letter grade scale. This scaling procedure will be discussed later on in the chapter.

There are three major categories of standard setting which are: (1) holistic judgments about the examination; (2) content judgments of individual test items; and (3) examinee test-performance-based judgments (Crocker & Algina, 1986). The 'holistic approach' sets standards by gathering a group of experts to review test content and to then set a cut-point based on their evaluation of the whole exam. With this overall view of the test the experts suggest how many items should be answered correctly by a low competency examinee at each particular level of interest. The judgments based on the 'item content approach' uses experts to assess each item individually. At this item by item level, experts would estimate what proportion of a "minimally acceptable" group would answer an item correctly. These proportions are summed for all items and a minimum passing score at a particular level of interest can be obtained. The 'examinee test performance based approach' would administer a test to a group of examinees that a group of experts deemed to be below the ability level being tested. This would allow the group of experts to set a minimum ability criteria by use of the mean or median from this lower ability group. Once the ability level has been established, it could then be used with the target population (Crocker & Algina, 1986). There are several different methods using judgments based on item content, one of which is called the "Extended Angoff method". This item based procedure involves the participation of a standard setting committee to predict how many examinees will attain certain levels of success. This standard setting committee is made up of a group of teachers in the subject area of interest. The extended Angoff asks "...each judge to state the probability that the 'minimally acceptable person' would answer each item correctly. In effect, judges would think of a number of minimally acceptable persons, instead of only one such person, and

would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities would then represent the minimally acceptable score" (Angoff, 1971, p. 515). Standard setting is used by the Ministry for a number of reasons. The first being that the interpretation of scores relies on cut-scores being determined for certain levels of achievement. The second reason is in order to ensure that from year to year exam results are comparable. This is a extensive process by which the Ministry uses the raw scores (number correct score) for every examinee. This is modified by standard setting committees which set cut-points in order to transform the number correct score onto the Ministry-approved letter grade scale.

Currently, the Ministry uses a modified Angoff method that requires two rounds of ratings which are the pre-cut and post-cut points. The 'first round' establishes a set of pre-cut points with each item being first reviewed by the standard setting committee before student results are known. The standard setting committee is usually made up of a group of 2 to 70 teachers depending on the subject area, as well as a marking chair. The marking chair for a particular subject will usually have been involved in the marking process for a number of years before being selected for this position. Prior to recording these cut-score estimates, the standard setting committee will have reviewed the examination individually. This review will consider difficulty level; the wording of the questions; the marking criteria for the written-response questions; the range of ability levels; and the letter grade criteria. The individual cut-points are averaged in order to determine what the "minimum" raw score will be for a student to pass the exam. This process will be repeated for each letter grade (A, B, C+, C, C-, F). The 'second round' is the post cut-score rating at which point the committee will receive the actual results of the

examinees. If at this point any of the items are found to be too easy or too difficult they will be discarded. Low discriminating items are also reviewed as well as items that may have more than one possible answer. At this stage the effects of their original pre-cuts on the examinee population are evaluated. The committee will then discuss the consequences of the current cut-points and then will make a second estimate of each cut-score. Pre-cuts and post-cuts are then reassessed and a final cut point is determined and submitted to the marking chair person. If the marking chair agrees with the final cut point, the cut-points will then be submitted for approval to the Ministry staff in the Assessment and Accountability branch and finally to the Board of Examiners. The Board of Examiners are professionals who represent "people of interest" in education such as the college and university systems. This board oversees and advises on policy decisions which include the standard setting committee. After the cut-points are established and agreement between groups has been met, transformation of the raw scores to the Ministry approved letter grade scale is undertaken by the Ministry staff (Examinations and Assessment Branch, 1992).

Scaling

Scaling is a process by which empirical observations are systematically assigned numbers in meaningful units (Crocker & Algina, 1986). Currently, the Ministry uses a standard setting committee to set cut-points in order to scale the raw scores (number correct) onto to the Ministry-approved letter grade scale.

The Provincial exams use this raw score scale which is not very meaningful or interpretable. In order for the Provincial exams to have interpretable meaning, the raw scores have to be transformed onto the well known Ministry-approved letter grade scale.

For example, a Provincial Math exam is out of a possible 120 points. If a cut-point for the "A" range was decided by the standard setting committee to be 100 out of a possible 120 points, then transformation onto the Ministry scale ("A" range cuts from 86 to 100) would be needed. If an examinee obtained a raw score of 110 on the math exam, this would be transformed onto the Ministry scale. This transformation can be mathematically undertaken by the formula,

$$MinScore = CutPt + \frac{M_c S_r}{T_r},$$

$$MinScore = 86 + \frac{10(14)}{20},$$

$$MinScore = 93,$$

where M_c is the number of raw score points above the cut-point and in this case, 110 is 10 points above the "A" level cut-point of 100. T_r is the cut-point range for the raw score points and S_r is the cut-point range for the Ministry letter grade in question. *CutPt* is the Ministry scale cut-point and *MinScore* is the actual score when transformed onto the Ministry scale. If $M_c S_r / T_r = 7$, then the transformed score on the Ministry scale would be 7 points above 86 which is 93. This transformation allows a comparison of previous exam scores to be made when using the same scale. Angoff (1971) states,

> ...it is important to recognize that raw scores as such have little if any generality, since they are a product of the items contained in the test.... Because of the natural and expected variation in difficulty from form to form, a raw score of given value will not always have the same meaning or represent the same level of ability. The form of the test would have to be specified and its characteristics known and kept in mind by the test

> user. The need to keep track of this additional information can prove to be cumbersome. The solution here is to adopt a reliable system of equating test forms that will make it possible to translate all forms into a common scale. But since, in this case, all but one of the forms would require some adjustment of the raw scores, it would seem less confusing to convert raw scores on all forms to an arbitrary scale that is different from any of the raw score scales" (p. 511-512).

Therefore, raw scores can be transformed to the Ministry-approved letter grade scale and this should increase the interpretability of the scores from the different exams in the same subject area (equating process).

Scaling IRT Estimates

Unlike classical test theory, IRT uses a theta (Θ) scale for score reporting. One of the problems with using the IRT models is that the scaling system is different and therefore not easily interpreted without some prior knowledge and experience. Theta scales generally have a mean of 0 and a standard deviation of 1 and usually range from -3 to 3. If item banks are to be used for the next generation of exams, then IRT methods could be used for item selection and test equating. Before this can occur, IRT analyses must first be investigated for use with Provincial exam data. A problem that arises with the use of IRT when the existing method used CTT is that the scores for the examinees are placed on two different scales and for interpretation purposes the two scores must be placed on a common scale (Wainer, p. 139). For example, in order to compare IRT and CTT (which includes standard setting) there must be a common scale in place. What would it mean if an examinee obtains a score of 90 (an A letter grade) through CTT and

standard setting and a 2.8 through IRT? The scale of preference would then be the more familiar scale of the Ministry. Before this transformation of the theta scale to the Ministry scale can be done, cut-points on the theta scale must be determined from the raw and Ministry-approved letter grade cut-points.

Equating

Test creators are interested in making comparisons of examinee scores across different tests that all measure the same underlying ability. This is important in making selection or pass/fail decisions of examinees who have written different tests and thus involves some sort of equating (Hambleton et al., 1991). The Ministry currently sets standards unique to each exam each year with the intent to equate the exam by adjusting the cut-points before transformation of raw score scale to the Ministry-approved letter grade scale (i.e. an A in Math 12 in January 1996 should mean the same level of achievement as an A in Math 12 in June 1998). Equating one exam with another currently relies on this process of standard setting. One assumption with this process is that from year to year the population has a similar distribution. Specifically, the Ministry does not equate their exams with any formal statistical methods.

When using an IRT model that fits the data set of interest, parameter invariance should hold. Parameter invariance features allow for item statistics that are not dependent on the group of examinees who wrote the exam and the examinee ability estimates are not dependent on the test items within an exam. If the Ministry generates exams from an item bank using IRT procedures, the property of invariance will allow for exam equating that may be attended to more formally through statistics as long as the item parameters

are known (Hambleton & Swaminathan, 1985). An important aspect of IRT is that an evaluation of the equating system can take place.

Summary

The British Columbia Policy, Evaluation and Analysis Branch within the Ministry of Education produces Grade 12 Provincial Examinations in over 15 subject areas for 5 administrations each year. A large amount of time and money is used to create new exams for each administration. Because of this the Ministry has investigated a system of item storage and retrieval called item banking. Previous exams can now be stored in the item bank and new exams can be generated from this bank of items.

Classical test theory (CTT) and raw score analyses are currently being employed by the Ministry for item and person estimation as well as the identification of grade level category cut-points in the attempt to equate each new exam. However, there is a problem that arises with the use of CTT for item difficulty and discrimination estimates when used in an item banking situation. The problem is the item estimates are dependent on the ability level of the group who wrote the exam. This means that, if a new group of students are tested that is quite different in overall ability, then the CTT item estimates will be different. From a measurement point of view, this poses a problem because the identification of the item difficulty estimates become meaningless and cannot be used to create comparable exams. Currently, the Ministry sets standards unique to each exam each year with the intent to equate the exam by adjusting the cut-points before transformation of raw score scale to the Ministry-approved letter grade scale. Equating one exam with another currently relies on this process of standard setting. The use of these raw scores (number correct) also poses a problem. These raw scores are dependent

on the items within a particular exam and this increases the difficulty in comparing examinees who have taken different exams. Therefore, in an item banking situation, expert committees will still be needed to identify raw score cut-points for each letter grade category with the attempt to equate each new exam.

Analysis based on item response theory can also be used for item and person estimates and exam equating. If IRT is used instead of CTT it may be possible to equate the exams without the need for continuous expert committee opinion. Item response theory overcomes the CTT dependency on the group of examinees who wrote the exam and on those items within that exam (Hambleton et al., 1991). This is important in order to be able to use item banks to create new exams for each administration and to accurately estimate examinee and item parameters. If accurate estimates are obtained and IRT model assumptions are met, then it would be possible to compare and equate the exam results from one administration to another in an item banking situation using IRT.

The purpose of this study was to investigate the feasibility of using binary logistic models with the dichotomously scored items and partial credit models with the polytomously scored items from the January administration of the 1996 British Columbia grade 12 Provincial Math Examination. In this study feasibility will be regarded as the practicality of using IRT in an item banking situation in place of the current method of CTT and standard setting. In order to be able to do this one must investigate the goodness of fit, parameter invariance and consistency of classification between IRT and CTT. This is important in order to be able to use item banks and IRT models to create new exams for each administration, and to accurately estimate examinee and item parameters. If accurate estimates are obtained and the IRT model assumptions are met,

then it would be possible to compare and equate the exam results from one administration to another in an item banking situation using IRT. Added to the study were comparisons of parameter estimates from the software programs PARSCALE, BILOG, and RUMM; comparisons across different models of IRT, and the types of information coming from the open-ended items (item characteristic response functions).

Chapter Three

Method

The scoring, analysis and reporting of the B.C. Provincial Examinations currently uses raw scores and classical test theory-based analysis. In order to investigate the feasibility of using IRT logistic based models including partial credit models, the software programs PARSCALE 3 (Muraki & Bock, 1996), BILOG (Mislevy & Bock, 1990), and a Rasch unidimensional measurement model (RUMM) (Sheridan, Andrich, & Luo, 1997) were used to analyze a Provincial response data set. The relationship between the raw scores, Ministry approved letter grade scale, and the proficiency estimates from the IRT based models were then compared. It should be noted that when using empirical data the underlying true scores and the dimensionality of the response data are not known. However, there are two estimates of the student achievement available to us which are the exam score and the school score which is an independent estimate of achievement provided by the teacher. Hence, this was a comparative study.

Empirical Data Set

This study used the response data of the 6141 students who wrote the January 1996 British Columbia Grade 12 Provincial Math Examination. The exam was made up of 50 multiple choice (dichotomous) items and 8 open-ended (polytomous) items.

The topics covered in the 1996 Provincial Math Exam were trigonometry; quadratic relations; exponential and logarithmic functions; polynomial functions; sequences and series; introduction to calculus; geometry; and problem solving. All topic areas were covered in the multiple choice section and all except for polynomial functions were covered in the open-ended section of the exam. The open-ended items were multi-

step problems that required the completion of a number of steps in which partial credit would be awarded on the basis of the number of steps completed. The 8 open-ended items ranged from one to four steps. Some of the items required graphing steps, while others required mathematical equation steps. It should be noted that the open-ended items were rearranged from the lowest values to the highest values in order to calibrate the items more easily with the software programs. Question 1 on the exam was changed to 2; Question 2 to 5; Question 3 to 6; Question 4a to 3; Question 4b to 1; Question 5 to 7; Question 6 to 4; and Question 7 to 8.

Computer Programs and Procedures

BILOG 3 (Mislevy & Bock, 1990) performs item analysis and test scoring with the 1PL, 2PL, and 3PL binary models which were used to analyze the dichotomous section of the exam. The IRT dichotomous estimates from BILOG were used in this study to make comparisons with PARSCALE and RUMM examinee ability estimates, item estimates, and CTT estimates.

The computer program RUMM (Sheridan et al., 1997) uses a Rasch (1PL) unidimensional model for measurement and was used to analyze the 50 multiple choice and 8 open-ended items from the 1996 Provincial math examination. Note that one of the open-ended items was dichotomous and therefore was included with the multiple choice (dichotomous) analysis. The item that was removed had only two categories and was therefore dichotomous. This study also reviewed the item characteristic response functions (ICRF) in detail and the model-data fit from the 1-parameter partial credit model.

PARSCALE 3 (Muraki & Bock, 1996) performs IRT based test scoring and item analysis for graded open-ended exercises and performance tasks. The response data set consists of a mixture of open-ended (polytomous) and multiple choice (dichotomous) items that were analyzed jointly using PARSCALE. The 1PL and 2PL logistic models were applied to the dichotomously scored items and the 1PCM and 2PCM models were applied to the polytomously scored items by PARSCALE. The 1PL/1PCM combined models were fit simultaneously on the same Θ scale and the 2PL/2PCM combined models were also fit simultaneously on the same Θ scale. Note that PARSCALE and BILOG are similar programs in the underlying analysis or algorithms except PARSCALE allows for an open-ended (polytomous) item analysis.

Dimensionality

In addition to the comparisons of test and item calibrations using the IRT based analysis, approaches for assessing unidimensionality were also evaluated. The IRT assumption of unidimensionality was assessed using the computer program DIMTEST (Stout et al., 1993). Stout's T statistic (Stout, 1987), along with a principal axis factor analysis from within the computer program DIMTEST were used in order to test whether or not the set of dichotomously scored math items for a particular group of examinees were unidimensional. Multidimensional scaling for the assessment of unidimensionality was also applied to the data set and evaluated in this study.

Another assumption of the IRT models that is related to dimensionality is that of speedeness. It is necessary that all examinees have enough time to try all the items so that their ability level affects their responses and not the failure to reach an item. It is implicit in the unidimensional assumption that only one ability is being measured and not

a second ability of speed in answering a question (Hambleton & Swaminathan, 1985). The assumption of speededness was evaluated in this study by calculating the amount of omission rates in the data set.

Item-Fit

Item-fit was also reviewed and comparisons were made between the different models. PARSCALE performs IRT based analyses on items and exercises that are dichotomously and polytomously scored and allows for a test of goodness-of-fit for both types of items which is similar to the method used by BILOG for dichotomously scored items. BILOG and PARSCALE evaluate fit item by item. PARSCALE uses a likelihood-ratio chi-square statistic to evaluate item-fit and the sum of these chi-squared statistics provides the likelihood-ratio chi-square statistic for the whole test (Samejima, 1997).

Parameter Invariance

Item invariance is a feature of IRT when a model fits the data set and was investigated by randomly splitting the population of examinees into two halves and comparing the estimates of item parameters. A more rigorous test of item invariance was also applied which involved taking the top half (high ability group) and the bottom half (low ability group) of the distribution of examinees using BILOG and then comparing the item estimates (Hambleton et al., 1991).

Ability invariance is another feature of IRT when the model fits the data and was investigated by splitting the exam into two smaller 20 item sub-tests. By splitting the exam, two estimations of ability for each examinee can be obtained and then compared. Again, a further analysis that is more rigorous was also applied and involved the splitting

of the test into the 20 hardest and 20 easiest items. This is more rigorous in the sense that the two tests varied in difficulty with no overlap compared to the previous random split of items. The above provided evidence about parameter invariance and was analyzed using Pearson product correlations and scatterplots (Hambleton et al., 1991).

Consistency of Classification

Once a model was fit to the data and scores were generated for each student, cut-points on the IRT theta scale were estimated by comparison to examination raw score and letter grade cut-point distributions. A comparison was then made between students located within a theta scale cut-point range and the Ministry approved letter grade scale. The extent of misclassification from one scale to the other was evaluated and should be one indication of the feasibility of analyses by both IRT logistic and partial credit models with Provincial exam data. This classification consistency was assessed using cross tabulation, Cohen's kappa (Systat 7.0), and Pearson product moment correlation coefficients. Cohen's kappa, "...may be interpreted as the increase in decision consistency that the tests provide over chance expressed as a proportion of the maximum possible increase over chance consistency" (Crocker & Algina, 1986, p. 201).

In order to use an IRT model with the Provincial exam data, there must be a link established between the Ministry's letter grade scale and the IRT scale. The ability estimates from an IRT model should be a good estimate of the true scores for each examinee. However, when using empirical data the underlying true scores are not known and therefore, the IRT ability estimates should be compared to the CTT raw scores for consistency of letter grade classification. Before this can occur, there must be agreement with the assumption that the Ministry grades are adequate estimates of the true underlying

ability. If one can agree with this conclusion, then an IRT model or models can be applied to the Provincial exam data and compared to the CTT estimates on the Ministry letter grade scale (consistency of classification). The factors such as unidimensionality, item fit, parameter invariance, and classification consistency were all used to assess the feasibility of analyses by both IRT binary and partial credit models with a Provincial exam data set.

Chapter 4

Results

The results of the investigation of feasibility using binary logistic models with the dichotomously scored items and partial credit models with the polytomously scored items are presented in three parts. These parts consisted of goodness of fit, parameter invariance, and consistency of classification. Note that the goodness of fit investigation which includes unidimensionality and item-fit was only used to help check for parameter invariance. Added to the feasibility study were comparisons of parameter estimates from one software program to the next; comparisons across different models of IRT, and the types of information coming from the open-ended items (item characteristic response functions).

Assessing Goodness of Fit

Unidimensionality. Unidimensionality is a fundamental assumption of the IRT models. As previously noted, the assumption is that there is only one underlying dimension that could account for the examinees responses on a particular test. If this assumption is met, then the interpretation of the results will be meaningful (Hambleton & Swaminathan, 1985). Stout's T statistic (Stout, 1987), along with a principal axis factor analysis from within the computer program DIMTEST were used in order to test whether or not the set of dichotomously scored math items for a particular group of examinees were unidimensional. "DIMTEST rejects the hypothesis of unidimensionality when $T > Z_a$, where Z_a is the upper 100 $(1-\alpha)$ percentile for the standard normal distribution, α being the desired level of significance" (Stout et al. 1993, p.3).

An exploratory principal axis factor analysis was first performed using tetrachoric correlations on the 50 multiple choice items from the 1996 Provincial Math Examination with a sample of 4000 students. A first factor was extracted with an eigenvalue of 5.37 (Table 1). A second factor was extracted with an eigenvalue of 1.62 and suggests that the data set may not be unidimensional. Factor loadings are shown in Appendix A and eigenvalues in Table 1. The second analysis that was performed on the data set was the Stout's T statistic. A T value of 3.84, $p<.0001$ was found to be significant thus rejecting the null hypothesis of essential unidimensionality. Along with DIMTEST there was also an evaluation of multidimensional scaling for the assessment of unidimensionality with the Provincial data set. This evaluation comes in part because the article by Oltman, Stricker, and Barrows (1990) state that, "multidimensional scaling appears to be a useful method for item-level analyses of test structure" (p. 21).

Table 1

Eigenvalues from Factor Analysis

Factor	Eigenvalue
1	5.37
2	1.62
3	0.73

The MDS analysis (Systat 7) was carried out for one, two and three dimensions. A one-dimension solution was chosen on the basis of increments in the variance accounted for by successive solutions. The results showed an r^2 of 0.91 for the first dimension, an r^2 of 0.94 with the second dimension added and an r^2 of 0.95 with the third dimension added which suggests that there is only one underlying dimension in the data set.

The findings about undimensionality are mixed in this study, however they are in line with the Hattie et al. (1996) article which showed disagreement between different tests of unidimensionality. The article alludes to the importance of the assumption of unidimensionality and states that because it is not a well addressed area, more research is needed.

Another assumption of the IRT models that is related to dimensionality is that of speedeness. It is necessary that all examinees have enough time to try all the items so that their ability level affects their responses and not the failure to reach an item. It is implicit in the unidimensional assumption that only one ability is being measured and not a second ability of speed in answering a question (Hambleton & Swaminathan, 1985). The assumption of speededness was evaluated in this study by calculating the amount of omission rates in the data set. Approximately 0.3% of the answers were found to be omitted across the entire population which shows that the assumption of speededness was reasonably well met. In order to make that conclusion a comparison was made between the percentage of omits found if half the population had missed the last five items in a row. If half of the population would have missed the last 5 items, then there would have been approximately 5% omitted data. If examinees would not have had enough time to complete the exam, then the number of omits would have been higher (closer to 5%).

Item-fit statistics. The item-fit statistics represents the extent to which the actual student responses match the predicted responses from an IRT model. It is important to assess item-fit in order to ensure that the most appropriate model is used to analyze the data (Lord, 1980). Chi-square statistics were used to evaluate the IRT item-fit. The results of the fit statistics for the 50 multiple choice items are presented in Table 2. These include

the fit statistics of the 1PL, 2PL, and 3PL models from BILOG and the 1PL and 2PL from PARSCALE and the 1PL model from RUMM. The results show the 3PL model from BILOG best fits the math data set with only 4 items rejected at $p < .001$. The 2PL model was found to be the second best fitting model from PARSCALE with only 19 items rejected. The 2PL model from BILOG had the third best fit overall with only 25 items rejected and this increased when using the 1PL model. The ICCs for the 50 multiple choice items with item fit statistics from the BILOG 2PL model are presented in Appendix B. In reviewing these ICCs in Appendix B aside from the chi-square results there was in general a good level of fit.

Table 2

50 Multiple Choice Item Rejection by Fit Statistics

	Bilog			Parscale		Rumm
	1PL	2PL	3PL	1PL	2PL	1PL
Number of items	37	25	4	50	19	39

$p < .001$

The results of the fit statistics for the 8 open-ended items are presented in Table 3. These include the fit statistics of the 1PCM and 2PCM (partial credit) models from PARSCALE and the 1PCM from RUMM. On the open-ended item section, the 1PCM and 2PCM models from PARSCALE rejected all 8 items and the 1PCM model from RUMM rejected 7 items (Table 3). Overall, the results may differ because the assessment of fit varies from software program to software program. Also, Hambleton et al. (1991)

Table 3

Open-ended Item Rejection by Fit Statistics

	Parscale		Rumm
	1PCM	2PCM	1PCM
Number of items rejected	8	8	7

$p<.001$

notes the sensitivity of item-fit statistics due to large sample sizes. The current study under investigation had a very large sample size and this should be taken into account by reviewing whether or not parmameter invariance was met. As mentioned earlier item-fit is used only as a check for parameter invariance.

Parameter Invariance

A model feature of IRT is that scores remain comparable even if items are added or deleted from a test (ability parameter invariance). The scale scores used in IRT allow for equivalent item statistics to be obtained from different groups of examinees taking a particular test (item parameter invariance). "The importance of the property of invariance of item and ability parameters cannot be overstated. This property is the cornerstone of item response theory and makes possible such important applications as equating, item banking, investigation of item bias, and adaptive testing" (Hambleton et al., 1991).

In order to assess item parameter invariance, the group of 6141 examinees were randomly divided into two groups and each half was then used to calibrate the item statistics. Correlations between item statistics from each half were then compared for the two separate runs of the PARSCALE program (Table 4a and 4c) and the means and

Table 4a

Subgroup Correlations for the 1PL/1PCM Combined Model.

