THE ESTIMATION OF DOMAIN SCORES THROUGH IRT METHODS ON A MATHEMATICS PLACEMENT TEST

by

HAYLEY KERN

(Under the Direction of Seock-Ho Kim)

ABSTRACT

This study applied item response theory (IRT) estimation methods to a mathematics placement test containing multiple-choice items. Issues that were examined include the following: 1) the selection of a best fitting model for the data from the three most widely used IRT models; 2) the estimation of ability and item parameters; 3) the effect of the number of items on domain score estimation; and 4) comparing IRT estimated domain scores to classical test theory (CTT) domain scores. The two-parameter logistic (2PL) and three-parameter logistic (3PL) models fit the data better than the Rasch model based on the -2 log likelihood values. Three alternate IRT ability estimation methods were considered. The ratios of the root mean squared errors were calculated for the classical test score and the IRT scale score. The results were displayed graphically and illustrate that CTT was more accurate than IRT in estimating an individual’s domain score.

INDEX WORDS: Domain score, Item response theory, Classical test theory, Scale scores
THE ESTIMATION OF DOMAIN SCORES THROUGH IRT METHODS
ON A MATHEMATICS PLACEMENT TEST

by

HAYLEY KERN

B.A., Clemson University, 2004

A Thesis Submitted to the Graduate Faculty of the University of Georgia in Partial
Fulfillment of the Requirements for the Degree

MASTER OF ARTS

ATHENS, GEORGIA

2007
THE ESTIMATION OF DOMAIN SCORES THROUGH IRT METHODS
ON A MATHEMATICS PLACEMENT TEST

by

HAYLEY KERN

Approved:

Major Professor: Seock-Ho Kim
Committee: Allan S. Cohen
Stephen F. Olejnik

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2007
ACKNOWLEDGEMENTS

First, I would like to extend gratitude to Natalia Alexeev for not only providing the data set, but also for supplying background information about the placement test. I would also like to thank my advisor and committee chair, Dr. Seock -Ho Kim, for his guidance and mentoring through the entire thesis process. Furthermore, I wish to thank Drs. Stephen F. Olejnik and Allan S. Cohen for serving on my thesis committee and guiding me through this process.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ACKNOWLEDGEMENTS</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION AND ITEM RESPONSE THEORY</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Domain Scores</td>
<td>3</td>
</tr>
<tr>
<td>Criterion-Referenced Tests</td>
<td>4</td>
</tr>
<tr>
<td>Classical Test Theory</td>
<td>5</td>
</tr>
<tr>
<td>Item Response Theory</td>
<td>7</td>
</tr>
<tr>
<td>Scale Scores</td>
<td>11</td>
</tr>
<tr>
<td>Maximum Likelihood Estimation</td>
<td>11</td>
</tr>
<tr>
<td>Bayes Estimation</td>
<td>12</td>
</tr>
<tr>
<td>Bayes Modal Estimation</td>
<td>14</td>
</tr>
<tr>
<td>Goodness-of-fit Statistic</td>
<td>15</td>
</tr>
<tr>
<td>Placement Tests</td>
<td>16</td>
</tr>
</tbody>
</table>
THE ESTIMATION OF DOMAIN SCORES THROUGH IRT METHODS ON A MATHEMATICS PLACEMENT TEST

PROCEDURE
   Instrumentation
   Sample
   Computer Program

RESULTS
   Model Selection
   Parameter Estimation
   Domain Score Estimation

SUMMARY AND DISCUSSION
   Summary
   Discussion

REFERENCES

APPENDICES
   A
   B
   C
   D
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Classical Item Statistics for the 26 Test Items</td>
<td>28</td>
</tr>
<tr>
<td>2. 2PL and 3PL Item Parameter Estimates of the 17 Multiple - Choice Algebra Items</td>
<td>31</td>
</tr>
<tr>
<td>3. A Summary of the More Accurate Estimation Method for Each Subset of Items Used in Both IRT Models Based on the Ratios of the Root Mean Squared Errors</td>
<td>34</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Classical/IRT-RMS Predictive Error Ratios: Algebra, Bayes Estimation,</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Six Items, 2PL Model ............................................................... ......35</td>
</tr>
<tr>
<td>2.</td>
<td>12 Items, 2PL Model ................................................................. ......36</td>
</tr>
<tr>
<td>3.</td>
<td>Algebra, Maximum Likelihood Estimation, Six Items, 2PL Model ............. ......37</td>
</tr>
<tr>
<td>4.</td>
<td>Algebra, Maximum Likelihood Estimation, 12 Items, 2PL Model ............. ......38</td>
</tr>
<tr>
<td>5.</td>
<td>Algebra, Bayes Modal Estimation, Six Items, 2PL Model ....................... ......39</td>
</tr>
<tr>
<td>6.</td>
<td>Algebra, Bayes Modal Estimation, 12 Items, 2PL Model ....................... ......40</td>
</tr>
<tr>
<td>7.</td>
<td>Algebra, Bayes Estimation, Six Items, 3PL Model ................................ ......41</td>
</tr>
<tr>
<td>8.</td>
<td>Algebra, Bayes Estimation, 12 Items, 3PL Model ................................ ......42</td>
</tr>
</tbody>
</table>
9. Classical/IRT-RMS Predictive Error Ratios: Algebra, Maximum Likelihood Estimation, Six Items, 3PL Model