	Subgroup 1 b-parameter 1	Subgroup 2 b-parameter 2	Full Sample b-parameter 3
b-parameter 1	1.00		
b-parameter 2	1.00	1.00	
b-parameter 3	1.00	1.00	1.00

Table 4b

Means and Standard Deviations of Item Difficulty (b-parameter) for the 1PL/1PCM Combined Model Estimates.

	Subgroup 1 b-parameter 1	Subgroup 2 b-parameter 2	Full Sample b-parameter 3
M	-0.47	-0.47	-0.47
SD	0.97	0.98	0.97

Table 4c

Subgroup and Total Exam Correlations for the 2PL/2PCM Combined Model.

	Subgroup 1		Subgroup 2		Whole Sample	
	a1	b1	a2	b2	a3	b3
a1	1.00					
b1	-	1.00				
a2	0.90	-	1.00			
b2	-	0.99	-	1.00		
a3	0.99	-	0.90	-	1.00	
b3	-	1.00	-	0.99	-	1.00

a= a-parameter
b= b-parameter

standard deviations are presented in Table 4b and 4d. These calibration runs included a 1PL/1PCM and 2PL/2PCM on the whole math examination. The means and standard deviations of item difficulty for the 1PL/1PCM are same for each group (M=-0.47 and

SD=0.97/0.98). The means and standard deviations of item difficulty and discrimination for the 2PL/2PCM were similar for each group (Table 4d). The correlations for

Table 4d

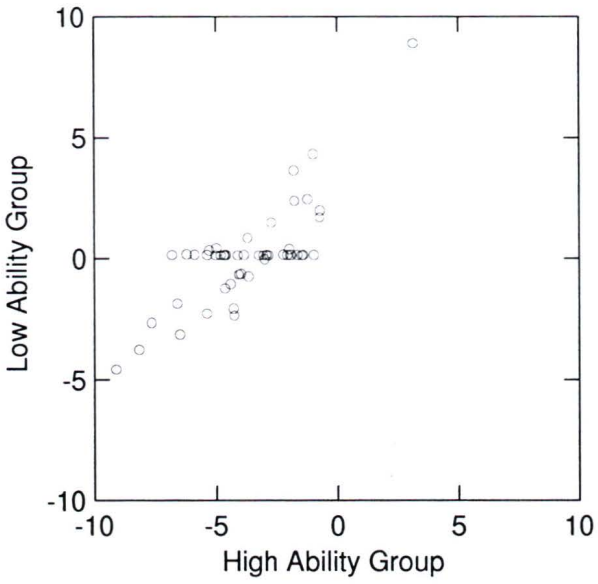
Means and Standard Deviations of Item Difficulty (b-parameter) and Discrimination (a-parameter) for the 2PL/2PCM Combined Model.

	Subgroup 1		Subgroup 2		Whole Sample	
	a1	b1	a2	b2	a3	b3
M	0.60	-0.57	0.62	-0.53	0.61	-0.56
SD	0.19	0.93	0.21	0.98	0.19	0.92

a= a-parameter
b= b-parameter

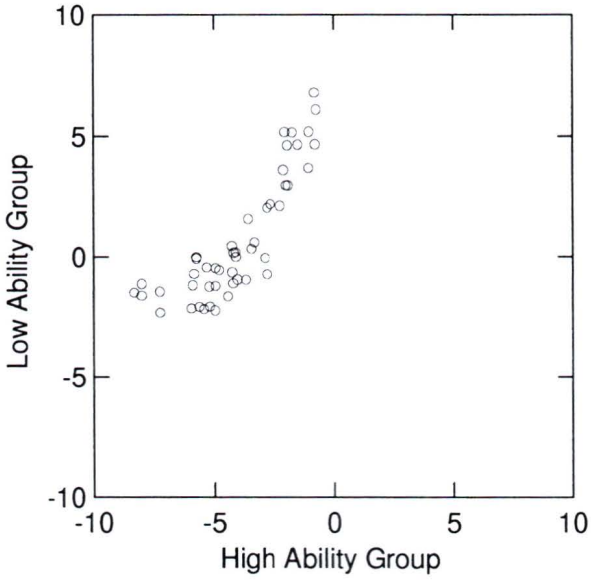
the item parameter invariance are presented in Tables 4a and 4c and show that on a random split of the population similar item estimates were calibrated and in this way item invariance holds (no correlations were less than 0.9). High correlations were found for both the 1PL/1PCM and 2PL/2PCM combined models with the 1PL/1PCM having slightly higher correlations of 1.00).

A more rigorous test of item invariance involved taking the top half of examinees (high ability group) and the bottom half (low ability group) of the distribution of examinees and then comparing item estimates using Bilog 1PL, 2PL, and 3PL models. The 50 multiple choice item difficulty estimates were calibrated from a sample of 2000 low ability students and 2000 high ability students and were plotted against each other in Figures 8a-8c (theta scale). The correlations between item difficulty estimates from BILOG for the 1PL was 0.79, 2PL was 0.82, and 3PL was 0.83 on the high/low split of ability groups (item invariance) and are plotted in Figures 8a-8c. The means and standard



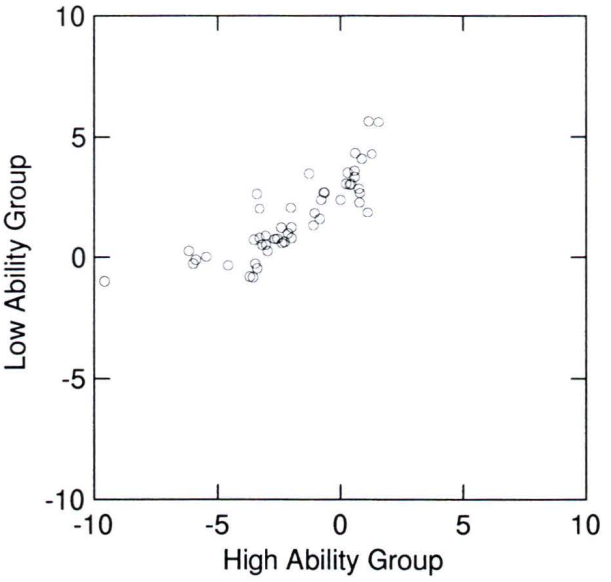
r=0.79

Figure 8a. Plot of 1PL Item Difficulty Estimates Based on Samples of High/Low Ability



r=0.82

Figure 8b. Plot of 2PL Item Difficulty Estimates Based on Samples of High/Low Ability



r=0.83

Figure 8c. Plot of 3PL Item Difficulty Estimates Based on Samples of High/Low Ability

deviations for the high/low ability split from the Bilog 1PL, 2PL, and 3PL models are found in Table 5. These plots of high/low splits (item invariance) suggests that there is a positive linear relationship that is offset upwards and in this way item parameter invariance was supported.

In order to assess the ability parameter invariance, 20 items were randomly selected for each run of the 1PL/1PCM and 2PL/2PCM models and examinee ability estimates were generated. Each examinee then has two separate ability estimates (one from each 20 item sub-test). The correlations between the two ability estimates for each examinee from the 1PL/1PCM and 2PL/2PCM combined models are presented in Table 6a and 6c while the means and standard deviations are presented in Table 6b and 6d. Higher correlations were found for the 2PL/2PCM (0.75 to 0.89) than the 1PL/1PCM

Table 5

Means and Standard Deviations for 1PL/2PL/3PL Split Group (High/Low Ability) Difficulty Estimates.

	1PL		2PL		3PL	
	High	Low	High	Low	High	Low
M	-3.65	0.10	-3.84	1.30	-1.85	1.69
SD	2.24	2.05	2.19	3.88	2.37	1.65

calibration run (0.74 to 0.85) and both suggest that on a random split of the items, ability invariance was supported. The suggestion of ability invariance is based on the high correlations found between examinee estimates even when only 20 of the 50 items were used to calibrate the examinees' ability. The above will be used as a baseline in order to have a comparison for a more rigorous test of parameter invariance that was analyzed next.

A more rigorous test for ability invariance was also applied by splitting the test into two 20-item tests of which one test contained the easiest items and the other contained the hardest items from the math exam. This is more rigorous in the sense that the two tests varied in difficulty with no overlap compared to the previous random split of items. Correlations between ability estimates from BILOG were found to be 0.71 for the

Table 6a

Correlations of Ability Sub-tests and Whole Exam for the 1PL/1PCM Combined Model.

	Ability 1	Ability 2	Ability 3
Ability 1	1.00		
Ability 2	0.74	1.00	
Ability 3	0.85	0.82	1.00

1= subtest 1
2= subtest 2
3= whole exam

Table 6b

Means and Standard Deviations of Ability Sub-tests and Whole Exam for the 1PL/1PCM.

	Ability 1	Ability 2	Ability 3
M	0.05	0.02	0.07
SD	1.13	1.04	0.89

1= subtest 1
2= subtest 2
3= whole exam

Table 6c

Correlations of Ability Sub-tests and Whole Exam for the 2PL/2PCM Combined Model.

	Ability 1	Ability 2	Ability 3
Ability 1	1.00		
Ability 2	0.75	1.00	
Ability 3	0.85	0.89	1.00

1= subtest 1
2= subtest 2
3= whole exam

Table 6d

Means and Standard Deviations of Ability Sub-tests and Whole Exam for the 2PL/2PCM.

	Ability 1	Ability 2	Ability 3
M	0.17	0.08	0.10
SD	0.86	1.23	1.02

1= subtest 1
2= subtest 2
3= whole exam

1PL, 0.71 for the 2PL, and 0.72 for the 3PL hard/easy split of items (ability invariance) and suggest that examinees estimates from two tests were moderately correlated. The moderate correlations from the forced selection of the 20 easiest items and the 20 hardest items from the exam for examinee ability calibration indicate ability invariance was supported.

Ability Estimates

In order to ensure estimates from different software programs and models were equivalent and to also assess the relationships between the estimates from IRT and CTT, comparisons of ability estimates from the 1PL, 2PL, and 3PL models from BILOG and 1PL/1PCM and 2PL/2PCM combined models from PARSCALE were analyzed. Also the 1PL/1PCM from RUMM along with the Ministry estimates were analyzed and are all presented in Table 7. Note that the correlations of ability estimates in Table 7 are based on estimates before score transformation onto the Ministry-approved letter grade scale. The next section on classification consistency introduces correlations between CTT and IRT ability estimates after score transformation onto the Ministry-approved letter grade scale. All IRT and Exam estimates had low correlations with the School estimates and high correlations with Exam estimates because the IRT estimates were based on the examinee responses to the Exam (Exam raw score also had low correlation with School estimate). The Provincial score is the combined results from the Exam score (40%) and the School score (60%). The School score is a separate estimate of ability based on the in-class achievement awarded by the student's teacher. The highest correlation was found using the PARSCALE combined 1PL/1PCM model (0.92) with the Exam score. The PARSCALE 1PL/1PCM model also had the highest correlation with the School estimate of 0.34 when compared to the other parameter estimates. The BILOG 3PL had the highest correlation with the Provincial estimate of 0.88 and it will be noted again that BILOG only estimates the ability of an examinee from the 50 multiple choice items - the open-ended items were not included. The RUMM estimates in Table 7 seem to have

Table 7

Correlations of Ability Estimates

	School	Exam	Prov	Par11	Par22	Bilog	Bilog	Bilog	Rum
School	1.00								
Exam	0.34	1.00							
Prov	0.44	0.87	1.00						
Par11	0.34	0.92	0.88	1.00					
Par22	0.34	0.91	0.88	1.00	1.00				
Bilog1	0.32	0.91	0.87	0.96	0.96	1.00			
Bilog2	0.33	0.91	0.88	0.96	0.96	1.00	1.00		
Bilog3	0.34	0.91	0.88	0.96	0.95	0.99	0.99	1.00	
Rumm	0.30	0.82	0.80	0.92	0.92	0.88	0.89	0.87	1.00

School = School estimate
Exam = Exam estimate
Prov = Provincial estimate(sum of School and Exam)
Par11 = PARSCALE 1PL/1PCM estimate
Par22 =PARSCALE 2PL/2PCM estimate
Bilog1 = BILOG 1PL estimate
Bilog2 = BILOG 2PL estimate
Bilog3 =BILOG 3PL estimate
Rum= RUMM 1PL/1PCM estimate

consistently lower correlations than the other estimates. This suggests overall that the 1PL/1PPC model from PARSCALE had the closest estimates to the current CTT estimates of Exam and Provincial. Overall, IRT is performing like the Exam score.

Classification Consistency

In order to use an IRT model with the Provincial exam data, there must be a meaningful link established between the Ministry's letter grade scale and the IRT scale. The ability estimates from an IRT model should be a good estimate of the true scores for each examinee. However, when using empirical data the underlying true scores are not known and therefore, the IRT ability estimates should be compared to the CTT raw scores for consistency of letter grade classification. Before this can occur, there must be agreement with the assumption that the Ministry grades are adequate estimates of the true

underlying ability. If one can agree with this conclusion, then an IRT model or models can be applied to the Provincial exam data and compared with the Ministry letter grade scale. In order to make this comparison the IRT theta scale cut-points were estimated by locating them on the examination score cut-point distributions (Ministry-approved letter

Table 8a

Correlations of Converted Ability Estimates from the Parscale 1PL/1PCM Combined Model.

	Ability 11	School	Provincial	Exam
Ability11*	1.00			
School	0.57	1.00		
Provincial	0.88	0.70	1.00	
Exam	0.98	0.57	0.88	1.00

*Ability 11 = Parscale 1PL/1PCM

Table 8b

Correlations of Converted Ability Estimates from the Parscale 2PL/2PCM Combined Model.

	Ability 22	School	Provincial	Exam
Ability22*	1.00			
School	0.63	1.00		
Provincial	0.90	0.79	1.00	
Exam	0.97	0.57	0.87	1.00

*Ability 22 = Parscale 2PL/2PCM

Table 9

Cohen's Kappa Coefficient for Transformed Ability Scores.

	1PL/1PCM	2PL/2PCM
Exam	0.89	0.82
School	0.20	0.23
Provincial	0.35	0.37

grades). The theta scores for all examinees around the cut-points (Exam score) were averaged. Then the average theta was taken as the cut-point. Once the cut-points were established, a comparison was made between the classification of letter grade status. This comparison was made between the Ministry-approved letter grades estimated by CTT (standard setting) and Ministry-approved letter grades from IRT. The ability estimates from the 1PL/1PCM (Ability 11) and 2PL/2PCM (Ability 22) models were scaled by the above method onto the Ministry-approved letter grade scale (A = 86-100%, B = 73-85%, C+ = 67-72%, C = 60-66%, C- = 50-59%, and F = 0-49%) and an assessment of classification consistency was then performed. This classification consistency was assessed using Pearson product moment correlations presented in Table 8a-8b and Cohen's Kappas (K) which are presented in Table 9. Cohen's kappa, "...may be interpreted as the increase in decision consistency that the tests provide over chance expressed as a proportion of the maximum possible increase over chance consistency" (Crocker & Algina, 1986, p. 201). Cohen's Kappa was found to be the highest (0.89) when using the combined 1PL/1PCM to calibrate the data set in comparison to the classification of letter grades from the Exam estimates. The reason for the high coefficient with the Exam is because the ability scores are based on the Exam results and not the School or the Provincial results. This high Kappa of 0.89 means that 89% of the total possible increase over chance consistency was observed for the decisions based on the two measures. In other words the 1PL/1PCM IRT estimates are performing like the Exam scores from CTT.

Table 10a-10f shows the cross tabulation of student classification from different estimates of ability. These different estimates include the 1PL/1PCM and 2PL/2PCM

combined models from PARSCALE which are compared to the Provincial, Exam, and School estimates. The highest consistency of classification was found using the 1PL/1PCM estimates compared to the Exam estimates (9% misclassification). The term misclassification will be used to represent students who were placed in a different letter grade group when comparing their IRT and CTT estimates. The highest consistency of classification between the School estimates was with the 2PL/2PCM estimates with 64% misclassification. The highest consistency of classification was with the 2PL/2PCM

Table 10a

Frequencies of 1PL/1PCM Ability Estimates (Rows) by Exam Scores (Columns).

K=.89

	F	C-	C	C+	B	A	Total
F	995	81	1	0	0	0	1077
C-	10	1037	124	0	0	0	1171
C	1	3	769	56	0	0	829
C+	0	0	82	778	103	0	962
B	0	0	0	7	1078	84	1169
A	0	0	0	0	8	924	932
Total	1006	1121	976	841	1189	1008	6141

Table 10b

Frequencies of 2PL/2PCM Ability estimates (Rows) by Exam Scores (Columns).

K=.82

	F	C-	C	C+	B	A	Total
F	924	93	1	0	1	0	1019
C-	82	978	161	0	0	0	1221
C	1	50	742	137	0	0	930
C+	0	0	72	626	94	0	792
B	0	0	0	78	1050	86	1214
A	0	0	0	0	44	921	965
Total	1007	1121	976	841	1189	1007	6141

Table 10c

Frequencies of 1PL/1PCM Ability Estimates (Rows) by School Scores (Columns).

K=.20

	F	C-	C	C+	B	A	Total
F	350	360	208	92	65	12	1087
C-	106	275	289	243	230	28	1171
C	70	60	138	199	308	54	829
C+	78	21	87	128	481	167	962
B	122	3	16	43	477	1508	1169
A	139	0	1	3	86	694	923
Total	865	719	739	708	1647	1463	6141

Table 10d

Frequencies of 2PL/2PCM Ability Estimates (Rows) by School Scores (Columns).

K=.23

	F	C-	C	C+	B	A	Total
F	375	373	175	61	29	6	1019
C-	115	287	360	248	194	17	1221
C	70	45	150	269	350	46	930
C+	62	13	43	91	470	113	792
B	121	1	10	37	526	519	1214
A	121	0	1	2	78	761	965
Total	866	719	739	708	1647	1463	6141

Table 10e

Frequencies of 1PL/1PCM Ability Estimates (Rows) by Provincial Scores (Columns).

K=.35

	F	C-	C	C+	B	A	Total
F	438	489	126	20	13	1	1087
C-	24	389	455	235	68	0	1171
C	1	36	206	325	258	3	829
C+	0	3	65	213	656	25	962
B	1	1	2	27	710	428	1169
A	8	0	0	0	54	861	923
Total	472	918	739	708	1759	1318	6141

Table 10f

Frequencies of 2PL/2PCM Ability Estimates (Rows) by Provincial Scores (Columns).

K=.37

	F	C-	C	C+	B	A	Total
F	451	488	67	4	8	1	1019
C-	2	425	594	170	30	0	1221
C	1	4	169	514	242	0	930
C+	0	1	23	127	629	11	792
B	2	0	0	5	826	381	1214
A	15	0	0	0	25	925	965
Total	472	918	853	820	1760	1318	6141

when compared to the Provincial estimates at 52% misclassification. The results suggest that the 1PL/1PCM had the highest consistency of classification and awarded the same letter grade to 91% of the 6141 students who wrote the math exam when compared to their estimate from CTT. As mentioned earlier, in order to use an IRT model with the Provincial exam data, there must be a link established between the Ministry's letter grade scale and the IRT scale. The ability estimates from an IRT model should be a good estimate of the true scores for each examinee. However, when using empirical data the underlying true scores are not known and therefore, the IRT ability estimates should be compared to the CTT raw scores for consistency of letter grade classification. Before this can occur, there must be agreement with the assumption that the Ministry grades are adequate estimates of the true underlying ability. If one can agree with this conclusion, then an IRT model or models can be applied to the Provincial exam data and compared to the CTT estimates on the Ministry letter grade scale (consistency of classification).

Item Parameter Estimates

In order to ensure that estimates from different software programs and models were equivalent and to also assess the relationships between the estimates from IRT and CTT- comparisons of ability estimates from the 1PL, 2PL, and 3PL models from BILOG and 1PL/1PCM and 2PL/2PCM combined models from PARSCALE and the 1PL/1PCM from RUMM along with the Ministry estimates were analyzed and are presented in Table 11. The correlations of item parameter estimates (difficulty or b-parameter) from CTT and IRT are presented in Table 11 and the raw score estimates are located in Appendix C. The correlations of item parameter estimates (discrimination or a-parameter) from CTT and IRT are presented in Table 12 and the actual estimates are in Appendix C.

Correlations between all CTT and IRT difficulty estimates were found to be high

Table 11

Correlations Between all CTT and IRT Difficulty Estimates for the Dichotomous Items.

	P-Val	Bilog 1	Bilog 2	Bilog 3	Par11	Par22	Rum11
P-Val	1.00						
Bilog 1	-0.99	1.00					
Bilog 2	-0.98	0.97	1.00				
Bilog 3	-0.94	0.93	0.94	1.00			
Par11	-0.99	1.00	0.97	0.93	1.00		
Par22	-0.97	0.96	1.00	0.93	0.96	1.00	
Rum11	-0.99	1.00	0.96	0.92	1.00	0.96	1.00

P-Val = CTT estimate
Par11 = PARSCALE 1PL/1PCM estimate
Par22 =PARSCALE 2PL/2PCM estimate
Bilog1 = BILOG 1PL estimate
Bilog2 = BILOG 2PL estimate
Bilog3 =BILOG 3PL estimate
RUM 11= RUMM 1PL/1PCM estimate

Table 12

Correlations Between all CTT and IRT Discrimination Estimates for Dichotomous Items.

	P-Bis	Bilog 2	Bilog 3	Par22
P-Bis	1.00			
Bilog 2	0.97	1.00		
Bilog 3	0.32	0.33	1.00	
Par22	0.98	0.99	0.32	1.00

between 0.92 and a high inverse relationship of -0.99. The high inverse relationship of the P-val estimates compared to the IRT estimates were found because of the different scales used. As the difficulty of the items increase, the P-val scale decreases and the IRT scale increases. The correlations between discrimination estimates are were found to be low (between 0.32 - 0.33) when compared with BILOG 3PL model estimates. PARSCALE 2PL/2PCM when compared to CTT point biserial was found to be the highest at 0.98.

The 1-Parameter Partial Credit Model

Most IRT applications have been developed for use with dichotomous data and applications of polytomous IRT models have not been widespread (Muraki, 1993). It is thus important then to investigate to what extent the results of an IRT analysis would be interpretable when a polytomous model such as the partial credit model (Masters, 1990) is applied to polytomous items. In order to investigate fully the use of the IRT models with Provincial exam data, a review of the polytomous (open-ended) items was employed.

The RUMM software program uses a 1-parameter partial credit model for open-ended items sections. In order to interpret how the open-ended items performed on the

exam the item category response functions (ICRF) should be analyzed. As stated earlier in the partial credit section, each step parameter is defined by two adjacent categories and therefore, the interpretation becomes less straightforward in marking out the regions of an underlying variable compared to the dichotomous difficulty estimates. Because of this, the partial credit analyses should be accompanied by an ICRF. The ICRFs for the 7 open-ended items are shown in Figures 9a-9g and are from the RUMM software program. In order to help with the interpretation, clarification of categories will be briefly discussed. The following ICRFs (Figures 9a-9g) show category 1 as labeled 0 because it had a score of 0, category 2 was labeled 1 because it had a score of 1, category 3 was labeled 2 because it had a score of 2, and category 4 was labeled 3 because it had a score of 3. Only 7 of the 8 open-ended items are shown because the first item only has 2 categories which makes that item dichotomous. The trace lines in Figures 9a-9c are from a partial credit model with three categorical responses for items 52, 53 and 54 of the math exam. In order to explain the meaning of the curves the next section will review several of the item's ICRFs.

The probability of obtaining category three (labeled on the curve as a score of 2) in item 52 (Figure 9a) is high for most of the ability range (-1 and up), while the second category is relatively low in the probability of its attainment and this is consistent with the frequencies in Table 13. The percent frequency of categorical responses is presented in Table 13 and shows 70% of the examinees obtaining the correct responses to category three (ie. a score of 2). The ICRF for item 52 (Figure 9a) also shows the third category covering the largest part of the ability scale (about 70%). Also, only 13% of the examinees obtained the correct response to the second category and 17% in the first

category or a score of 0. Item 52 may then be considered relatively easy and demonstrates examinees either getting a score of 2 (70%) or a score of 0 (17%) or a score of 1 (13%) and in this regard acts more like a dichotomous item (most examinees obtaining a score of 0 or 2). Another way to interpret the information is by person location. A person location (ability) of 0 would have a 0.70 probability of obtaining a score of 2; a 0.10 probability of 0, and a 0.13 probability of obtaining a 1.