10. Classical/IRT-RMS Predictive Error Ratios: Algebra, Maximum Likelihood Estimation, 12 Items, 3PL Model

11. Classical/IRT-RMS Predictive Error Ratios: Algebra, Bayes Modal Estimation, Six Items, 3PL Model

12. Classical/IRT-RMS Predictive Error Ratios: Algebra, Bayes Modal Estimation, 12 Items, 3PL Model

13. Scatter Plot of the Relationship Between the IRT Estimated Domain Score Using the Bayes Estimation Method and the True Domain Scores of the 17 Algebra Items; 2PL Model

14. Scatter Plot of the Relationship Between the IRT Estimated Domain Score Using the Maximum Likelihood Estimation Method and the True Domain Scores of the 17 Algebra Items; 2PL Model

15. Scatter Plot of the Relationship Between the IRT Estimated Domain Score Using the Bayes Modal Estimation Method and the True Domain Scores of the 17 Algebra Items; 2PL Model

16. Scatter Plot of the Relationship Between the IRT Estimated Domain Score Using the Bayes Estimation Method and the True Domain Scores of the 17 Algebra Items; 3PL Model
17. Scatter Plot of the Relationship Between the IRT Estimated Domain Score Using the Maximum Likelihood Estimation Method and the True Domain Scores of the 17 Algebra Items; 3PL Model …………………………………………………………… 54

18. Scatter Plot of the Relationship Between the IRT Estimated Domain Score Using the Bayes Modal Estimation Method and the True Domain Scores of the 17 Algebra Items; 3PL Model …………………………………………………………… 55
CHAPTER 1
INTRODUCTION AND ITEM RESPONSE THEORY

Introduction

The present study was designed to compare the methods of classical test theory (CTT) and item response theory (IRT) to estimate domain scores on a random sample of multiple-choice items on a mathematics placement test. Pertinent issues looked at were the selection of a best fitting model from among the three most widely used IRT models and the comparison between CTT and IRT with respect to precision and accuracy with estimating domain scores. The empirical results from the mathematics placement test are presented.

Domain scores have been receiving more attention from the educational community because of their usefulness in score interpretation. These scores can enable school administrators, teachers, parents, and students to see clearly what the student does and does not understand within a given domain. Previous research has focused on the ability and accuracy in estimating these scores (Bock, Thissen, & Zimowski, 1997; Pommerich, 2006; Pommerich, Nicewander, & Hanson, 1999; Schulz, Kolen, & Nicewander, 1999; Schulz & Lee, 2002). Specifically, the issue of obtaining the most precise estimate has been examined when comparing CTT and IRT methods. IRT has typically been shown to be a more accurate estimator of domain scores than percent-
correct scores on the test (Bock et al., 1997; Pommerich, 2006; Schulz et al., 1999; Schulz & Lee, 2002).

In the present study, the mathematics placement test was measured by multiple-choice items. While the majority of previous research has focused primarily on the use of domain scores as indicators of progress or where an examinee currently stands in regard to knowledge of the subject at hand (Bock et al., 1997), the present study examined the issue of placement into a sequence of mathematics courses at a university. Another topic that was covered in this study was the identification of the most accurate way in which to estimate the domain score for an individual. Previous research has shown that IRT provides a more accurate and precise estimate when compared to CTT (Schulz et al., 1999; Pommerich, 2006; Schulz & Lee, 2002).

In this study, item parameters were calibrated using the two-parameter logistic (2PL) model as well as the three-parameter logistic (3PL) model. The model selection was based on the value of the $-2$ log likelihood, with smaller values indicating a better model fit to the data. Overall, the 3PL model would provide the most accurate estimates because it’s $-2$ log likelihood value was the smallest of the three IRT models. The one-parameter logistic (1PL) model was not considered in this study because the 2PL and 3PL models provide more accurate parameter and domain score estimates. The fit value of the 1PL model was larger than the fit value of either the 2PL or 3PL, indicating the 2PL and 3PL models are a better fit for the data. This analysis allows for the comparison of the two models to study the similarities and differences between them regarding their parameter estimates and the domain score estimates. This study investigated the effects of
domain score estimation using a small number of items in a domain in an attempt to see if the IRT estimation method was more accurate than the classical test estimate. The present study used concepts and methods put forth in a previous study by Bock et al. (1997) which used the concepts of random item sampling, the ratio of the root mean squared errors for the test score and the IRT scale score, as well as implementing alternative estimation methods. The following information in this first chapter presents definitions of and information about domain scores. This chapter also explores the two most customary types of estimation methods for domain scores and puts an emphasis on the alternate methods of IRT estimation.