The ICRF in Figure 9b shows a more difficult item and should be interpreted in this way because of the shift to the right of the location parameter which is -0.20 compared to item 52 (-0.99). Examinees with an ability level of zero would have the greatest probability of getting the first part of this question correct. An examinee with an ability level or person location of 2 would have a 0.7 probability of obtaining a score of 2 and would have a 0.2 probability of obtaining a score of 1 on item 53.

The ICRF for item 54 in Figure 9c is a very difficult item with a relatively low probability of getting the second category of the item correct. This is shown by the first and third category intersecting above the second category. This item illustrates that most

Table 13

% Frequency of Categorical Responses.

Item Label	Category Responses				
	0	1	2	3	4
Item 52	17	13	70		
Item 53	28	43	29		
Item 54	73	15	12		
Item 55	11	28	31	30	
Item 56	32	23	24	21	
Item 57	22	14	12	52	
Item 58	34	19	09	23	15

of the examinees would not obtain the correct answer (only 29% obtained a score of 2) except for the high ability group (theta of 1.5 - 3). The frequency of categorical responses in Table 13 shows for item 54 that 73% would obtain a zero score and indicating that this is a difficult item. The ICRF for item 54 demonstrates that a student with a theta of 2 would have a 0.6 probability of obtaining category 3 (a score of 2). The trace lines in Figures 9d-9f are from a partial credit model with four categorical responses. Item 55 in Figure 9d shows a relatively easy first and second category with the fourth category covering a wide ability range. The item in Figure 9e acts more like a dichotomous item in that the second and third category are collapsed. This shows that a low-medium ability examinee would have a high probability of getting the answer wrong and a medium-high ability examinee would have a high probability of getting the entire answer to the question correct. Category two and three in Figure 9f are also collapsed

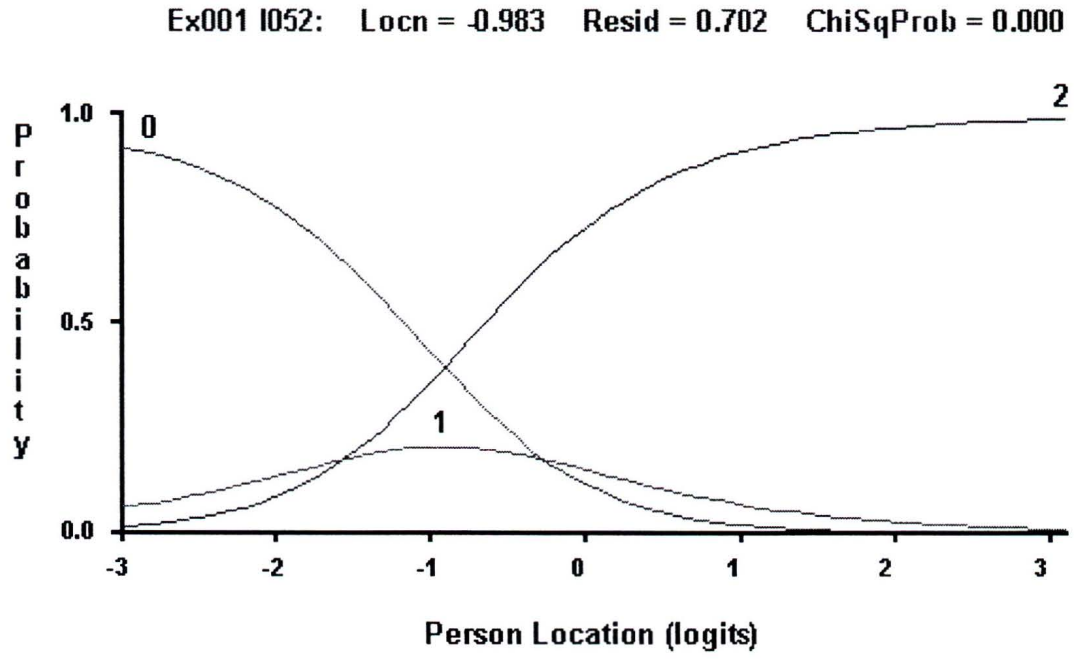


Figure 9a. ICRF of Item 52 from the 1996 Provincial Math Examination

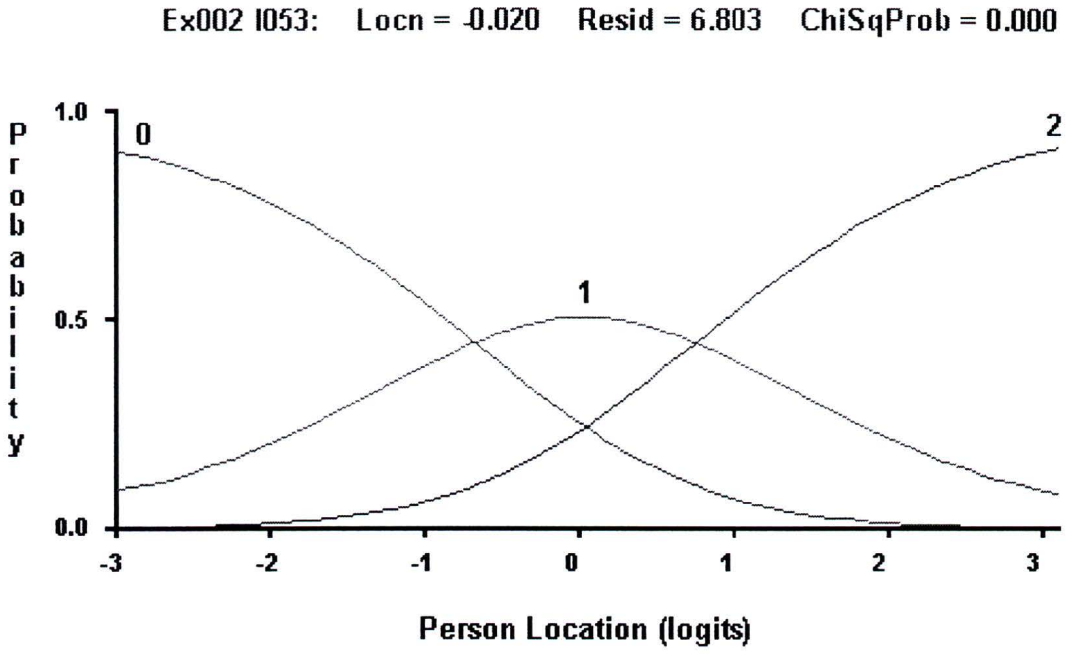


Figure 9b. ICRF of Item 53 from the 1996 Provincial Math Examination

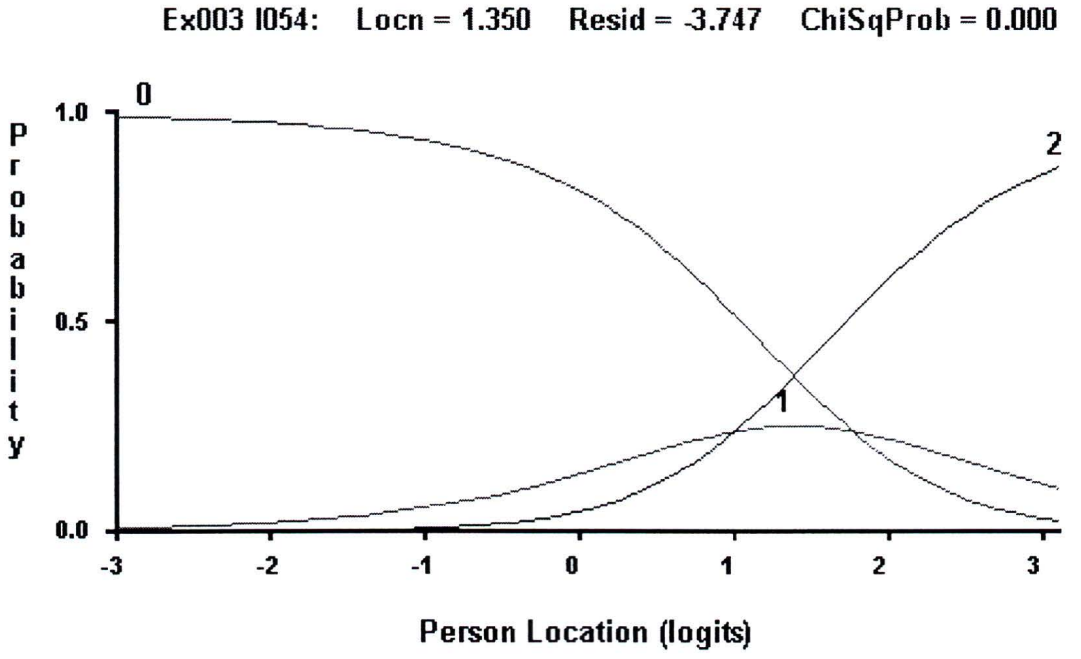


Figure 9c. ICRF of Item 54 from the 1996 Provincial Math Examination

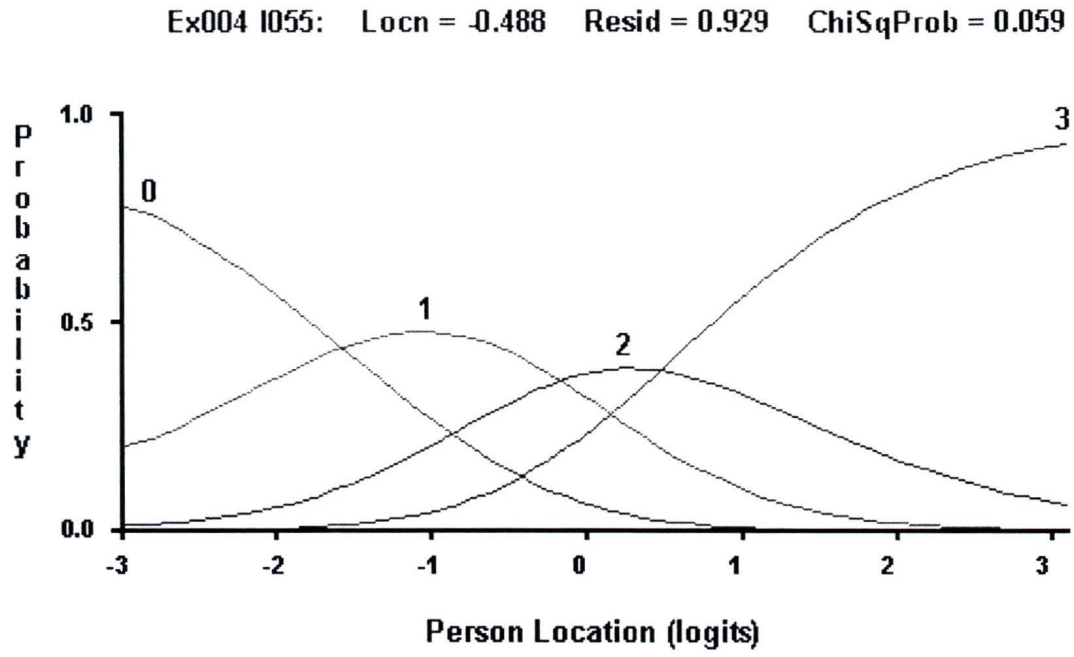


Figure 9d. ICRF of Item 55 from the 1996 Provincial Math Examination

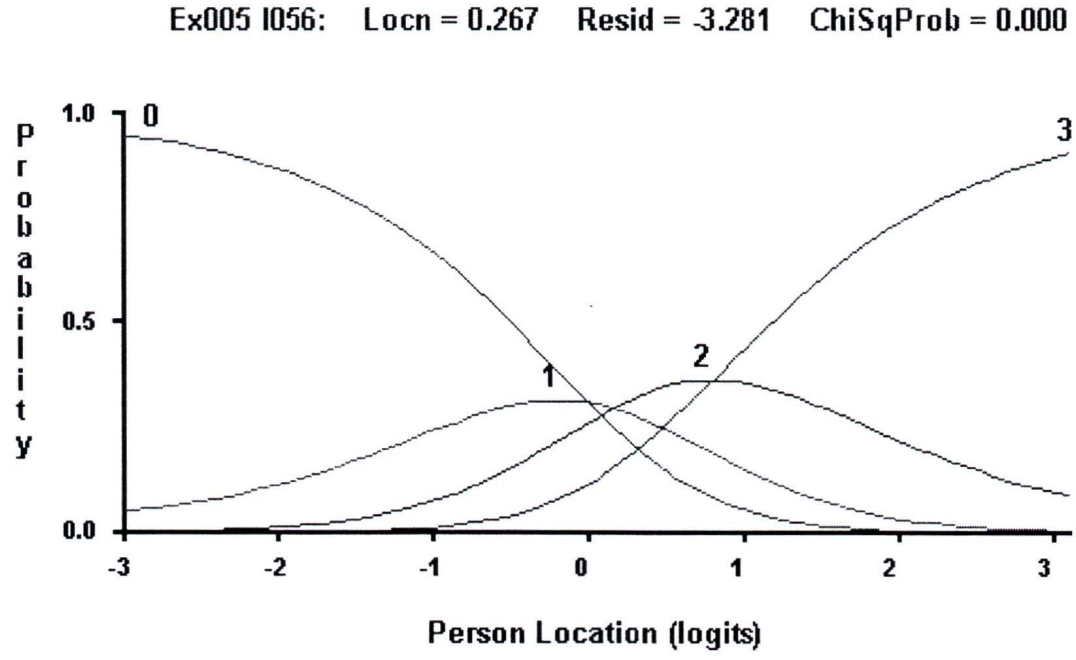


Figure 9e. ICRF of Item 56 from the 1996 Provincial Math Examination

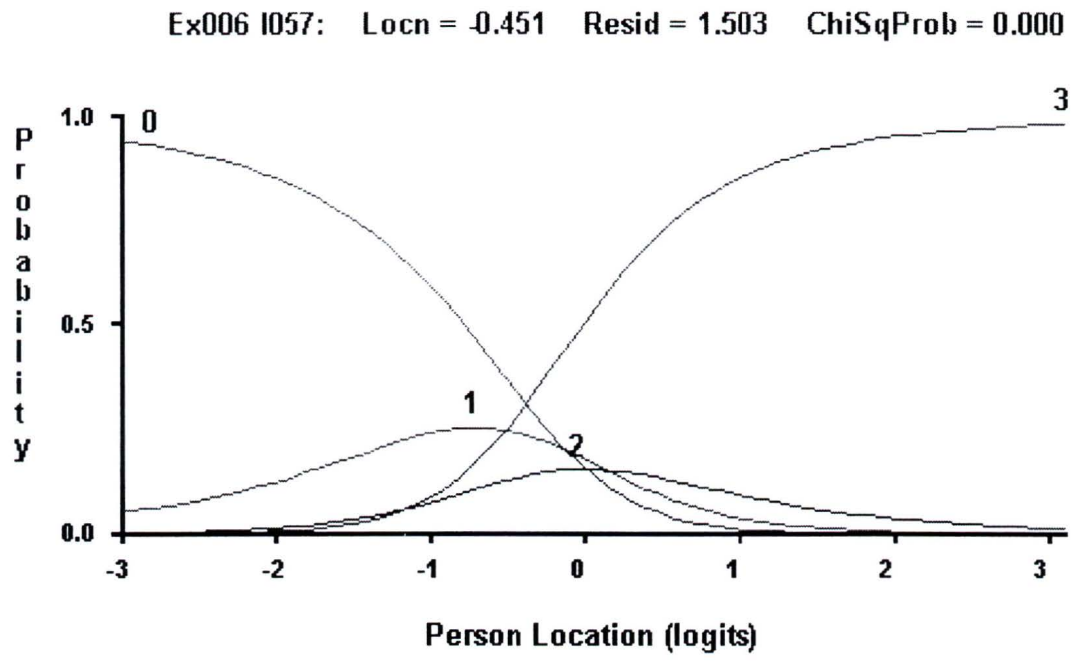


Figure 9f. ICRF of Item 57 from the 1996 Provincial Math Examination

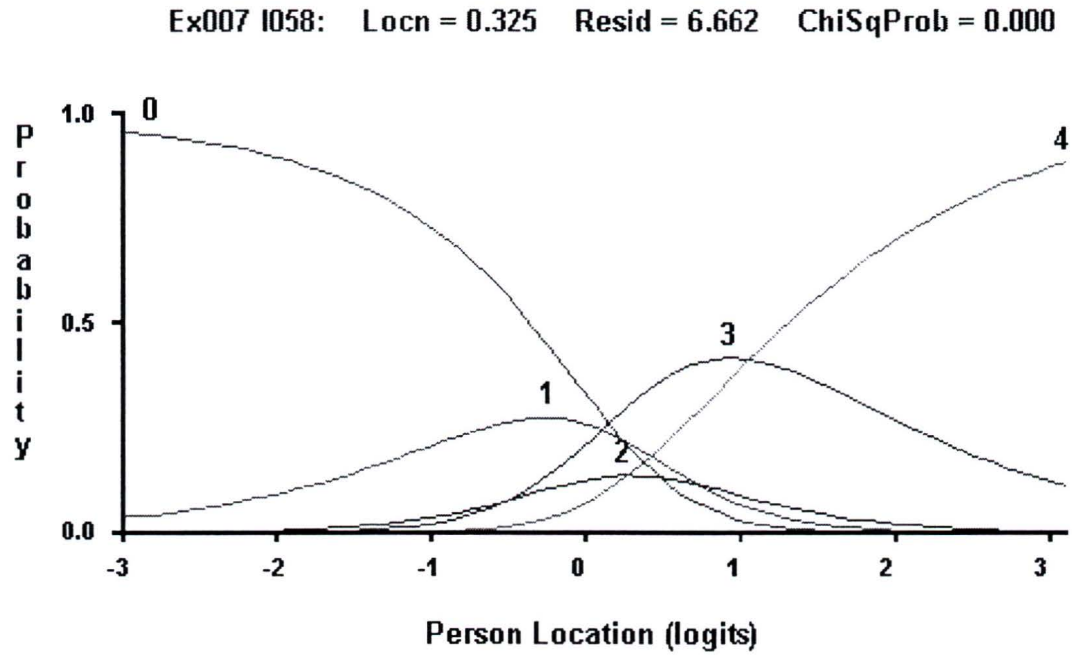


Figure 9g. ICRF of Item 58 from the 1996 Provincial Math Examination.

showing similar information to item 56 except for the fact that it was easier. The trace lines in Figure 9g are from a partial credit model with five categorical responses. Examinees with ability estimates from -3 to 0 all have a high probability of obtaining a score of 0 whereas examinees with ability estimates from 0 to 1 have a moderate probability (0.4) of success on attaining the correct answer to the first three

Table 14

Item-location and Category Parameters.

Item	Location	Category			
	b	d_1	d_2	d_3	d_4
Item 52	-0.99	0.65	-0.65		
Item 53	-0.02	-0.74	0.74		
Item 54	1.35	0.37	-0.37		
Item 55	-0.49	-1.18	0.26	0.92	
Item 56	0.27	-0.33	-0.17	0.50	
Item 57	-0.45	0.27	0.43	0.80	
Item 58	0.33	-0.15	0.35	-0.91	0.70

parts of the item (fourth category). Examinees with ability estimates between 1 and 3 have a high probability of obtaining a score of 4. Table 14 shows the item location and category parameters for items 52 to 58 (refer to Chapter 3 within the partial credit section for details). The fit statistics or chi-square results for the 7 open-ended items are presented in Table 15. All items were rejected as fitting the model except for item 55, $p<.001$. The frequency distribution of ability estimates for the entire examinee population is presented in Appendix D. This graph shows an equal distribution centering around zero of ability estimates for the 6141 examinees who answered the 7 open ended math items.

Table 15

Location and Fit of the 7 Open-ended Items.

Label	Location	SE	Fit	ChiSq	Prob
Item 52	-0.99	0.02	0.76	26.63	0.000
Item 53	-0.02	0.02	6.90	68.19	0.000
Item 54	1.35	0.02	-3.71	66.60	0.000
Item 55	-0.49	0.02	1.06	5.28	0.047
Item 56	0.27	0.02	-3.20	40.97	0.000
Item 57	-0.45	0.01	1.55	21.86	0.000
Item 58	0.33	0.01	6.73	70.29	0.000

$p<.001$

Item 52, 54, 56 and 57 were reviewed using the ICRFs and interpreted to be mostly dichotomous in the sense that the middle categories had few occurrences. This resulted in low ability examinees getting the question incorrect at every level while high ability examinees obtained the correct and final answer to the entire question. Item 53, 55 and 58 were also reviewed using the ICRFs and were found to act more polytomous in nature. This means that the middle item categories showed more occurrences of examinee attainment. Item-fit was assessed using a chi-square statistic and resulted in 6 of the 7 items being rejected as not fitting the 1-parameter partial credit model. The ICRFs from the 1-parameter partial credit model were found to be very meaningful for the interpretation of how the item categories were operating within an exam.

Chapter Five

Summary, Discussion, and Conclusion

This study investigated the feasibility of using of an item response theory (IRT) based analysis with response data that consists of both dichotomous and polytomous data. In order to investigate the feasibility of using IRT based models including partial credit models, the software programs PARSCALE, BILOG and RUMM were used to analyze the January 1996 British Columbia Grade 12 Provincial Mathematics Examination. Once the items were calibrated, the investigation of feasibility of using both dichotomous and polytomous item response models was implemented. The investigation of feasibility consisted of studies of goodness of fit, parameter invariance, and consistency of classification. Added to the feasibility study were comparisons of parameter estimates from one software program to the next; comparisons across different models of IRT, and the information derived from the open-ended items (item characteristic response functions).

Unidimensionality was assessed and is a fundamental assumption of the IRT models. Multidimensional scaling suggested that there is only one underlying dimension in the data set. However, Stout's T statistic and factor analysis rejected the assumption that the data set was unidimensional. Therefore, IRT based on the results of unidimensionality could not be supported. The findings about undimensionality are mixed in this study, however they are in line with the Hattie et al. (1996) article which showed disagreement between different tests of unidimensionality. Better detection procedures will have to be investigated via future research in order to increase confidence

in the assessment of unidimensionality and the possibility of misfit must be kept in mind when considering the results of the studies of parameter invariance.

Another assumption of the IRT models that is related to dimensionality is that of speedeness. It is necessary that all examinees have enough time to try all the items so that their ability level affects their responses and not the failure to reach an item. It is implicit in the unidimensional assumption that only one ability is being measured and not a second ability of speed in answering a question (Hambleton & Swaminathan, 1985). The assumption of speededness was evaluated in this study by calculating the amount of omission rates in the data set. It was found that speededness was not a factor with this math examination.

Item-fit and overall model-fit (undimensionality) is the extent to which the actual student responses match the predicted responses from an IRT model. There are some important features of interest if an item response model fits the data set and the pool of items all measure one underlying ability. The first is that the item parameter estimates obtained from the application of IRT models are independent of the sample of examinees to which a test is administered (item parameter invariance). Also, any set of items can be chosen from an item pool and the examinee estimates obtained are independent of that item selection (ability invariance) (Wainer, 1990). The item-fit statistics show the 2PL and 3PL models as the better fitting models than the 1PL model with the multiple choice section of the exam. This is in part because the low level students writing the exam have accumulated enough knowledge to allow for an educated guess. The *c* or lower asymptote (guessing) parameter is set at zero for the 1PL and is not estimated, however the 3PL model has a varying lower asymptote that would allow for better fit if the

examinees of low ability were to guess on the items. The 1PL model also assumes equal discrimination across all items. Both the 2PL and 3PL can account for discrimination because the slope is estimated for each item separately. The items in this study were found to vary in discrimination (slope) between low and high level students which accounts for why the 2PL and 3PL models were better fitting with this Provincial data set.

The 1PCM and 2PCM models from PARSCALE did not fit the open-ended section of the exam, although the 1PCM from RUMM did show one item as fitting. It should be noted that different software programs and models yielded different amounts of fitting items and therefore the exams should be calibrated using only one software program and model in order to ensure consistent results. Also, Hambleton et al. (1991) notes the sensitivity of item-fit statistics due to large sample sizes. The current study under investigation had a very large sample size and this should be taken into account when reviewing the number of rejection rates of which were high for most models.