**Domain Scores**

While there are a few variations in definition of a domain score at the individual or group level as well as the type of item (i.e., open-ended or multiple-choice), the basic definition of a domain score is the expected proportion correct that an examinee could obtain on a domain of items. A domain of items can be comprised of any set of items which has been clearly set by an expert. The ideal situation is to have a well-defined domain that is comprised of the skills and abilities that are possible items which may appear on the test, including those items that may not be potential test items (Pommerich et al., 1999). Domain scores aid in student qualification and assessment, however they are not intended to show rank or compare students. These scores aim to show what a student knows on a specific ability or trait, not simply what subject area the student did well in on a particular assessment.
Estimation of domain scores revolves primarily around utilizing the most precise and accurate estimation method. The two most frequently used methods are CTT and IRT. The method of CTT is also known as the true score theory and examines the examinee’s true score, or percent-correct. This means that CTT considers the examinee’s raw test score, which is the sum of the scores for the correct items (Baker, 2001). While CTT makes weak assumptions about the test items, the main objective of this theory is to focus on the test as a whole (Yen & Fitzpatrick, 2006). On the other hand, IRT considers the examinee’s ability level or skill level, when predicting the probability of a correct response on a given test item. IRT seeks to use the observed responses to the test items as raw data estimating to see where an individual stands on the latent characteristic. This type of information can also be used further to gauge where the examinee is in other subject-areas as well as make important decisions regarding performance and placement (Hulin, Drasgow, & Parsons, 1983). Pommerich (2006) stated that IRT estimation offers a flexible method, when estimating domain scores, and that these IRT-based estimates tend to be more accurate than the observed percent-correct score. Bock et al. (1997) also stated that when predicting domain scores, the IRT estimated score typically shows greater accuracy than the test score.

**Criterion-Referenced Tests**

The concept of a domain score grew out of the concept of a criterion-referenced test (CRT). Because a criterion-referenced measure provides information about an individual’s performance, it is important that standards of performance and the purpose of the test are established before the criterion-referenced measure is constructed.
The basic objective of a criterion-referenced measurement is that it evaluates student achievement based on a specific criterion standard reflecting the level at which the student is proficient, but the objective is not to illustrate how the student performed when compared to others (Glaser, 1971b; Popham 1978). Similar to the use of domain scores, a CRT shows what the individual student can do and cannot do, it is not meant to compare students or rank order them (Popham & Husek, 1971).

Nitko (1980) stated that a criterion-referenced assessment is primarily used to capture the individual students’ status with respect to a well-defined domain. These tests can be referred to as domain-referenced tests. The goal of these tests is to gain accurate and specific information about the domain itself in order to interpret the domain score for the individual. While the terms “domain-referenced” and “criterion-referenced” can be seen in previous research being used interchangeably, it is imperative to bear in mind that domain-referenced refers to the association between a well-defined domain and the test items that are representative of the well-defined domain, while criterion-referenced refers to the way an examinee’s score is interpreted, i.e., with respect to a criterion (Martuza, 1977).

**Classical Test Theory**

CTT and IRT models are generally used in similar circumstances and can be analogous. Both theories have the similar main goal of explaining a defined latent variable. CTT basically breaks down the observed score into a true score and an error score. The observed score $X_i$ of an examinee, $i$, can be denoted as $X_i$. The expectation of
the propensity distribution is defined by the examinee’s true score, \( T_i \). This is properly stated as

\[
E(X_i) = t_i .
\]  (1)

The error score for person \( i \), will be represented as \( E_i \) and is defined as the difference between the examinee’s observed score and the true score:

\[
E_i = X_i - t_i .
\]  (2)

Equation 2 derives from the definition that the error score has expectation zero:

\[
E(E_i) = E(X_i) = E(X_i - E(t_i)) = t_i - t_i = 0 .
\]  (3)