If item-fit has been achieved, then the assumption of parameter invariance should hold and examinee statistics will not be based on the particular set of questions in an exam and the item statistics will not be based on a particular group of examinees (Hambleton & Swaminathan, 1985). Invariance of item parameter estimates was analyzed by splitting the examinees into a high ability group and a low ability group and then comparing item parameter estimates (rigorous test of high/low students). Hambleton et al. (1991) states that in order for true invariance to be shown the scatterplot should be in a straight line with a slope of 1 and an intercept of 0. "However, the item parameters for items that are administered to two separate groups may appear to be different. This

apparent discrepancy arises because of the arbitrary fixing of the metric for Θ (or b). A linear relationship, however, exists between the item parameters and ability parameters in the two groups" (Hambleton & Swaminathan, 1985, p.203). The slope of all three scatter plots in Figures 8a - 8c of high/low splits (item invariance) shows a positive linear relationship that is offset upwards and therefore supports the assumption of item invariance. Note that the examinees were divided into extreme ability groups. The low ability group was used to estimate the item difficulty levels for the 50 multiple choice questions and then the high ability group was used for the item calibration. The correlations in Figures 8a-8c represent the relationship between the two separate difficulty estimates from each group. These high correlations (0.79 to 0.83) provide evidence of parameter invariance.

Invariance of ability parameter estimates was checked by splitting the exam into two 20-item tests. One test consisted of 20 easy items , whereas the second test consisted of 20 hard items. Each examinee was then calibrated on the two tests and two separate estimates of ability were obtained (one for each sub-test). Comparisons by Pearson product correlations were made between the two ability estimates and differences were found between the ability estimates. However, invariance of ability parameter estimates was supported as seen by the moderate to high correlations between the ability estimates obtained from both exams.

In order to assess the relationships between the estimates from IRT and CTT, comparisons of ability estimates from the 1PL, 2PL, and 3PL models from BILOG and 1PL/1PCM and 2PL/2PCM combined models from PARSCALE were analyzed. Added to that assessment was the 1PL/1PCM from RUMM and the Exam, School, and

Provincial estimates. Comparisons of different model ability estimates before score transformation to letter grades showed that PARSCALE had the highest association with EXAM, while RUMM at the lowest association with EXAM estimates. However, all software and models had a strong association overall with the Exam estimates of ability. Once again, this suggests that the estimates were not equal and if IRT models are to be used with the math examinations, the same model and software program should be used in order to ensure consistency in results. It should be noted that a strong association was expected between the Exam estimates and the IRT estimates because the IRT estimates were based on the Exam results.

In order to use an IRT model with the Provincial exam data, there must be a meaningful link established between the Ministry's letter grade scale and the IRT scale. The ability estimates from an IRT model should be a good estimate of the true scores for each examinee. However, when using empirical data the underlying true scores are not known and therefore, the IRT ability estimates were compared to the CTT raw scores for consistency of letter grade classification. The relationship between the raw scores, Ministry-approved letter grades, and the proficiency estimates from the IRT based models were compared next (consistency of classification). The consistency of classification results suggested that the 1PL/1PCM model combination had closer estimates than the 2PL/2PCM model combination with regards to the current system under CTT. It should be noted that the 1PL/1PCM model estimates should be closer to the exam scores because it is a non-linear function of those number correct scores. In order to replace the current system which uses raw scores and CTT based analysis with an IRT approach, the results of estimation of examinee ability levels must be consistent. The 1PL/1PCM model

combination should then be used because of its high consistency of classification when compared to the CTT Exam estimates and in this way is feasible.

The last section addressed the issues of the open-ended items and partial credit models. Master's partial credit model (1PCM) was developed from the 1-parameter (Rasch) model but is generalized for use with unidimensional polytomous data (1982). Muraki's generalized partial credit model (2PCM) is a unidimensional item response model for polytomous data (1997). The difference between the 1PCM and the 2PCM model is that the discriminating or slope parameter (a_j) is estimated by the latter. Both of these models were reviewed in detail, however, only the 1-parameter Rasch model was applied using the software program RUMM because it does not employ the generalized partial credit model. The computer program RUMM uses a Rasch unidimensional model for measurement and was used to analyze the 7 open ended items from the 1996 Provincial math examination. All of the item ICRFs were reviewed and over half were interpreted dichotomous in the sense that the middle categories had few occurrences. This resulted in low ability examinees getting the question incorrect at every level while high ability examinees obtained the correct and final answer to the entire question. Item-fit was assessed using a chi-square statistic and resulted in 6 of the 7 items being rejected as not fitting the 1-parameter Rasch model. The ICRFs from the 1-parameter partial credit model were found to be very meaningful for the interpretation of how the item categories were operating within an exam.

Conclusion

Overall, the feasibility of using IRT models with the January administration of the 1996 British Columbia grade 12 Provincial Math Examination was supported. This

feasibility investigation supports a number of conclusions: 1) There is a definite need for better tests for assessing unidimensionality. 2) The item-fit statistics show better a better fit of items with the 3PL model then the 1PL and 2PL models. Although the 1PL and 2PL models did not fit as well as the 3PL model there was not much difference when it came to the parameter invariance investigation which is of greater importance. Thus feasibility support of item-fit came from the conjunction with the parameter invariance investigation. This support was warranted because Hambleton et al. (1991) states that if item-fit has been achieved, then the assumption of parameter invariance should hold. Given the bi-directionality of the assumption, it can be assumed that the finding of parameter invariance implies item-fit. This section also highlights the fact that different software programs using the same models rejected different numbers of items as fitting. Therefore only one software program should be selected and used with the Provincial exam data. 3) Parameter invariance was approached which supports the feasibility investigation and in this way the assumptions underlying the item response models should hold. "The importance of the property of invariance of item and ability parameters cannot be overstated. This property is the cornerstone of item response theory and makes possible such important applications as equating, item banking, investigation of item bias, and adaptive testing" (Hambleton et al., 1991). 4) Estimates from the different software programs were found to be vary and if IRT models are to be used with the math examinations the same model and software program should be used in order to ensure consistency in results (not directly related to feasibility). 6) Feasibility was also supported with the use of the 1PL/1PCM model combination over the 2PL/2PCM model combination because of the high consistency of classification when compared to the

classical Exam estimates. 7) Lastly, it was found that the interpretation of how the open-ended items were behaving can be derived effectively by reviewing the item characteristic response functions and this should improve item development and selection for future exams.

The factors such as unidimensionality, item-fit, parameter invariance, and classification consistency were all used to assess the feasibility of analyses by both IRT binary and partial credit models with a Provincial math exam data set. Note that unidimensionality and item-fit were reviewed only has a check for the more important assumption of parameter invariance. Most of these factors have been largely met and most models found to produce meaningful results.

The recommendations for IRT use with a current Provincial exam program that is already in place would be for the 1PL/1PCM combined model within the software program RUMM to be used with the Provincial Math data in order to maintain consistency in measurement. The use of the 1PL/1PCM is supported by the consistency of classification studies and the parameter invariance investigation. The RUMM program was chosen over the PARSCALE program based on the ease of use (windows-based) and the production of item characteristic response functions inherent in the RUMM program. The PARSCALE program also has a limited number of categories and cannot estimate items with over 20 categories. Note that PARSCALE simply could not handle the data well. However, this is a complex area and there is no simple answer. These findings lend support to the feasibility of using both dichotomous and polytomous IRT applications with the Provincial math examinations. These IRT models could then be used to

facilitate the use of item banks to develop and equate Provincial math exams that use both multiple choice and open-ended items.

Most of the preliminary analyses in regards to IRT (model assumptions) have come back positive. However, future research should investigate the equating of actual item bank produced exams. This research will have to include an exam that has been generated from a Provincial item bank. If the current system of CTT remains, then the creation of these new exams would still require a standard setting committee to be assembled each time an exam is created in order to set cut-points for each letter grade category (informal equating). Future research then, should examine whether or not analyses based on item response theory may be used to determine these cut-points that have been traditionally selected based on expert opinion. If this is found possible, then exams may be equated one administration to another in an item banking situation through statistical procedures.

References

Anderson, J. O., Muir, W., Bateson, D. J., Blackmore, D., & Rogers, W. T. (1990). *The impact of Provincial examinations on education in British Columbia: General report.* Victoria, B.C. British Columbia Ministry of Education.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Eds.), *Educational Measurement* (2nd) (pp. 508-600). New York: Macmillan.

Bock, R. D., Thissen, D., & Zimowski. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, 34, 197-211.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory.* NewYork: Holt, Rinehart & Winston.

De Ayala, R.J. (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Counseling and Development*, 25, 172-189.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement,* 31, 295-311.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8, 341-349.

British Columbia, Ministry of Education. (1992). *The exam development process.* Victoria: Ministry of Education, B.C.

British Columbia, Ministry of Education. (1994). *Guidlines for student reporting for the kindergarten to grade 12 education plan.* Victoria: Ministry of Education, B.C.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Klumer Nijhoff.

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

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

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Linn, R.L., & Gronlund, N.E. (1995). *Measurement and assessment in teaching* (7th ed.). Toronto, Ont: Prentice-Hall.

Livingston, S.A. (1988). Reliability of test results. In J.P. Keeves (Eds.), *Educational Research, Methodology, and Measurement: An International Handbook* (pp. 386-392). Toronto, Ont: Pergamon Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillside, New Jersey: Lawrence Erlbaum Associates.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-173.

Masters, G. N. (1990). Partial credit model. In H. J. Walberg & G. D. Haertel (Eds.), *The International Encyclopedia of Educational Evaluation* (pp. 388-393). NewYork: Pergamon Press.

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.

McKinley, R. L. (1989). An introduction to item response theory. *Measurement and Evaluation in Counseling and Development*, 22, 37-57.

Mislevy, R. J., & Bock, R. D. (1990). *PC BILOG 3: Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software, INC.

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., & Bock, R. D. (1996). *PARSCALE 3: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks* [Computer program]. Chicago, IL: Scientific Software.

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. 153-164). New York: Springer-Verlag Inc.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement*, 28, 99-117.

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

Oltman, P., Stricker, L., & Barrows, T. (1990). Analyzing test structure by multidimensional scaling. *Journal of Applied Psychology*, 75, 21-27.

Samejima, F. (1997). Graded response model. In Wim J. van der Linden & Ronald K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85-100). New York: Springer-Verlag Inc.

Sheridan, B., Andrich, D., & Luo, G. (1997). RUMM [computer program] Murdoch University.

SPSS Inc. (1996). SYSTAT 7.0. Chicago: SPSS Inc.

Stout, W., Douglas, J., Junker, B., & Roussos, L. (1993). DIMTEST [computer program]. Department of Statistics, University of Illinois at Urbana-Champaign.

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillside, NJ: Lawrence Erlbaum Associates.

APPENDIX A

Factor Loading Matrix

Table A-1

Factor Loading Matrix

Items	Factor 1 loadings	Factor 2 loadings
1	.25	-.19
2	.26	-.26
3	.24	-.26
4	.33	-.26
5	.32	-.27
6	.30	-.22
7	.30	-.17
8	.32	-.26
9	.31	-.24
10	.32	-.18
11	.32	-.13
12	.30	-.12
13	.37	-.18
14	.36	-.16
15	.33	-.15
16	.34	-.11
17	.35	-.09
18	.37	-.08
19	.38	-.14
20	.35	-.14
21	.32	-.06
22	.38	-.04
23	.38	-.07
24	.35	-.08
25	.36	.00
26	.37	.03
27	.37	.06
28	.39	.05
29	.32	.05
30	.32	.04
31	.35	.12
32	.36	.14
33	.35	.10
34	.37	.04
35	.34	.10
36	.32	.15

Table A-1 cont'd

Factor Loading Matrix

Items	Factor 1 loadings	Factor 2 loadings
37	.36	.13
38	.34	.12
39	.31	.14
40	.31	.20
41	.32	.20
42	.31	.22
43	.31	.22
44	.27	.25
45	.27	.34
46	.31	.29
47	.31	.25
48	.25	.22
49	.24	.30
50	.24	.31

APPENDIX B

Item Characteristic Curves

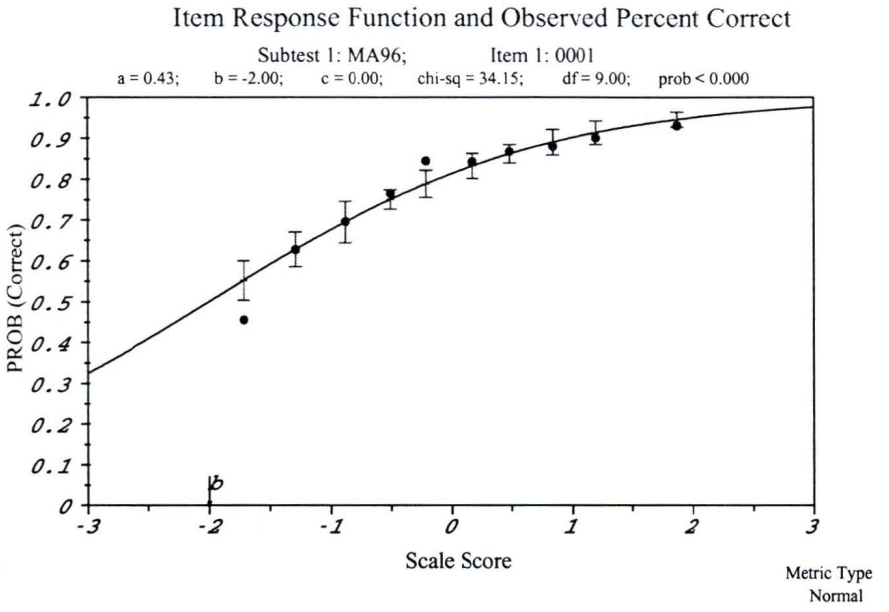


Figure B1. 2PL model ICC showing item fit from BILOG for item 1.

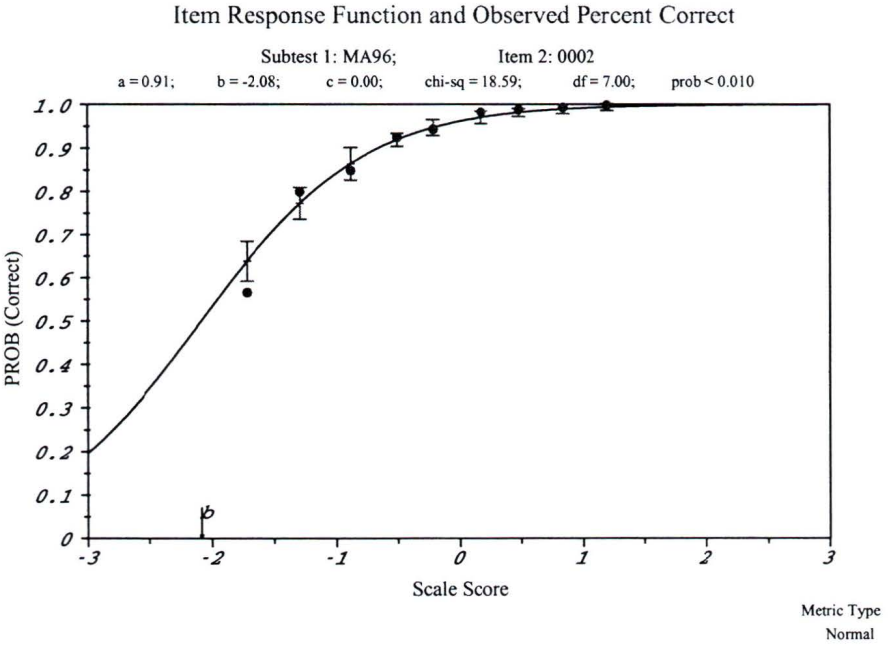


Figure B2. 2PL model ICC showing item fit from BILOG for item 2.

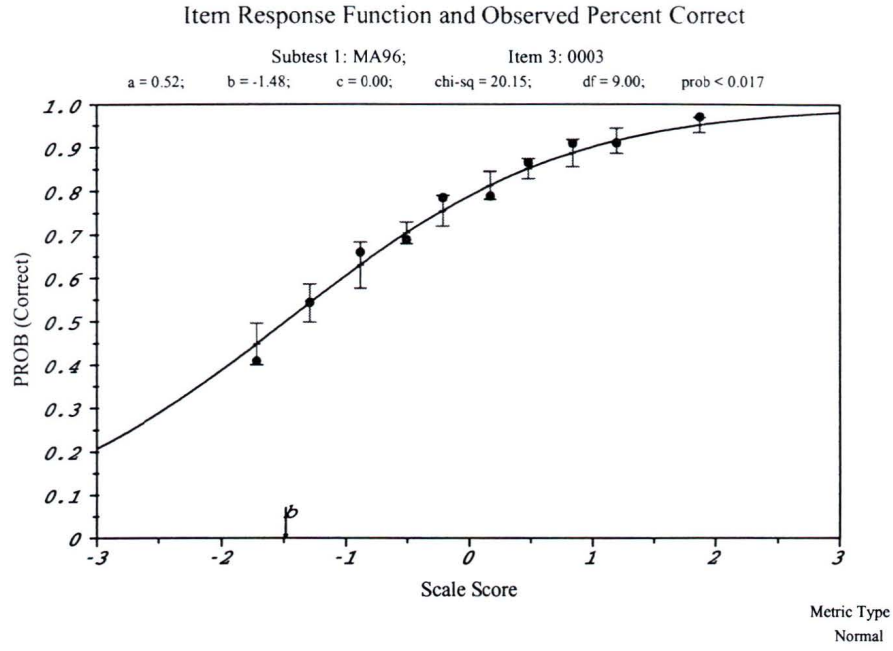


Figure B3. 2PL model ICC showing item fit from BILOG for item 3.

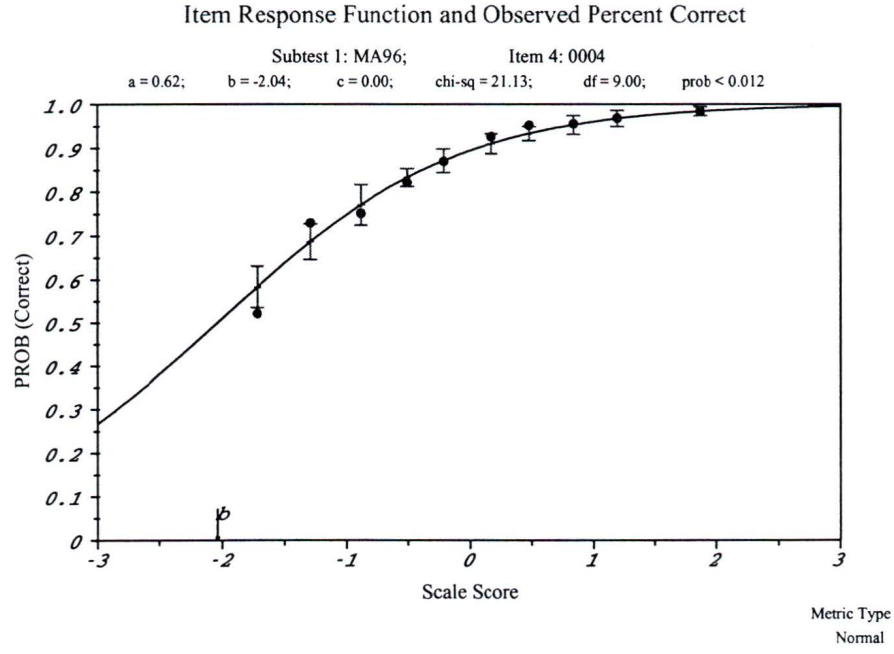


Figure B4. 2PL model ICC showing item fit from BILOG for item 4.

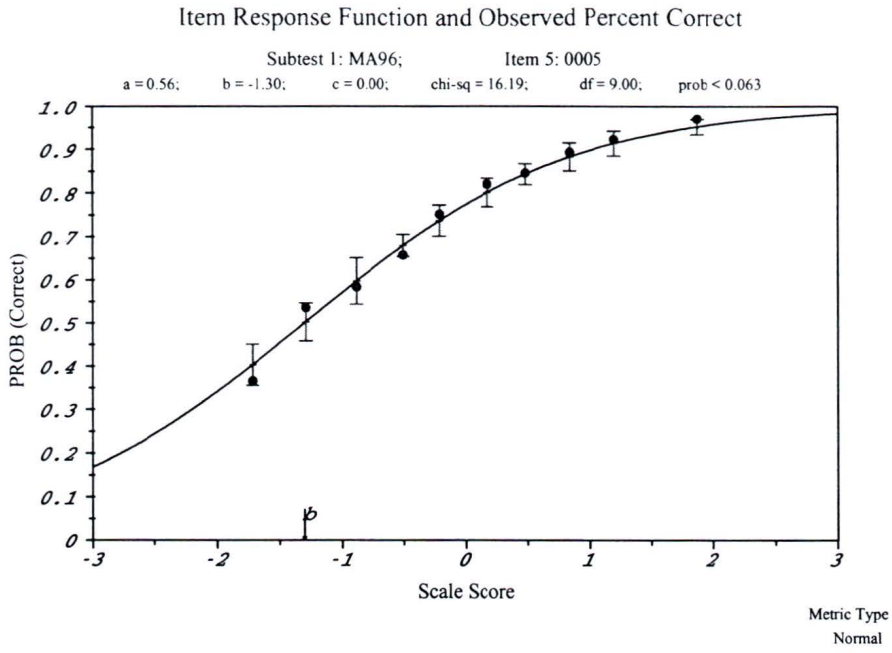


Figure B5. 2PL model ICC showing item fit from BILOG for item 5.

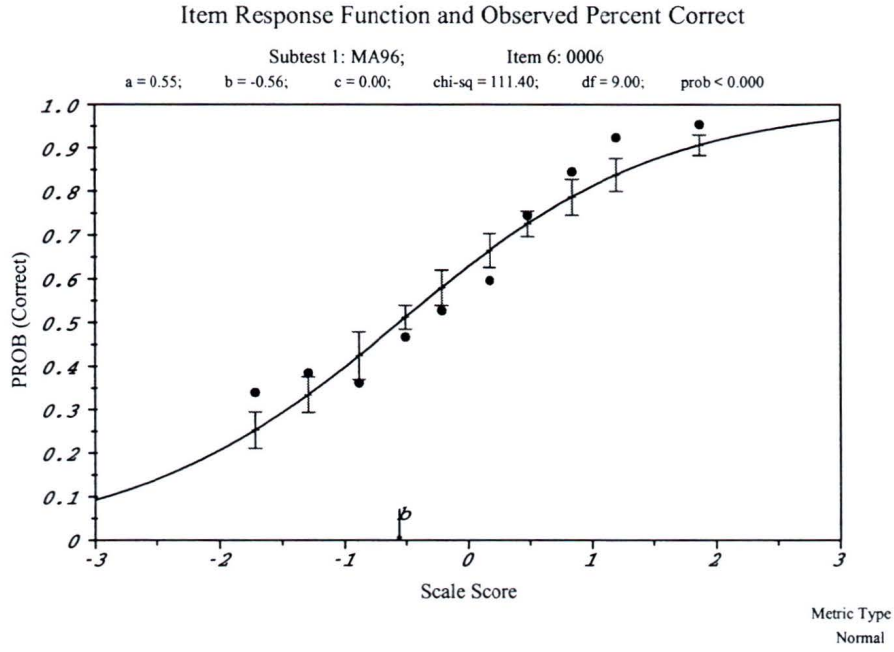


Figure B6. 2PL model ICC showing item fit from BILOG for item 6.

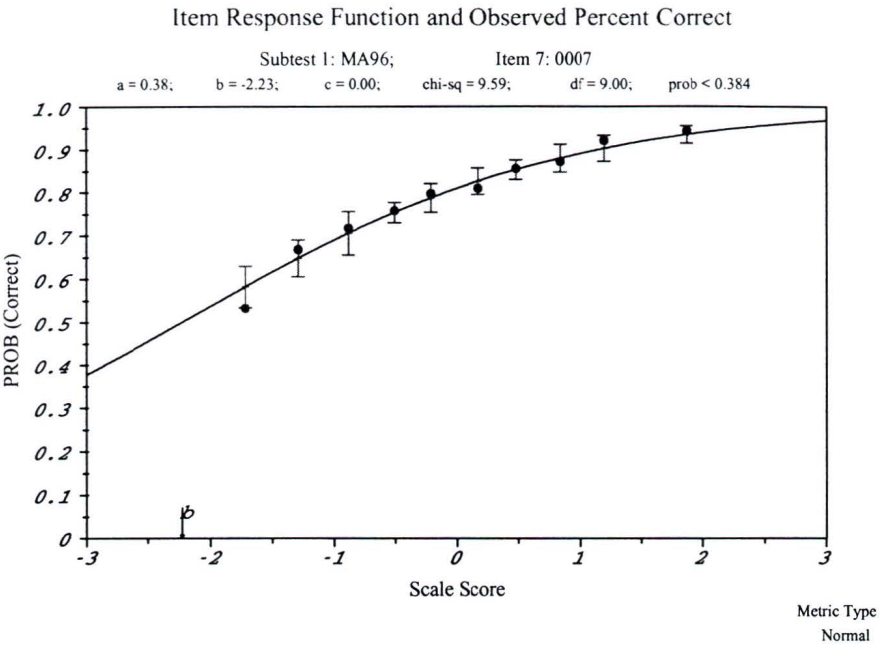


Figure B7. 2PL model ICC showing item fit from BILOG for item 7.

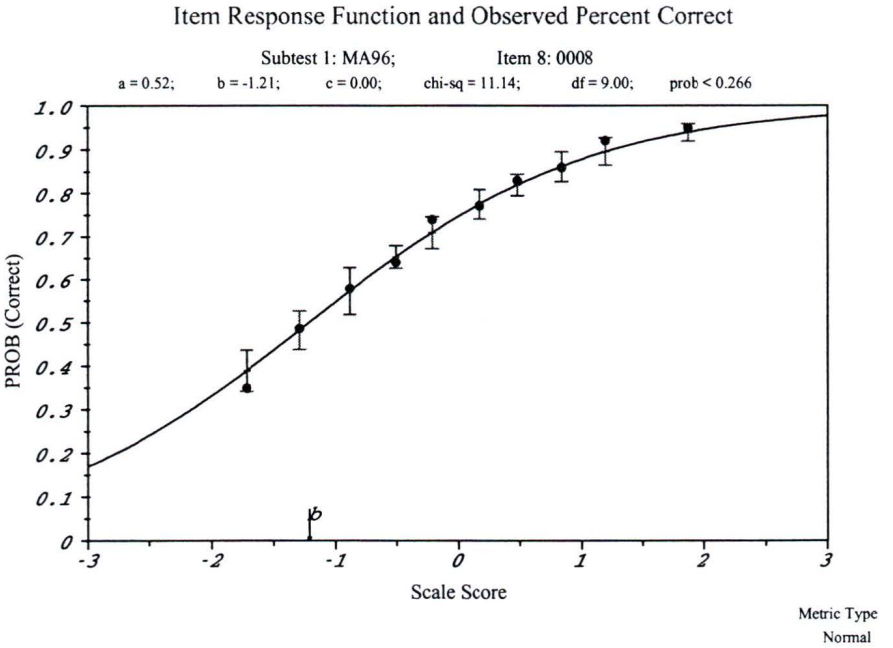


Figure B8. 2PL model ICC showing item fit from BILOG for item 8.

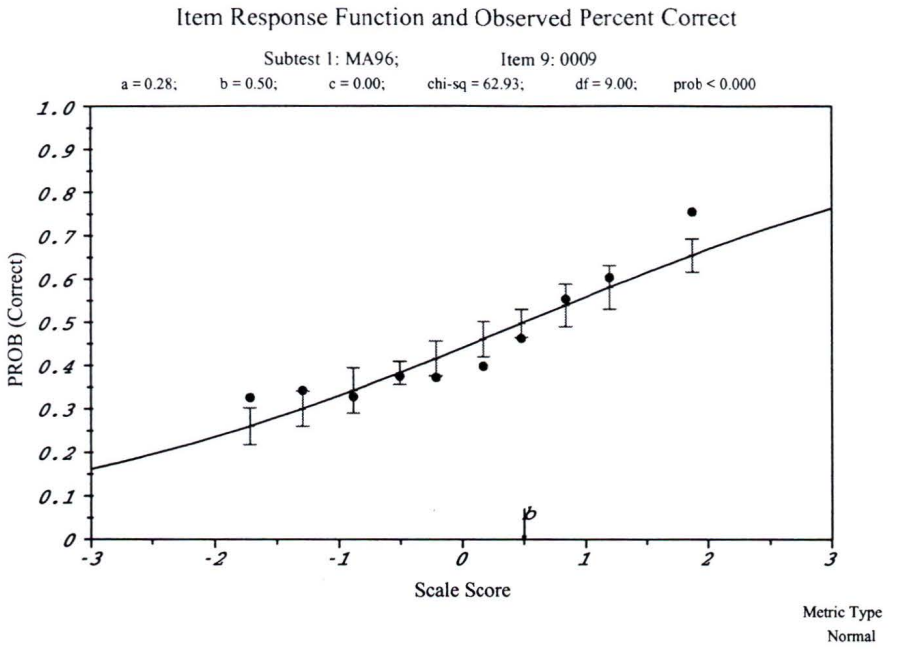


Figure B9. 2PL model ICC showing item fit from BILOG for item 9.

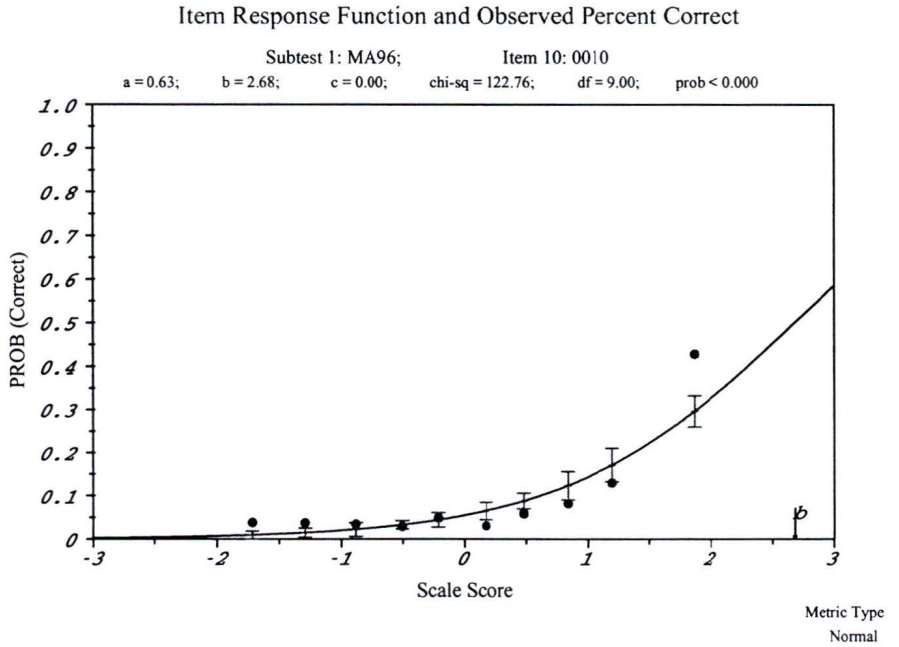


Figure B10. 2PL model ICC showing item fit from BILOG for item 10.

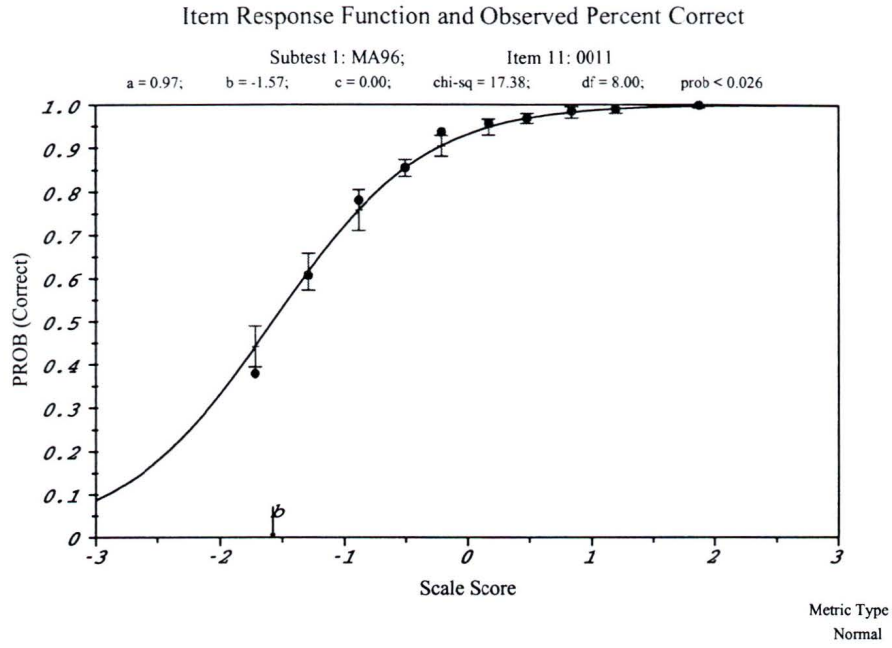


Figure B11. 2PL model ICC showing item fit from BILOG for item 11.

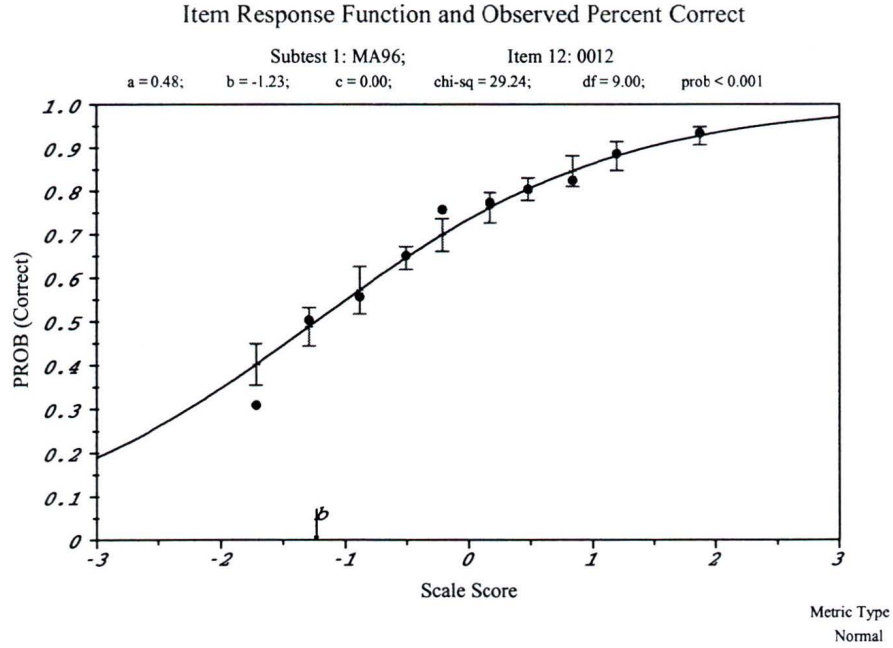


Figure B12. 2PL model ICC showing item fit from BILOG for item 12.

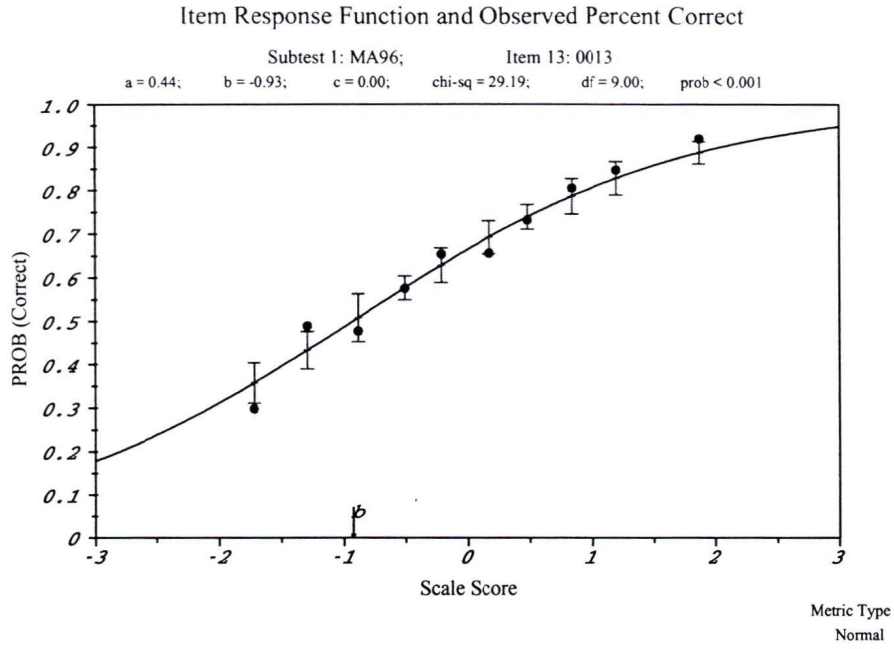


Figure B13. 2PL model ICC showing item fit from BILOG for item 13.

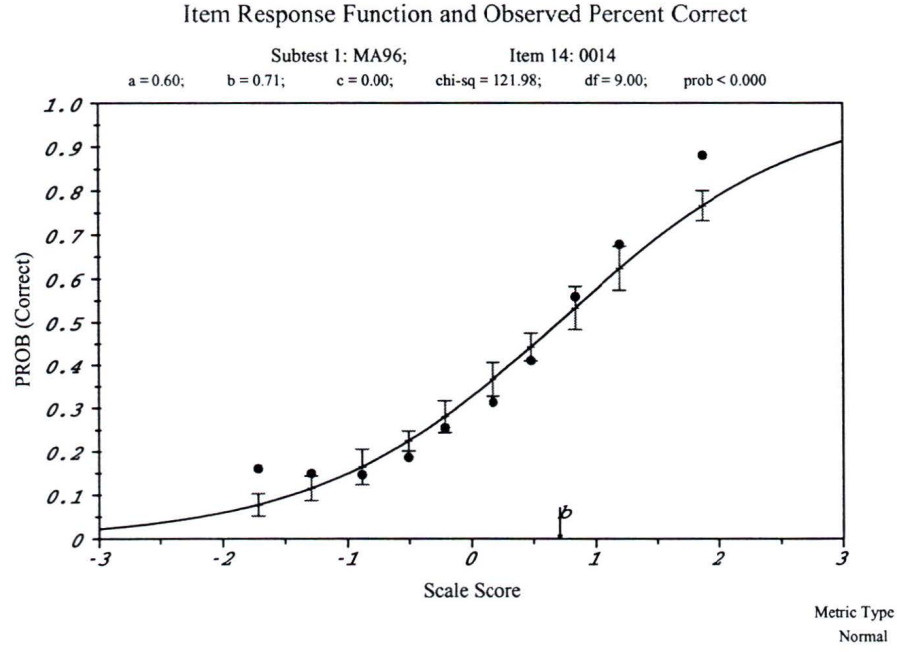


Figure B14. 2PL model ICC showing item fit from BILOG for item 14.

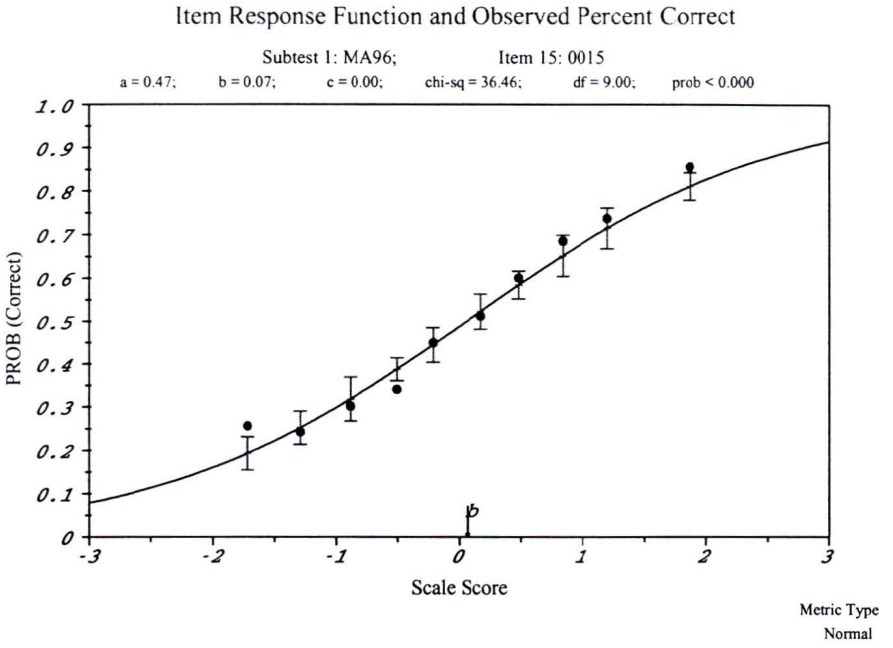


Figure B15. 2PL model ICC showing item fit from BILOG for item 15.

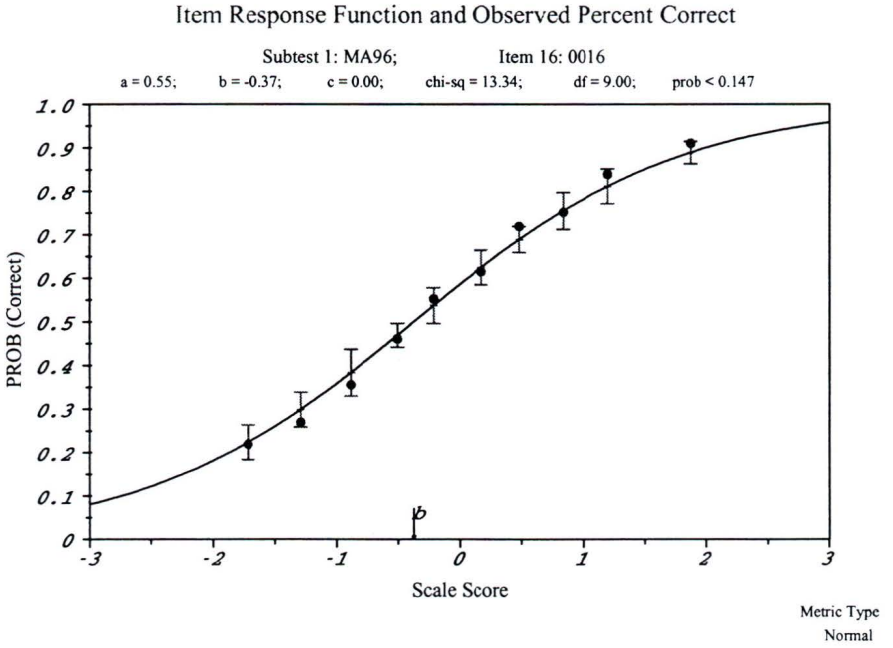


Figure B16. 2PL model ICC showing item fit from BILOG for item 16.

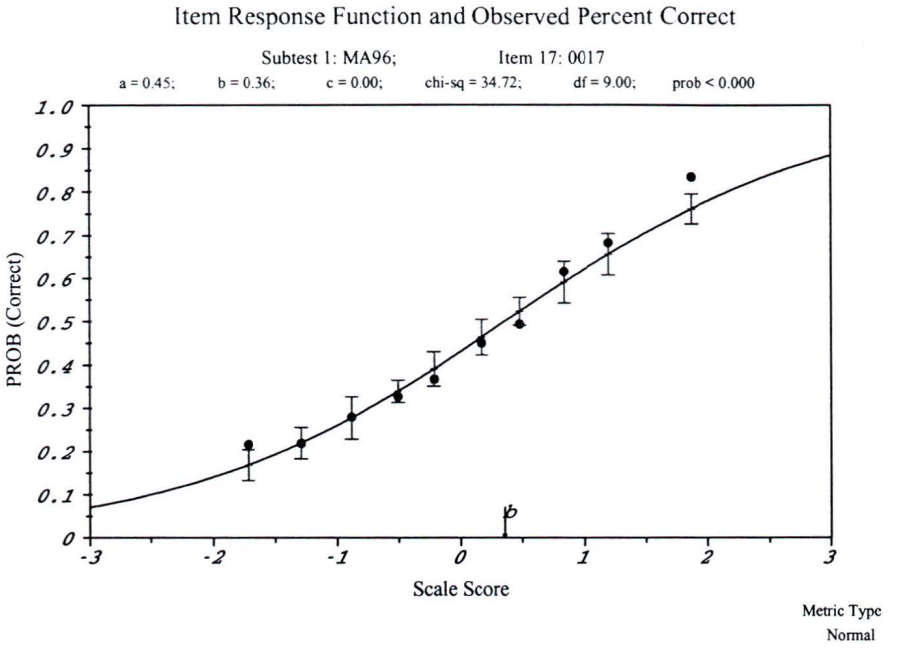


Figure B17. 2PL model ICC showing item fit from BILOG for item 17.

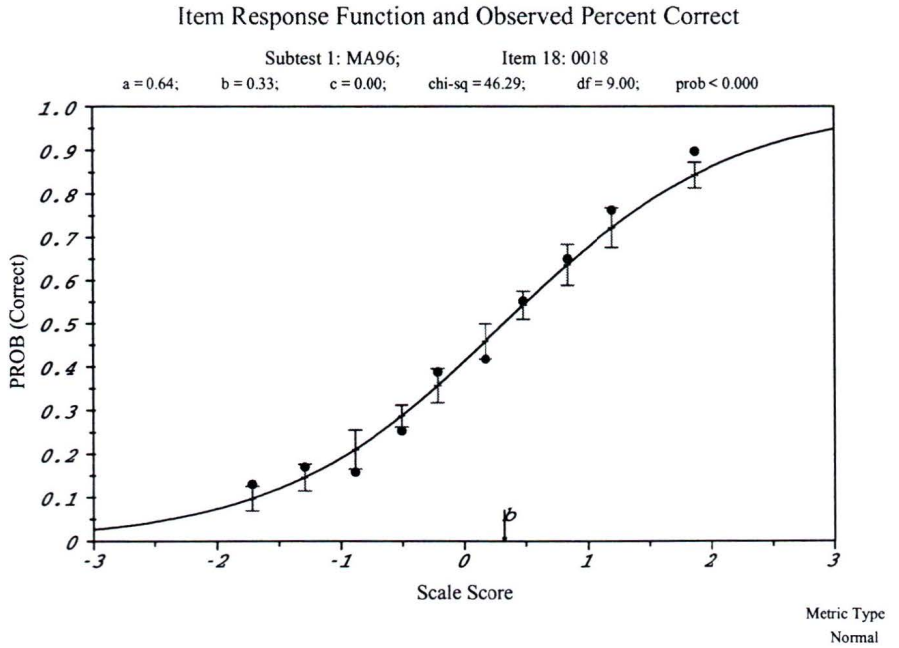


Figure B18. 2PL model ICC showing item fit from BILOG for item 18.

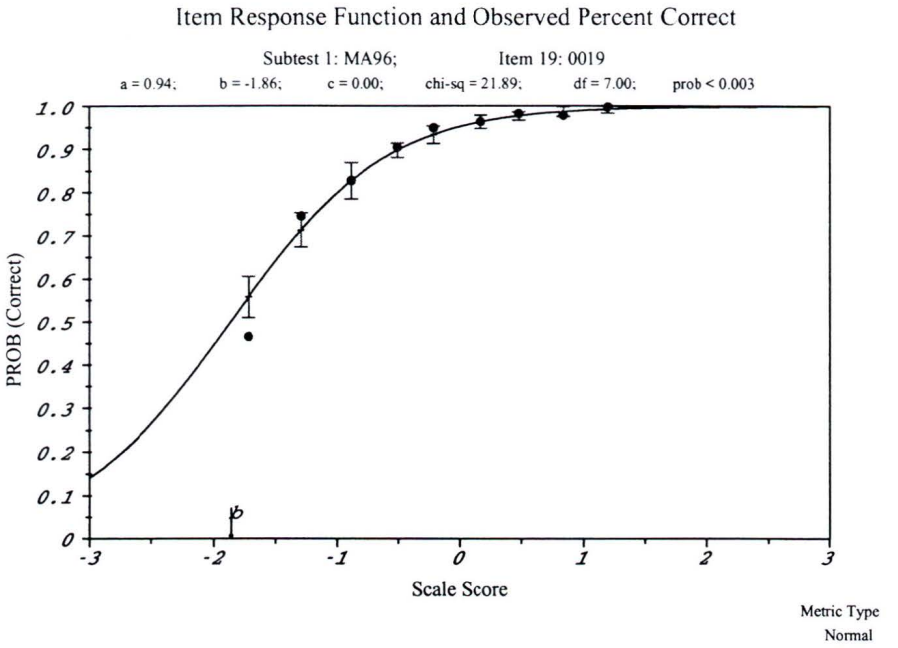


Figure B19. 2PL model ICC showing item fit from BILOG for item 19.

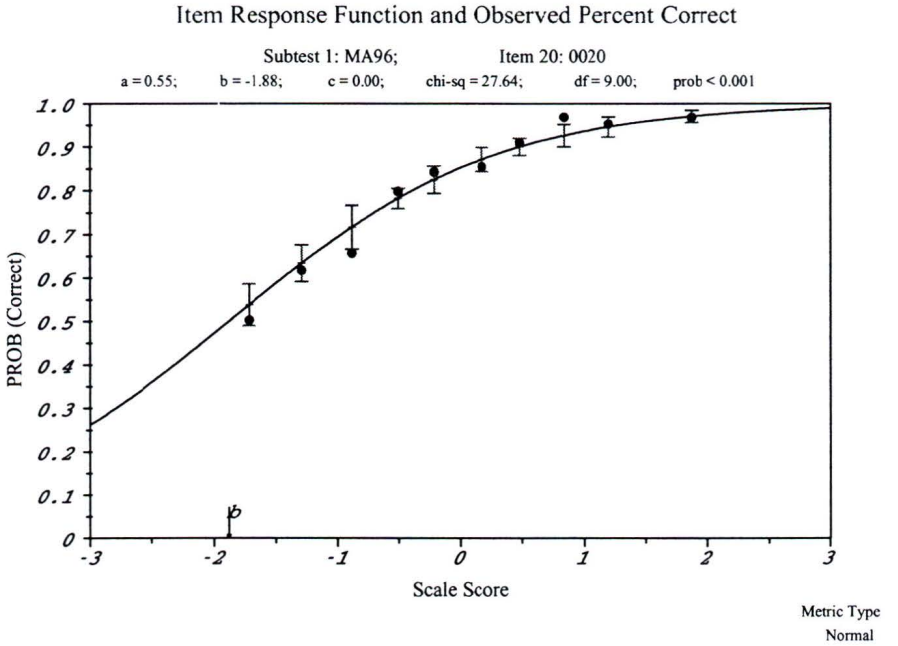


Figure B20. 2PL model ICC showing item fit from BILOG for item 20.

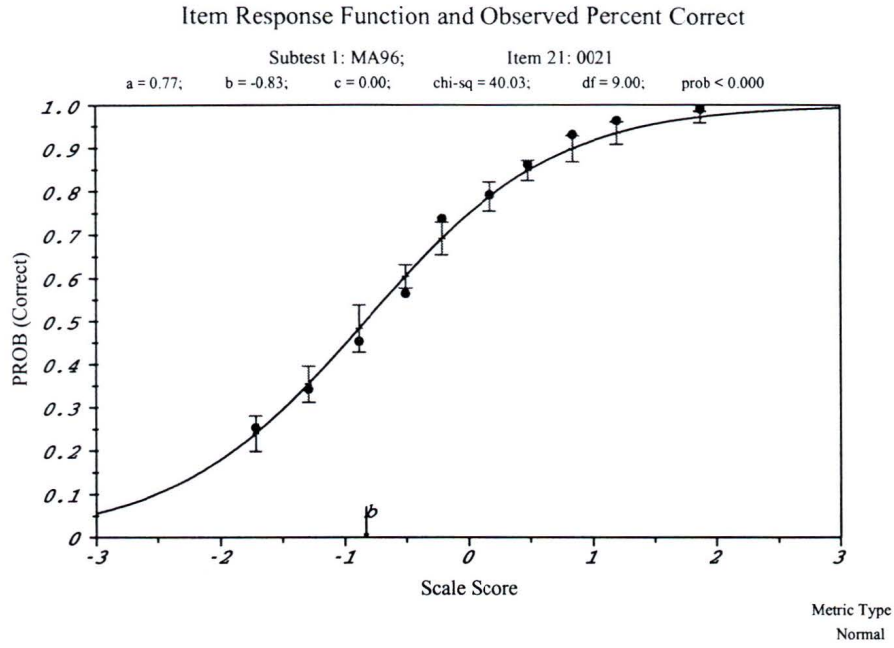


Figure B21. 2PL model ICC showing item fit from BILOG for item 21.

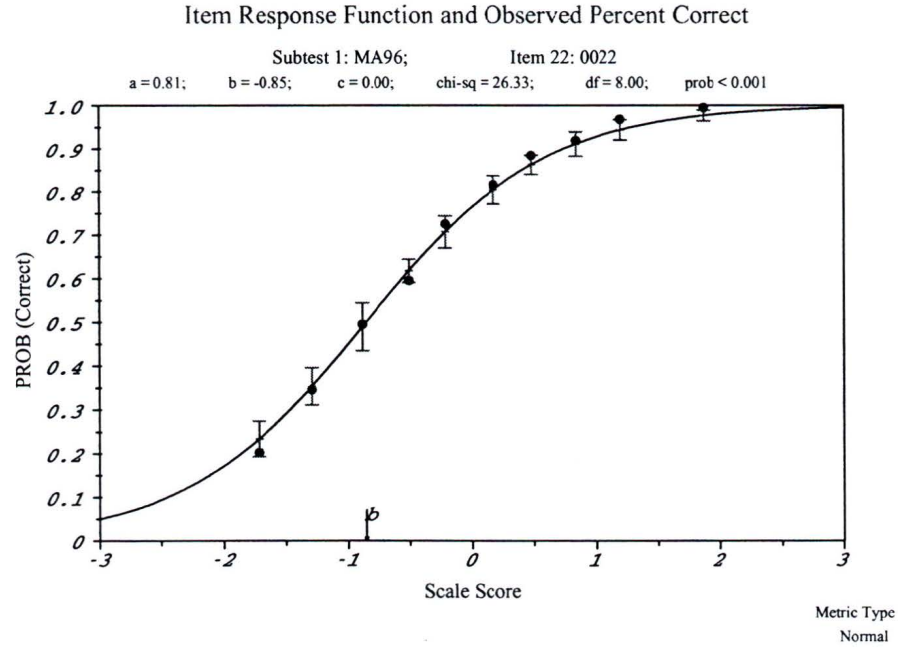


Figure B22. 2PL model ICC showing item fit from BILOG for item 22.

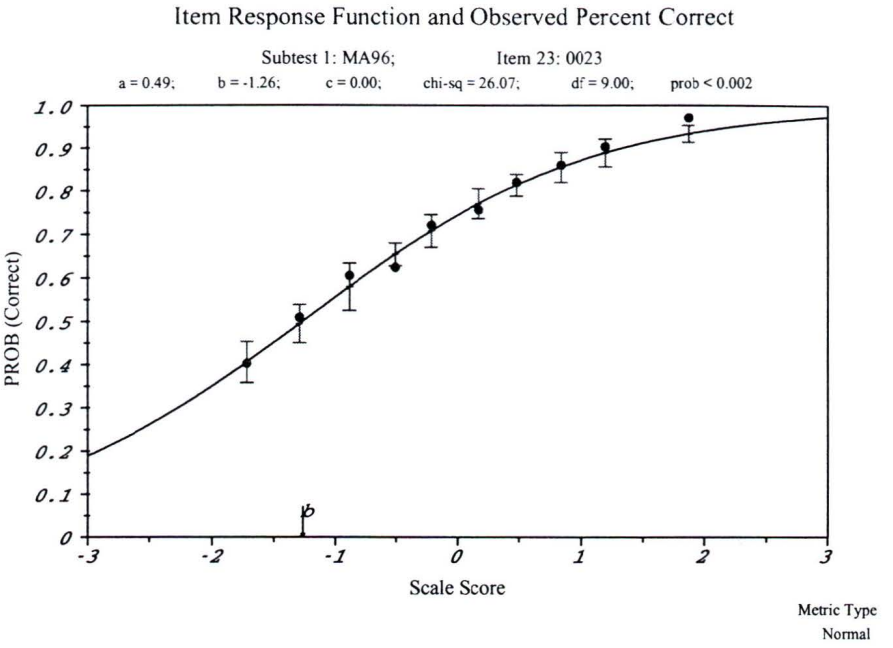


Figure B23. 2PL model ICC showing item fit from BILOG for item 23.

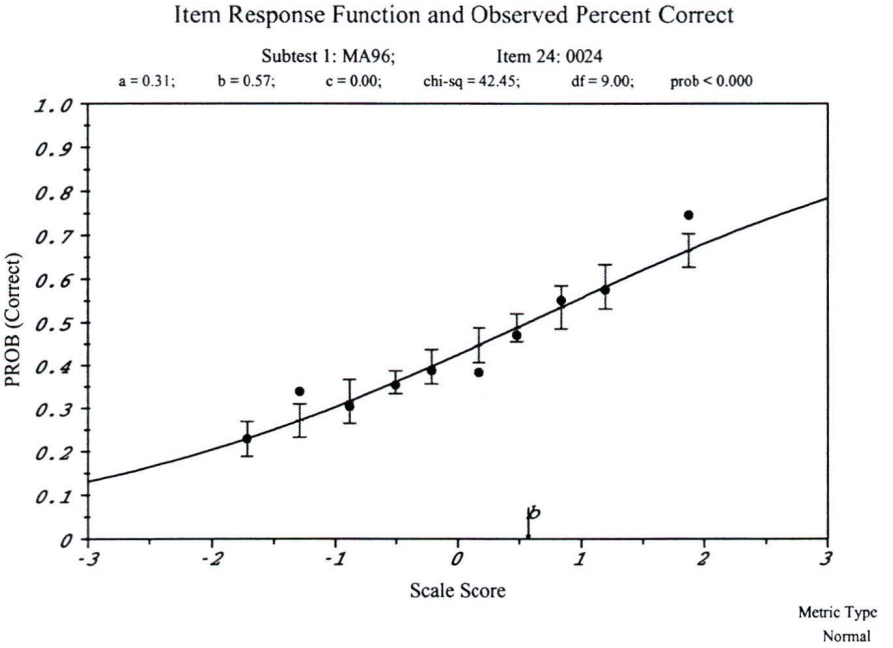


Figure B24. 2PL model ICC showing item fit from BILOG for item 24.

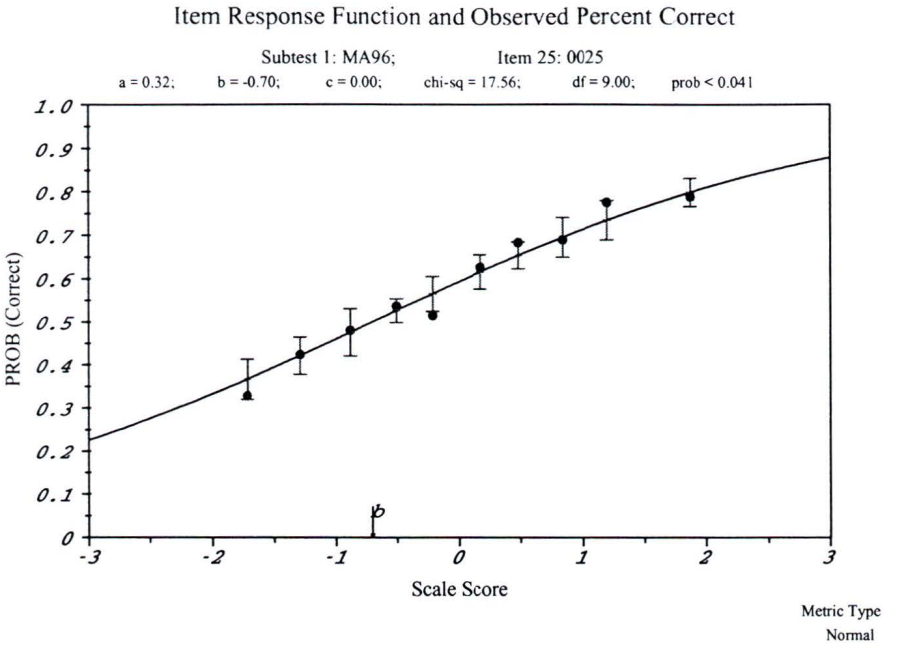


Figure B25. 2PL model ICC showing item fit from BILOG for item 25.

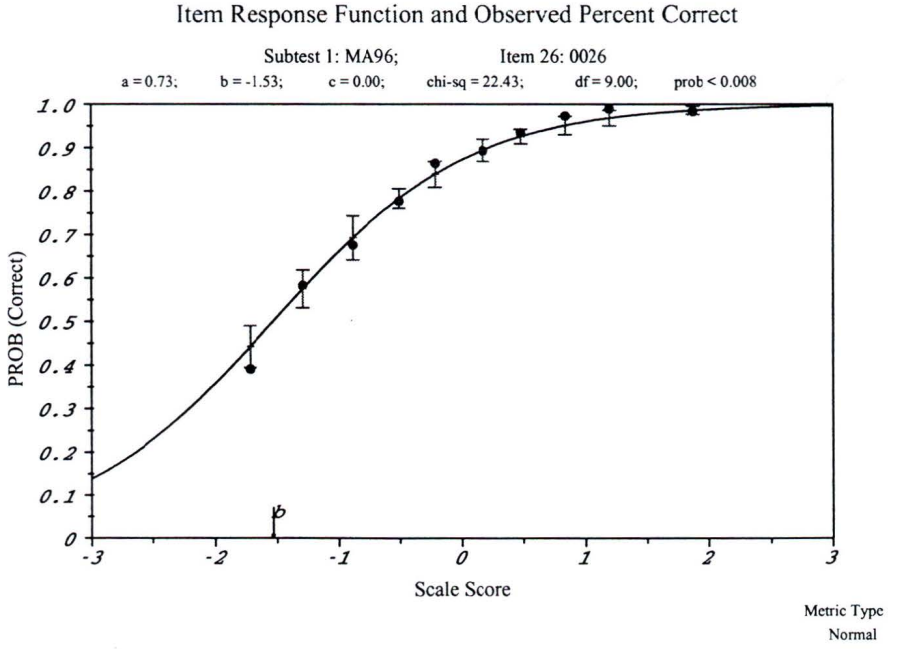


Figure B26. 2PL model ICC showing item fit from BILOG for item 26.

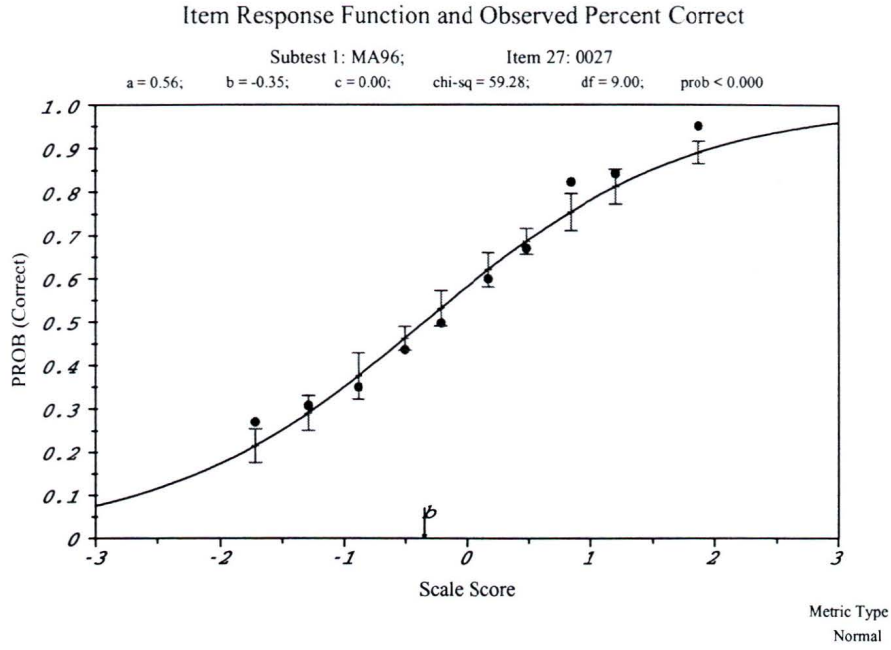


Figure B27. 2PL model ICC showing item fit from BILOG for item 27.

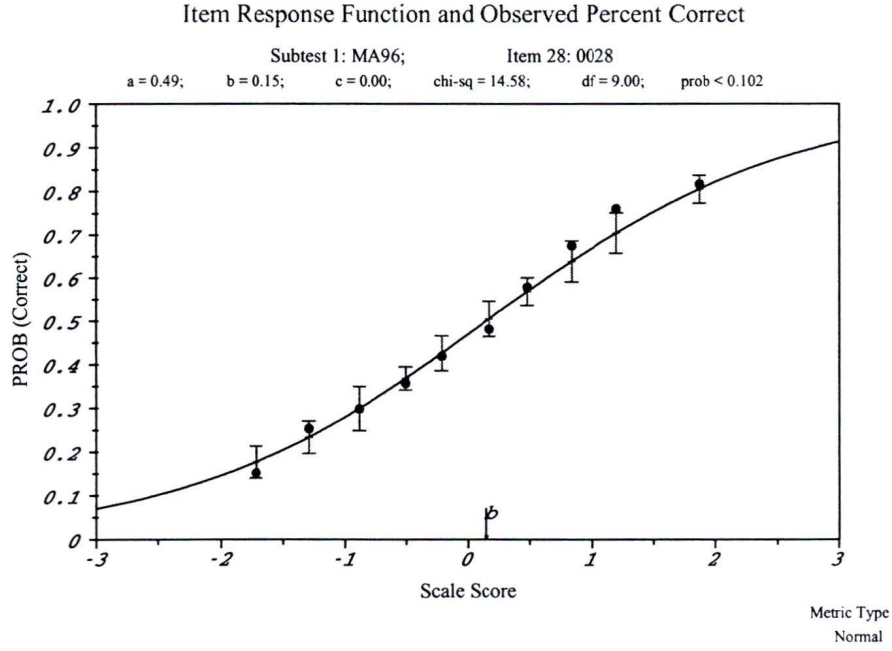


Figure B28. 2PL model ICC showing item fit from BILOG for item 28.

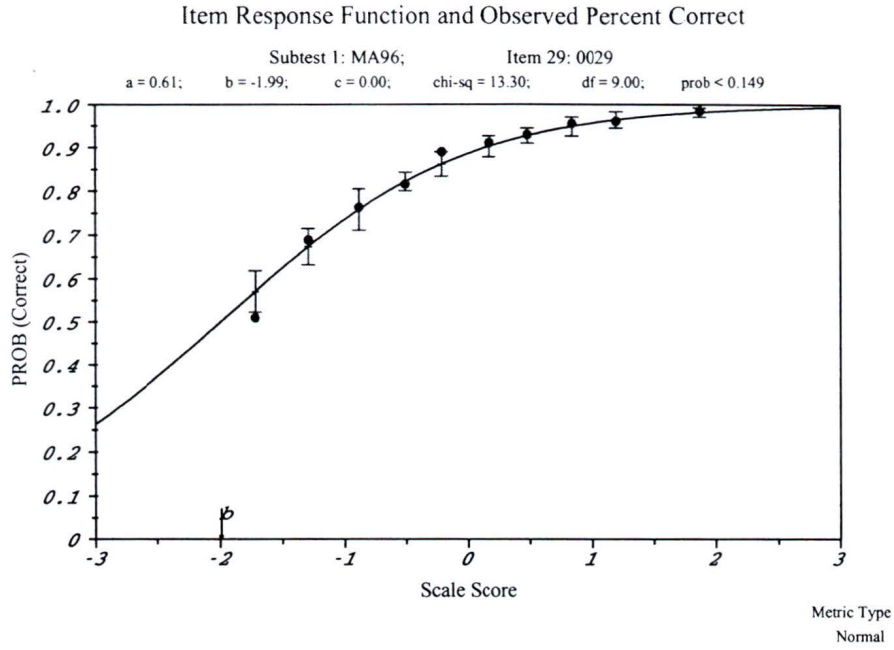


Figure B29. 2PL model ICC showing item fit from BILOG for item 29.

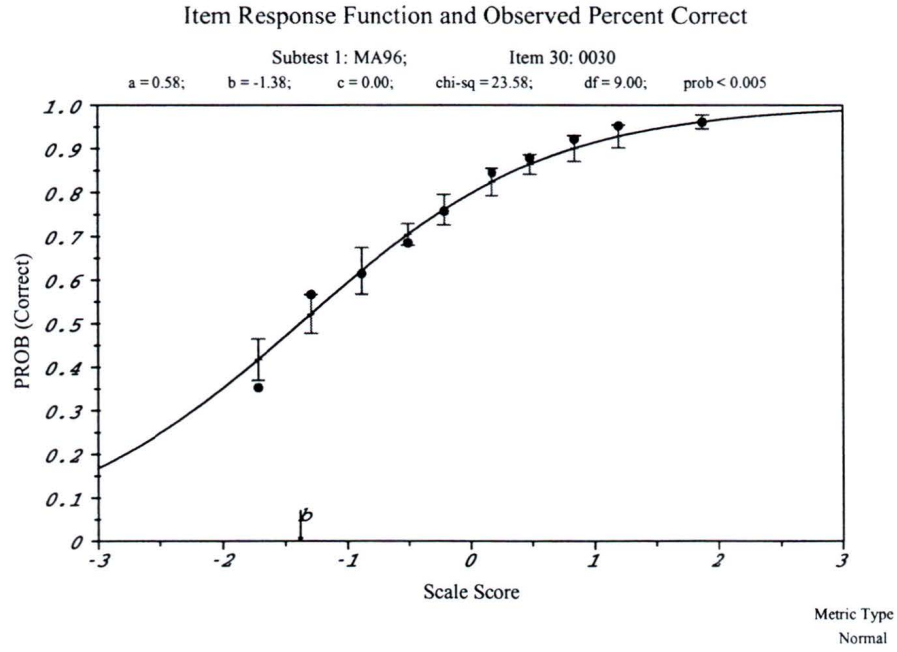


Figure B30. 2PL model ICC showing item fit from BILOG for item 30.

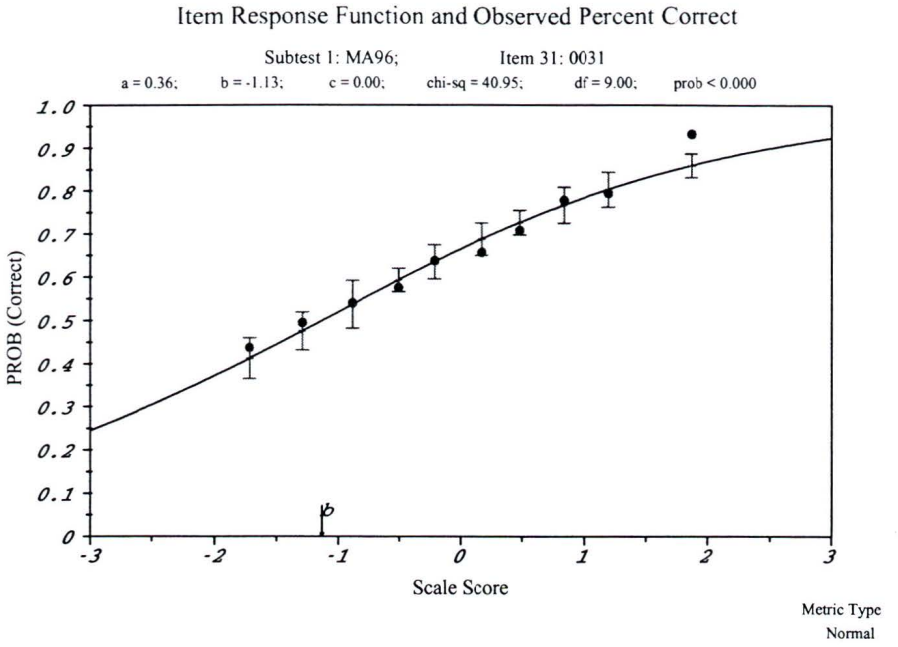


Figure B31. 2PL model ICC showing item fit from BILOG for item 31.

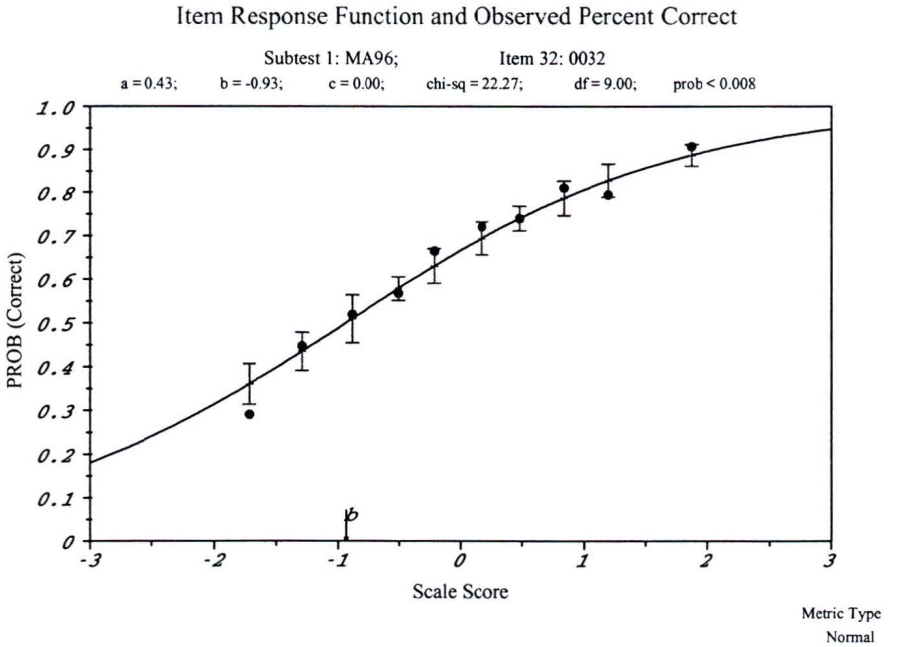


Figure B32. 2PL model ICC showing item fit from BILOG for item 32.

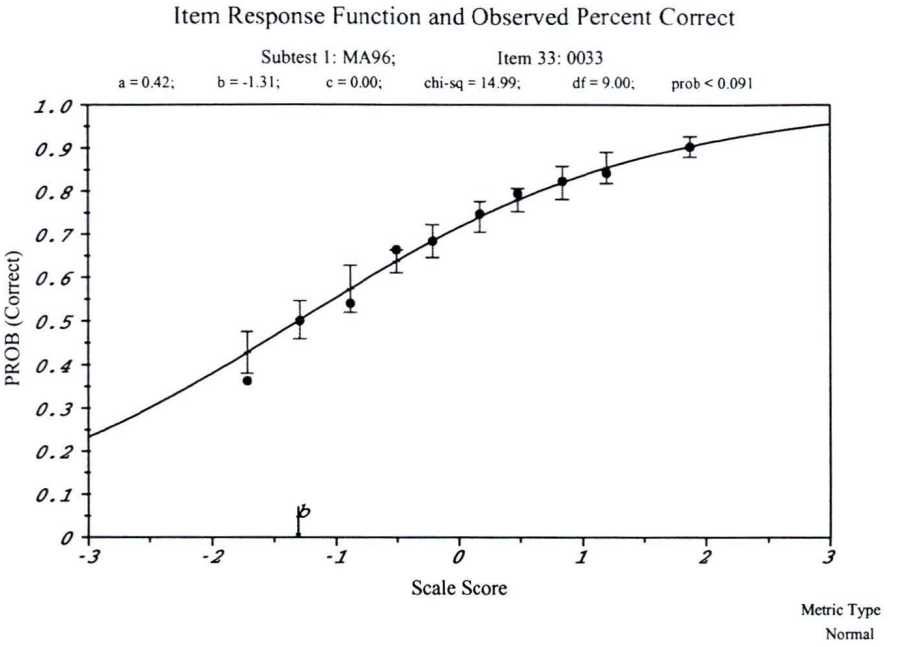


Figure B33. 2PL model ICC showing item fit from BILOG for item 33.

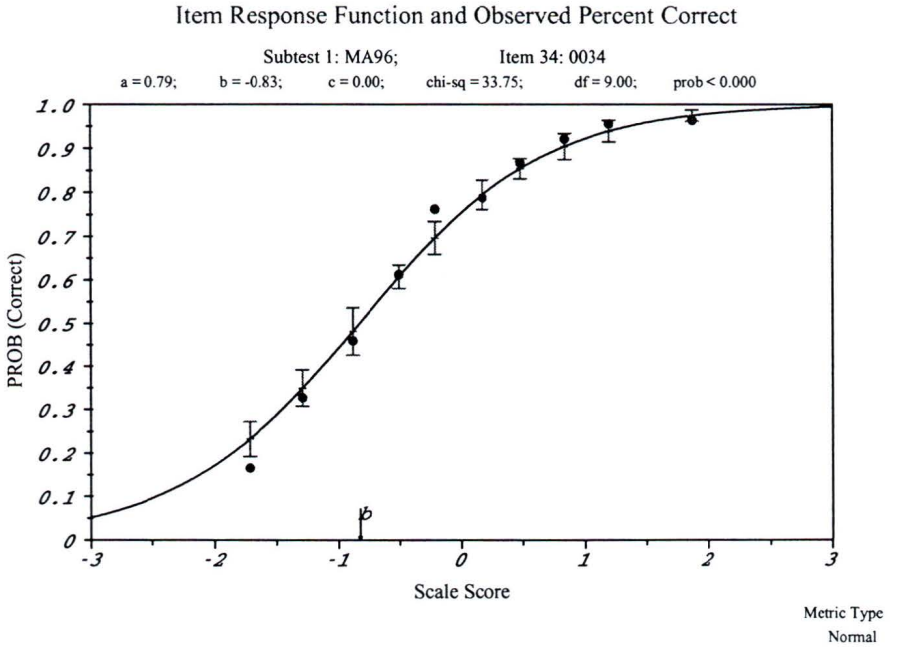


Figure B34. 2PL model ICC showing item fit from BILOG for item 34.

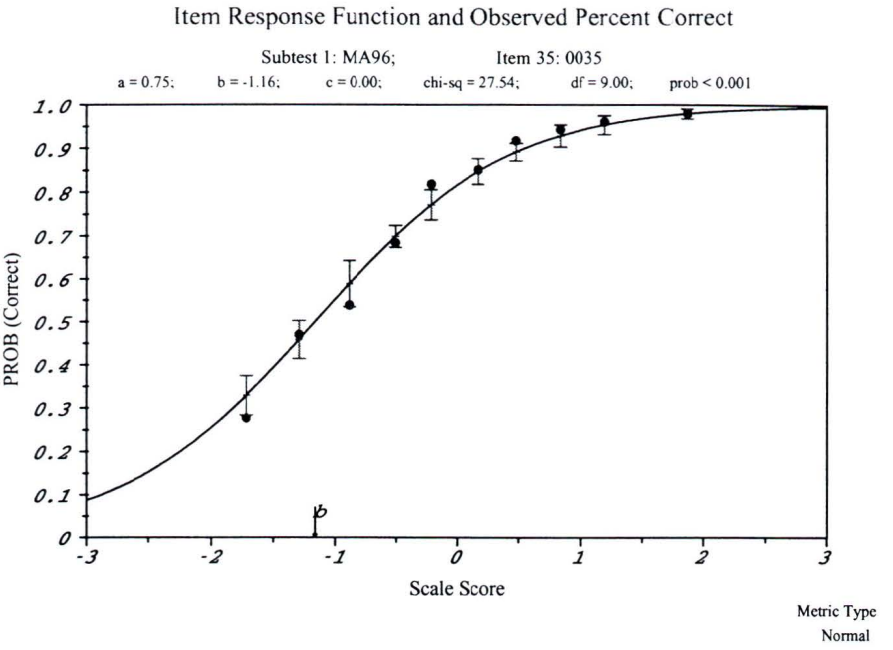


Figure B35. 2PL model ICC showing item fit from BILOG for item 35.

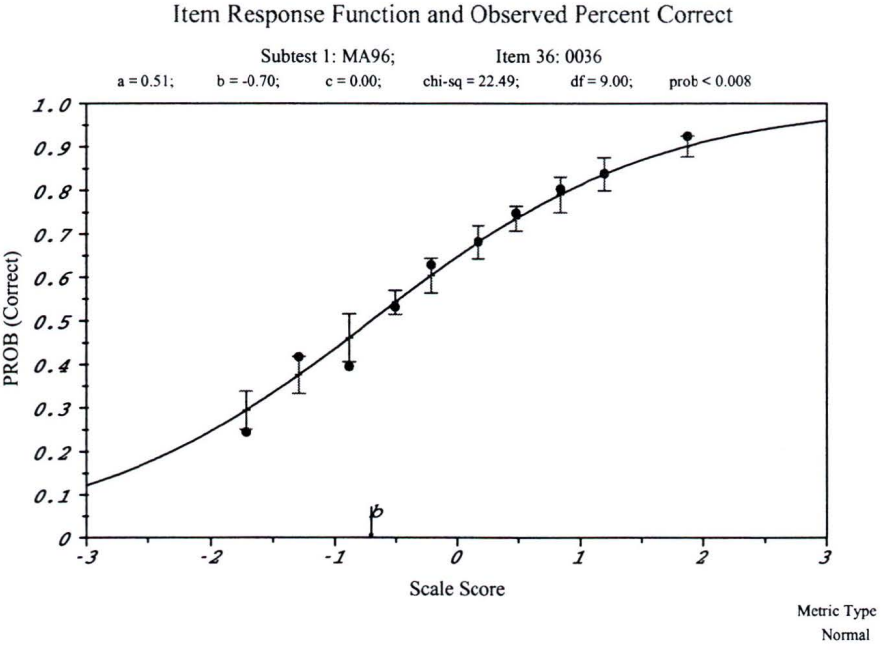


Figure B36. 2PL model ICC showing item fit from BILOG for item 36.

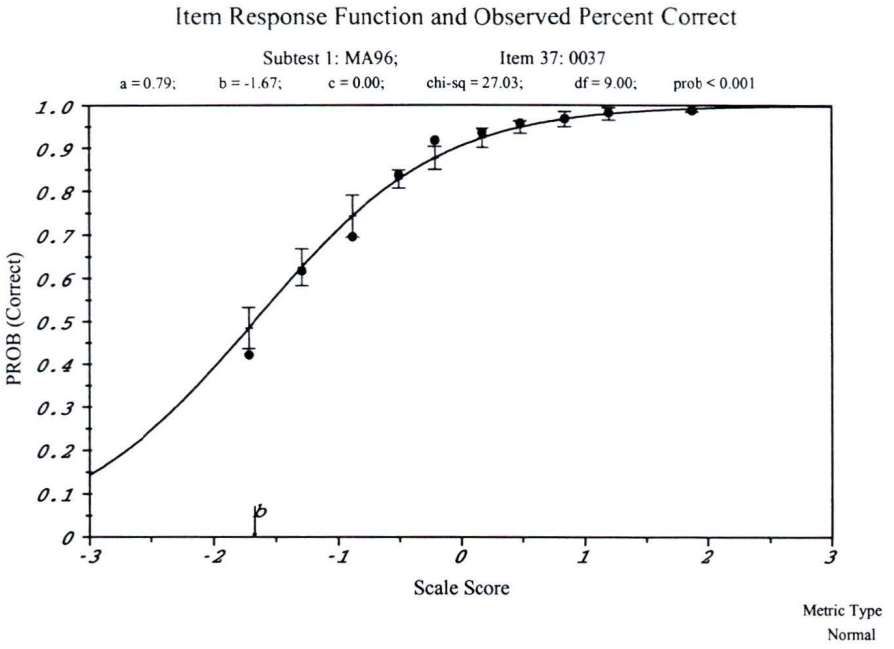


Figure B37. 2PL model ICC showing item fit from BILOG for item 37.

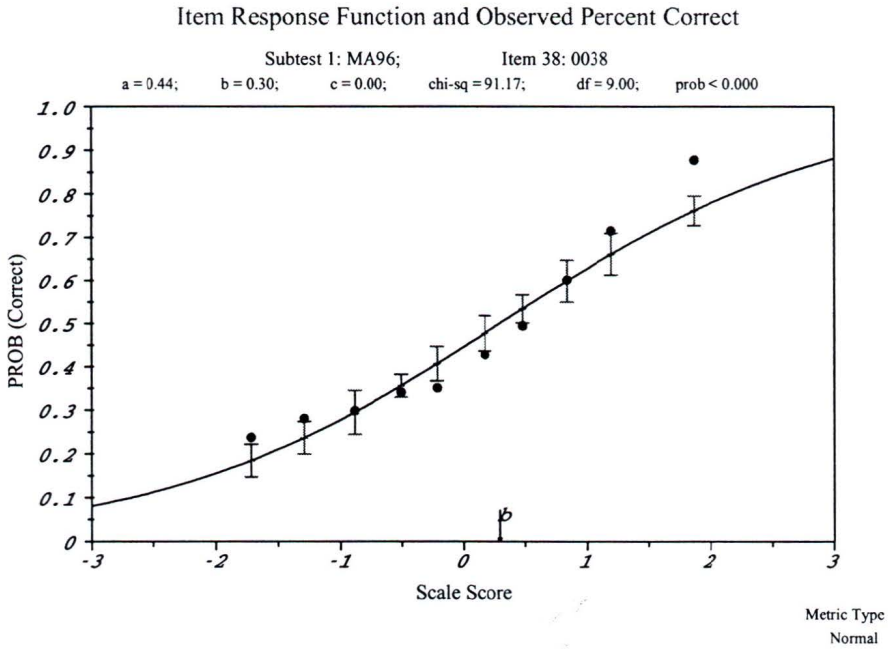


Figure B38. 2PL model ICC showing item fit from BILOG for item 38.

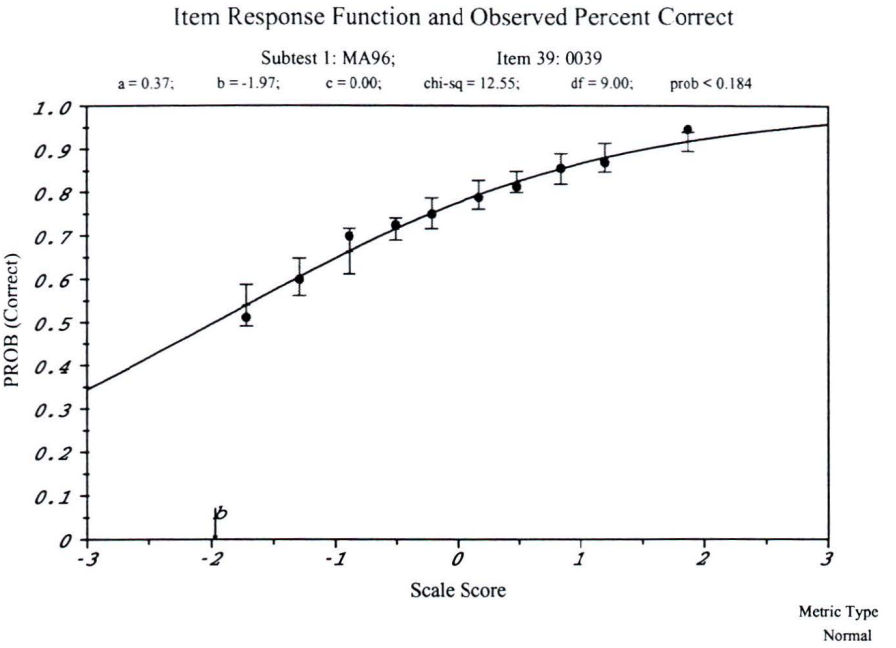


Figure B39. 2PL model ICC showing item fit from BILOG for item 39.

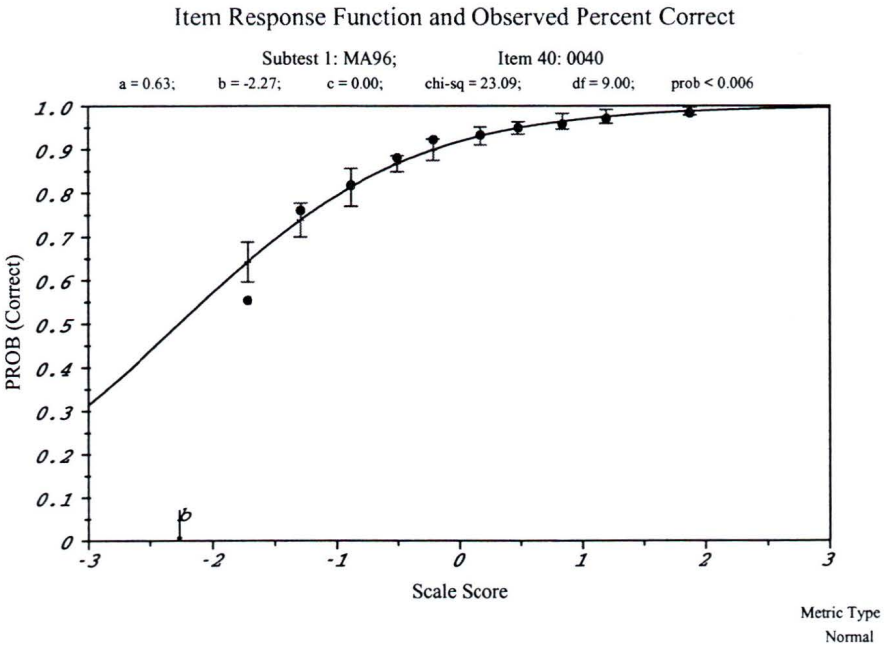


Figure B40. 2PL model ICC showing item fit from BILOG for item 40.

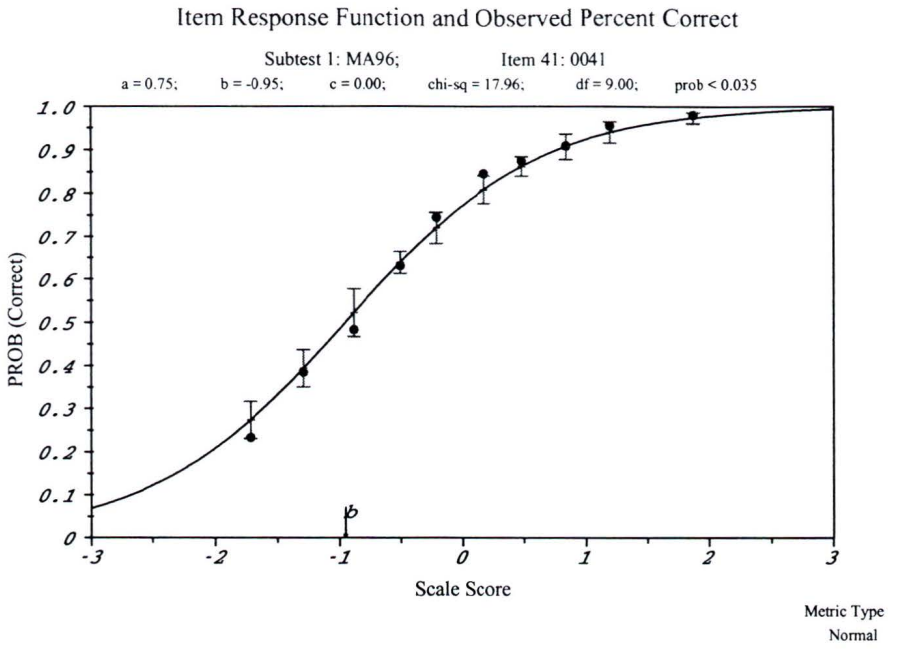


Figure B41. 2PL model ICC showing item fit from BILOG for item 41.

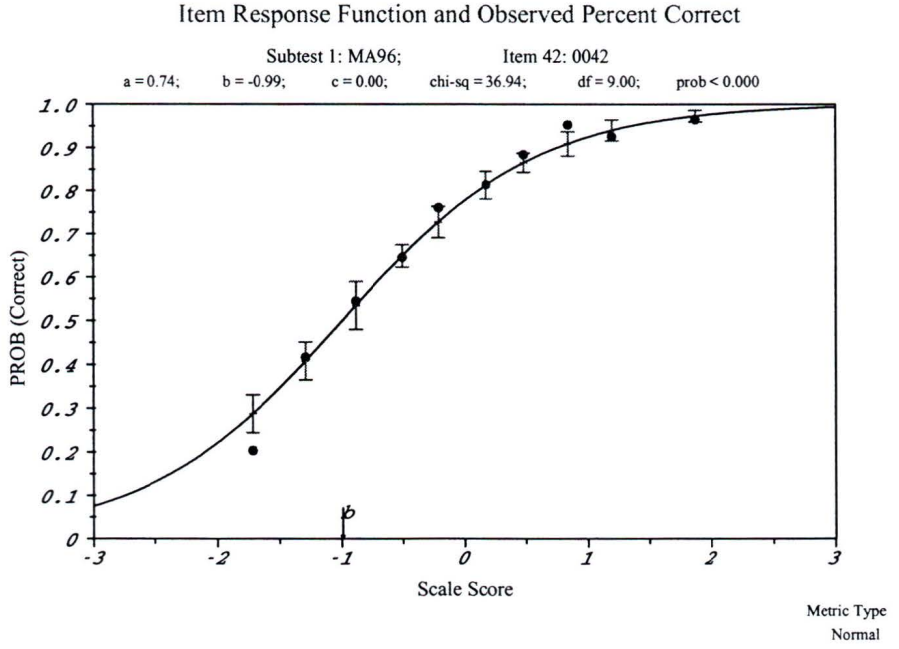


Figure B42. 2PL model ICC showing item fit from BILOG for item 42.

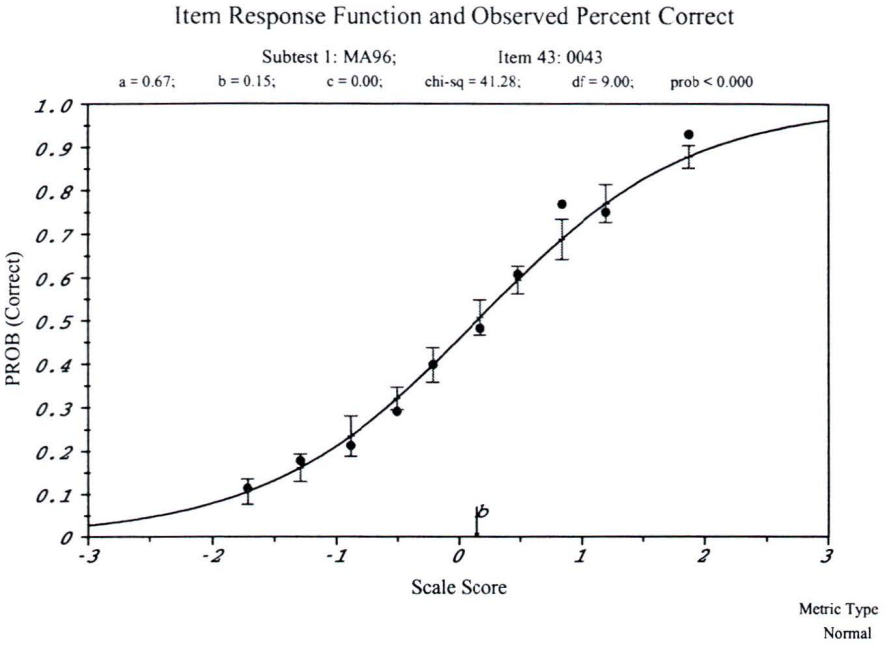


Figure B43. 2PL model ICC showing item fit from BILOG for item 43.

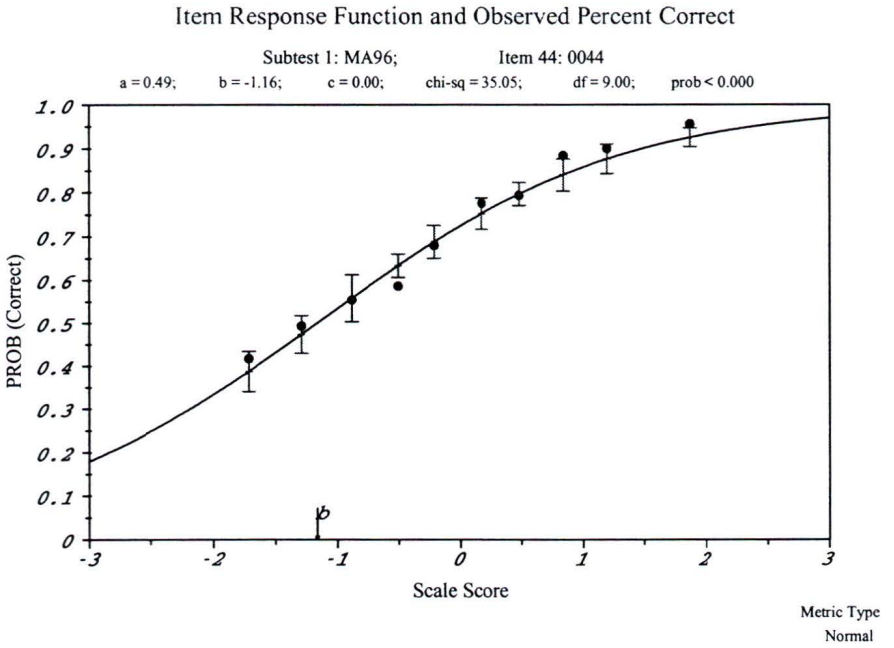


Figure B44. 2PL model ICC showing item fit from BILOG for item 44.

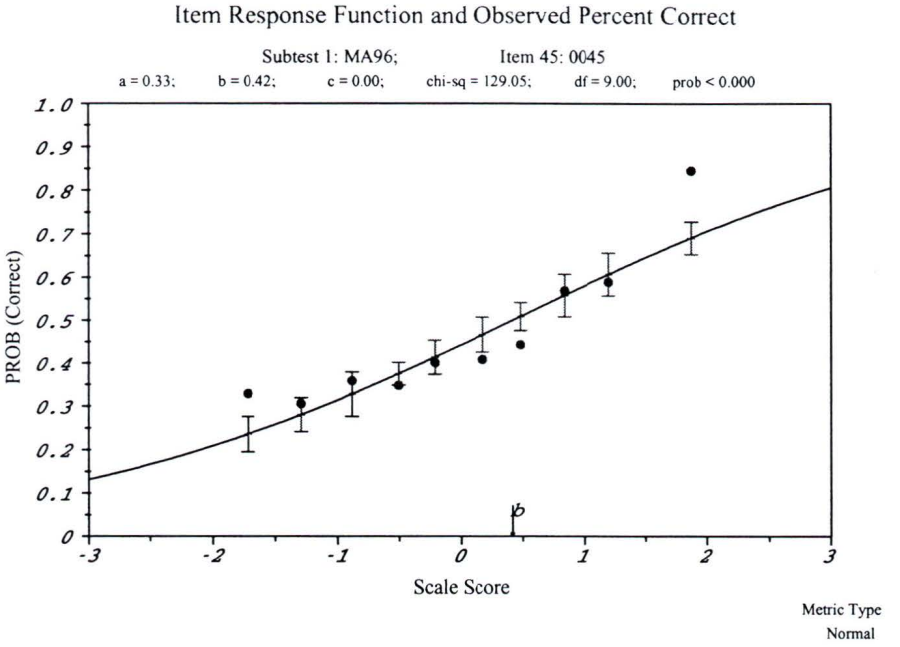


Figure B45. 2PL model ICC showing item fit from BILOG for item 45.

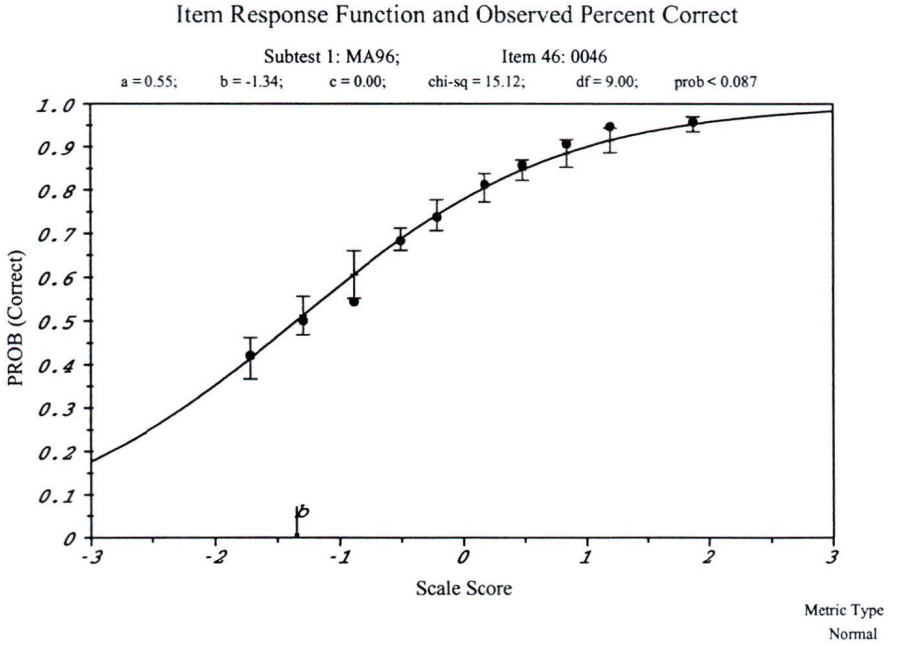


Figure B46. 2PL model ICC showing item fit from BILOG for item 46.

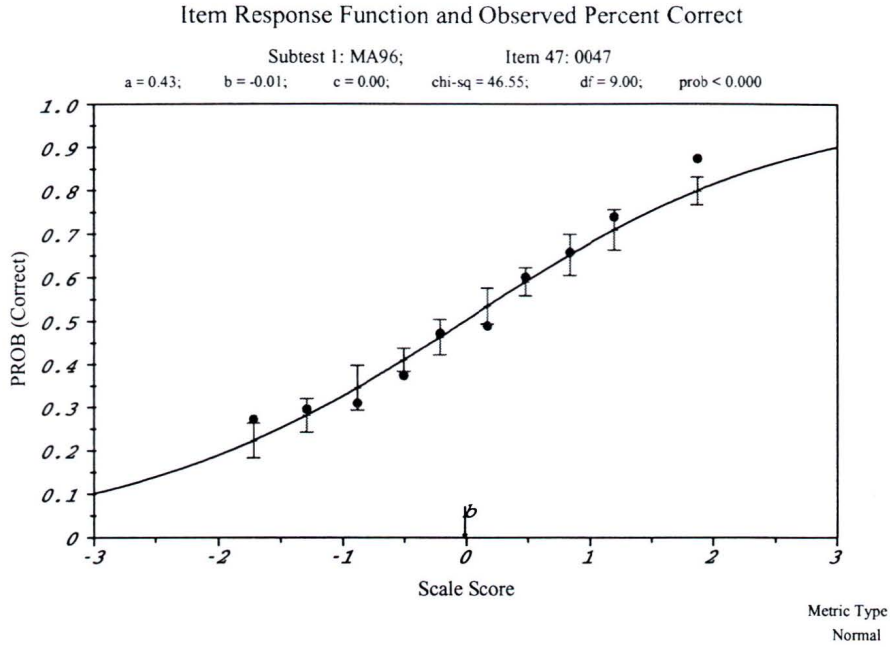


Figure B47. 2PL model ICC showing item fit from BILOG for item 47.

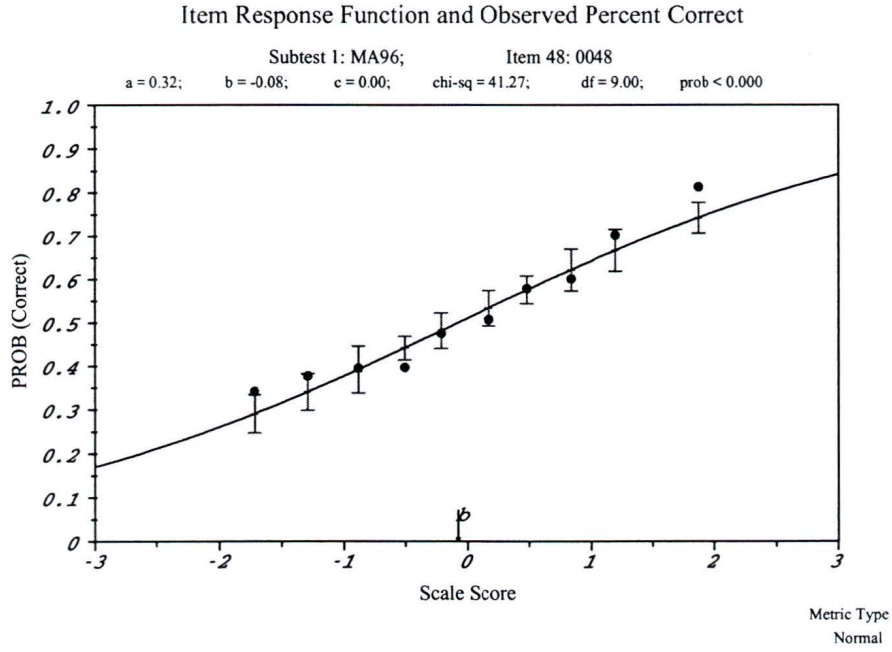


Figure B48. 2PL model ICC showing item fit from BILOG for item 48.

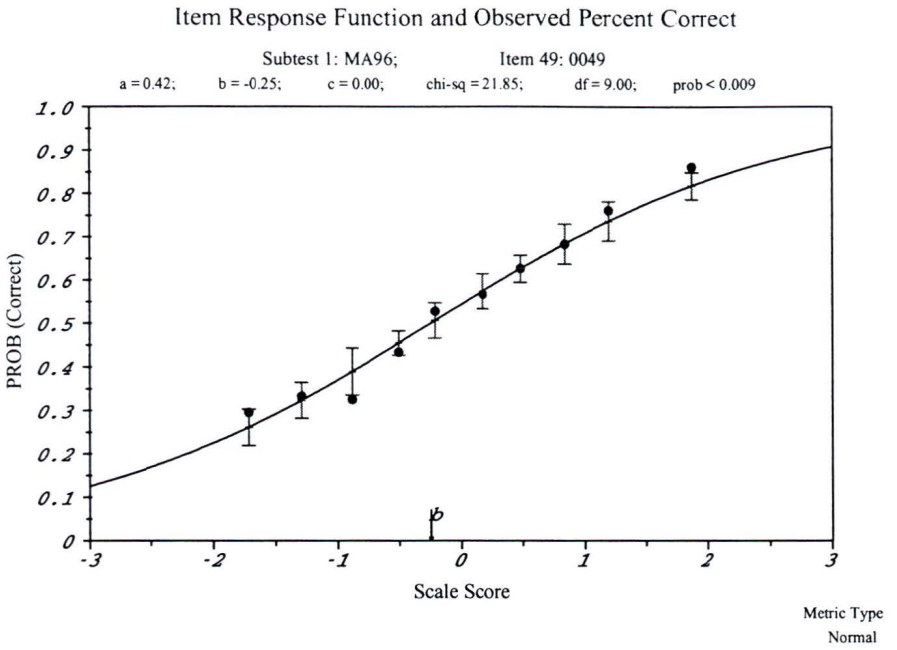


Figure B49. 2PL model ICC showing item fit from BILOG for item 49.

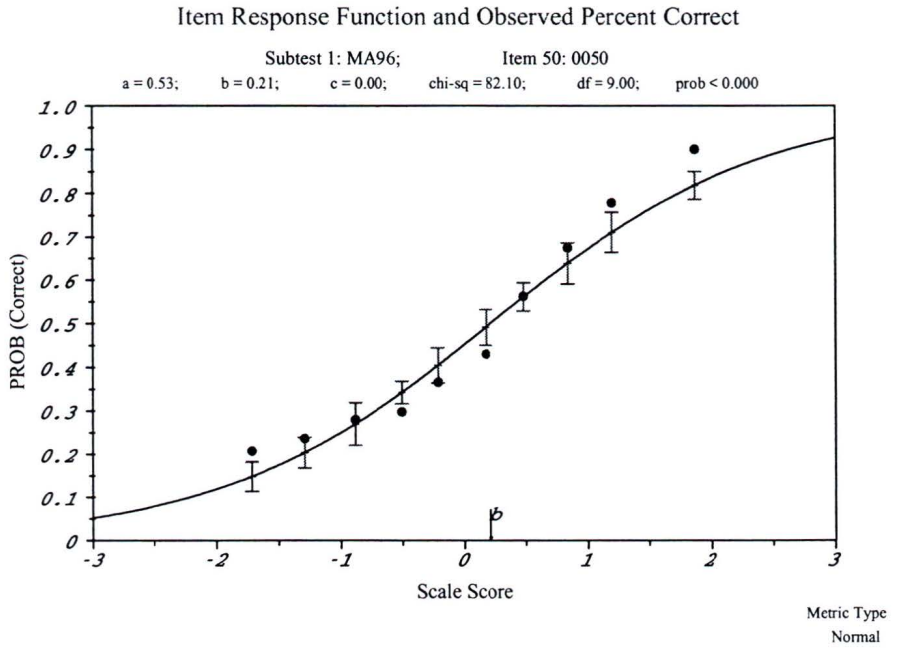


Figure B50. 2PL model ICC showing item fit from BILOG for item 50.

APPENDIX C

Item Characteristics

Table C-1

Difficulty estimates from CTT and IRT.

No	Pval	Bilo1	Bilo2	Bilo3	Par11	Par22	Rum11
1	0.79	-1.75	-2.00	-1.80	-1.33	-1.77	-0.85
2	0.92	-3.11	-2.08	-2.06	-2.47	-1.84	-2.16
3	0.76	-1.50	-1.48	-1.11	-1.12	-1.31	-0.65
4	0.86	-2.35	-2.04	-1.79	-1.83	-1.81	-1.40
5	0.74	-1.38	-1.30	-0.88	-1.01	-1.17	-0.55
6	0.61	-0.58	-0.56	0.38	-0.33	-0.46	0.14
7	0.79	-1.76	-2.23	-1.84	-1.33	-2.08	-0.84
8	0.72	-1.23	-1.21	-0.84	-0.87	-1.05	-0.41
9	0.45	0.31	0.50	1.42	0.42	0.53	0.87
10	0.09	3.12	2.68	1.98	2.80	2.64	3.42
11	0.86	-2.40	-1.58	-1.51	-1.88	-1.36	-1.52
12	0.71	-1.18	-1.23	-1.03	-0.83	-1.07	-0.37
13	0.65	-0.81	-0.93	-0.40	-0.53	-0.78	-0.05
14	0.36	0.81	0.71	0.92	085	0.70	1.36
15	0.49	0.07	0.07	0.61	0.22	0.11	0.70
16	0.57	-0.38	-0.37	0.06	-0.16	-0.29	0.32
17	0.44	0.34	0.36	0.88	0.45	0.39	0.93
18	0.43	0.40	0.33	0.58	0.50	0.35	1.02
19	0.90	-2.81	-1.86	-1.80	-2.22	-1.64	-1.89
20	0.82	-1.98	-1.88	-1.12	-1.52	-1.64	-1.07
21	0.70	-1.08	-0.83	-0.33	-0.76	-0.70	-0.30
22	0.70	-1.15	-0.85	-0.47	-0.81	-0.71	-0.36
23	0.72	-1.23	-1.26	-0.32	-0.87	-1.11	-0.41
24	0.43	0.39	0.57	1.36	0.49	0.58	0.94
25	0.59	-0.47	-0.70	-0.14	-0.24	-0.60	0.24
26	0.82	-1.97	-1.53	-1.13	-1.51	-1.28	-1.10
27	0.57	-0.36	-0.35	0.36	-0.14	-0.27	0.34
28	0.47	0.15	0.15	0.42	0.29	0.19	0.78
29	0.85	-2.27	-1.99	-1.83	-1.76	-1.79	-1.33
30	0.76	-1.52	-1.38	-1.04	-1.13	-1.23	-0.66
31	0.65	-0.84	-1.13	0.46	-0.55	-0.99	-0.08
32	0.65	-0.82	-0.94	-0.64	-0.53	-0.79	-0.05
33	0.70	-1.11	-1.31	-1.00	-0.78	-1.14	-0.29
34	0.70	-1.10	-0.83	-0.77	-0.77	-0.69	-0.30
35	0.76	-1.51	-1.16	-0.88	-1.12	-1.01	-0.68
36	0.63	-0.69	-0.70	-0.31	-0.42	-0.59	0.06
37	0.85	-2.27	-1.67	-1.50	-1.77	-1.44	-1.37
38	0.45	0.27	0.30	0.96	0.39	0.33	0.86
39	0.76	-1.51	-1.97	-1.47	-1.13	-1.83	-0.64
40	0.89	-2.65	-2.27	-2.20	-2.08	-2.03	-1.67

No.	Pval	Bilo1	Bilo2	Bilo3	Par11	Par22	Rum11
41	0.72	-1.23	-0.95	-0.74	-0.87	-0.79	-0.43
42	0.72	-1.27	-0.99	-0.92	-0.91	-0.84	-0.46
43	0.47	0.20	0.15	0.35	0.33	0.19	0.84
44	0.70	-1.12	-1.16	0.00	-0.79	-0.96	-0.32
45	0.45	0.30	0.42	1.25	0.41	0.44	0.86
46	0.75	-1.41	-1.35	-0.72	-1.04	-1.15	-0.58
47	0.50	0.00	-0.01	0.73	0.16	0.04	0.63
48	0.51	-0.04	-0.08	1.04	0.12	-0.02	0.58
49	0.54	0.20	-0.25	0.45	-0.01	-0.17	0.47
50	0.46	0.23	0.21	0.71	0.36	0.25	0.84

Pval - CTT P-val estimates
Bilo1 - 1PL difficulty estimate
Bilo2 - 2PL difficulty estimate
Bilo3 - 3PL difficulty estimate
Par11 - 1PL/1PCM difficulty estimate
Par22 - 2PL/2PCM difficulty estimate
Rim11- 1PL/1PCM difficulty estimate

Table C-2

Discrimination estimates from CTT and IRT.

No.	P-bis	Bilog2	Bilog3	Par22
1	0.37	0.43	0.44	0.47
2	0.57	0.91	0.86	0.87
3	0.43	0.52	0.56	0.57
4	0.46	0.62	0.63	0.67
5	0.45	0.56	0.62	0.58
6	0.46	0.55	1.24	0.60
7	0.32	0.38	0.40	0.40
8	0.43	0.53	0.58	0.57
9	0.25	0.28	0.96	0.29
10	0.44	0.63	1.48	0.61
11	0.61	0.97	0.97	1.08
12	0.42	0.49	0.50	0.53
13	0.38	0.44	0.51	0.47
14	0.48	0.60	1.25	0.66
15	0.41	0.47	0.75	0.50
16	0.46	0.55	0.70	0.59
17	0.40	0.45	0.82	0.47
18	0.52	0.64	0.93	0.71
19	0.60	0.94	0.92	0.96
20	0.43	0.55	0.64	0.61
21	0.56	0.77	1.06	0.86
22	0.59	0.81	1.03	0.90
23	0.41	0.50	0.68	0.53
24	0.29	0.31	0.78	0.34
25	0.30	0.32	0.36	0.32
26	0.53	0.74	0.83	0.86
27	0.46	0.56	0.97	0.60
28	0.42	0.49	0.59	0.54
29	0.46	0.61	0.60	0.65
30	0.47	0.58	0.63	0.62
31	0.32	0.36	0.68	0.39
32	0.39	0.43	0.47	0.46
33	0.36	0.42	0.44	0.45
34	0.58	0.79	0.81	0.87
35	0.56	0.75	0.83	0.83
36	0.43	0.51	0.59	0.55
37	0.55	0.79	0.81	0.90
38	0.38	0.44	1.11	0.46
39	0.32	0.37	0.39	0.38
40	0.46	0.63	0.61	0.66
41	0.55	0.75	0.82	0.84

No.	P-bis	Bilog2	Bilog3	Par22
42	0.55	0.74	0.75	0.80
43	0.54	0.67	0.85	0.74
44	0.41	0.49	0.81	0.55
45	0.29	0.33	1.23	0.35
46	0.45	0.55	0.65	0.61
47	0.38	0.43	0.85	0.48
48	0.30	0.32	0.72	0.35
49	0.37	0.42	0.61	0.46
50	0.45	0.53	1.05	0.57

P-bis - CTT Point-biserial estimates
Bilo2 - 2PL discrimination estimate
Bilo3 - 3PL discrimination estimate
Par22 - 2PL/2PCM discrimination estimate

APPENDIX D

Frequency Distribution

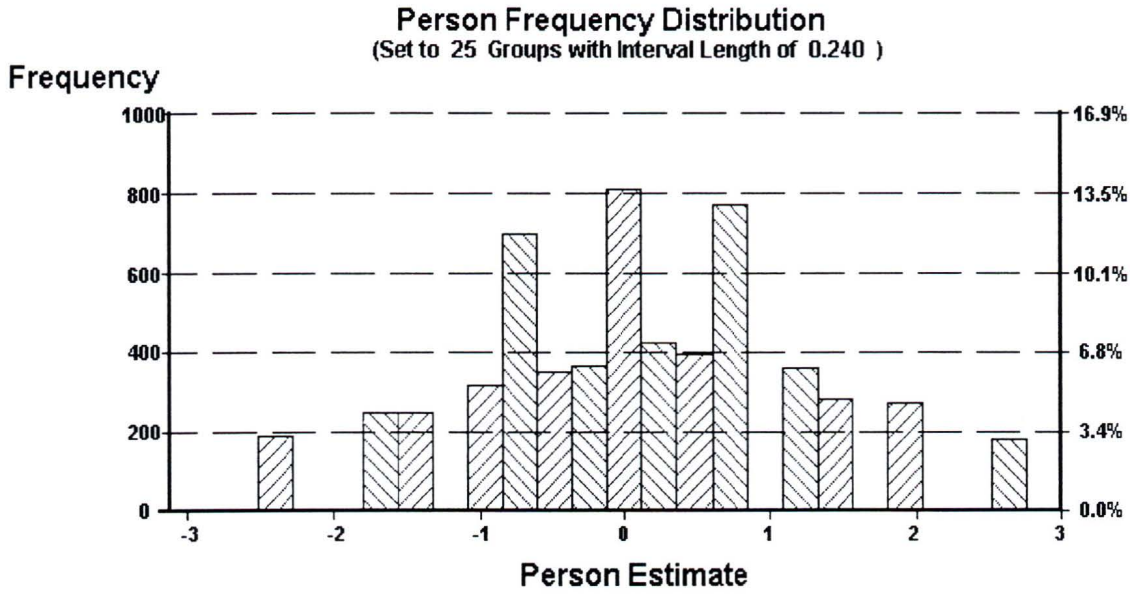


Figure D-1. Frequency distribution of person estimates (ability).

VITA

Surname: Boughton Given Names: Keith Andrew

Place of Birth: Toronto, Ontario, Canada

Educational Institutions Attended:

University of Victoria	1993 to 1998
University College of the Cariboo	1991 to 1993

Degrees Awarded:

B.Sc.	University of Victoria	1996

Honours and Awards:

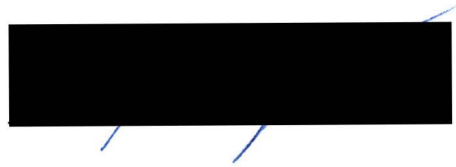
Graduate Teaching Fellowship	1997 to 1998
Graduate Teaching Fellowship	1996 to 1997

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 shall not be allowed without my written permission.

Title of Thesis:

Item Response Theory: The Application of Both Dichotomous and Polytomous Item Response Models to a Provincial Exam Data Set.

Author 

Keith Andrew Boughton
August 26, 1998