CTT is not used as a way to analyze and interpret the test scores for an individual, but rather it provides a way to look at the properties of the test scores in relation to the population of people. Under the assumption that the people sampled from the population are done so randomly, the true score is no longer a fixed value but rather it now becomes a random variable, which produces the CTT equation:

\[
X = T + E ,
\]  (4)

where the variable \( X \) is equal to the examinee’s observed score, the variable \( T \) represents the true score, and \( E \) is the random error component. One of the advantages of CTT is the ability to look at the relationship between the three variables in Equation 4. Since this equation provides information about the quality of the test scores, the reliability of the scores is the main concept within CTT. The reliability of the observed test scores is \( \rho_{xy}^2 \) and from this we can derive the following equation:

\[
\rho_{xy}^2 = \frac{\sigma_T^2}{\sigma_X^2} .
\]  (5)
The definition of Equation 5 is the ratio of the true score variance to the observed score variance. From this equation, it is possible to show how the variance of the observed scores can be equal to the summation of the variance of true scores and the variance of the error scores:

$$\rho_{xT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{(\sigma_T^2) + \sigma_E^2}. \quad (6)$$

When the proportion of test score variance becomes lower, the reliability of the test scores becomes higher. The reverse effect can also occur: When the proportion of test score variance increases and the reliability of the test scores decreases. An important note to remember is that the correlation between the observed scores and the true scores is also the square root of the reliability.

**Item Response Theory**

In comparison to CTT, item response theory (IRT) focuses mainly on the item scores on a test and uses the assumptions underlying the mathematical association between the ability of the examinee and the response given on a test item. The basic idea of IRT comprises the issue of the individual test items rather than the total score of these items. IRT also takes into account the ability level of the examinee and the probability of answering a test item correctly (Hutchinson, 1991). Basically, IRT is a way to examine the relationship between the individual’s ability level and the subsequent response to a test item (Lord, 1980). An examinee’s ability is represented by the Greek letter theta, $\theta$, and at each ability level, there exists a probability of a correct response to that item. This probability is noted as $P(\theta)$. Quite typically, $P(\theta)$ will be larger for high ability examinees and lower for examinees with low ability (Baker, 2001).
An excellent way to graphically achieve understanding of the relationship that Lord (1980) described is to examine one of the central components of this method, the item characteristic curve (ICC). This curve illustrates the relationship between the ability scale and the probability of an examinee’s correct response to a test item. An ICC is present for every individual item on a particular test. The ICC is graphically displayed as an S-shaped curve. Because of its form, the slope can change as a function of the examinee’s ability level. It reaches a maximum value when the ability level is equal to the difficulty of the item.

There are two main components of the ICC: item discrimination and item difficulty. Item discrimination describes how well an item can differentiate between examinees with low ability and those examinees with higher abilities (Baker, 2001). One of the technical aspects to consider is the steepness, or flatness, of the ICC. If the curve is steep, this shows an item that can discriminate well. If the curve appears flat, then this indicates that the item is not able to discriminate very well, because the probability of a correct response at a high level of ability is quite similar to the probability of a correct response at a lower level of ability. Item difficulty is the percentage of correct responses in CTT and on the ICC it indicates the point on the ability scale at which the curve changes from increasing to decreasing.

The basic mathematical model for the ICC is the cumulative form of the logistic function. There are two popular forms of the ICC that are used widely. These are the normal ogive model and the logistic ogive model (Hambleton & Swaminathan, 1985). For the purposes of the present study, only the logistic ogive model will be discussed.
The probability function for the logistic ogive model is:

$$P_i(\theta) = P_i(a_i, b_i, \theta) = \Psi(Z_{ij}) = \frac{e^{Z_{ij}}}{1 + e^{Z_{ij}}} = \frac{1}{1 + e^{-Z_{ij}}}, \quad (7)$$

where $Z_{ij} = a_i(\theta - b_i)$ and $a_i$ specifies the point on the ability scale where $P_i(\theta) = .5$. The discrimination parameter is shown by $a_i$* and it is also the reciprocal of the standard deviation of the logistic function.

The most basic IRT model, the Rasch or one-parameter logistic (1PL) model, only takes into consideration the difficulty of the item. Item discrimination is not considered:

$$P_i(\theta_j) = \frac{e^{\theta_j - b_i}}{1 + e^{\theta_j - b_i}} = \frac{1}{1 + e^{-(\theta_j - b_i)}}, \quad (8)$$

The ability parameter is represented by $\theta_j$ and the difficulty parameter is represented by $b_i$.

Moving to the two-parameter logistic (2PL) model, the item discrimination parameter, $a$, is now considered:

$$P(\theta_j) = \frac{1}{1 + e^{-L}} = \frac{1}{1 + e^{-(\theta_j - b_i)}}, \quad (9)$$

In Equation 9, $e$ is the constant which is equal to 2.718. The difficulty parameter is represented by $b$ and the discrimination parameter is noted by $a$. The logistic deviate is also equal to $L$, which is also equal to $a(\theta - b)$. And again, the ability level is still represented by $\theta$. 
The three-parameter logistic (3PL) model includes the difficulty parameter, the discrimination parameter, and the guessing parameter, \( c \):

\[
P(\theta_j) = c_i + (1 - c_i) \left( \frac{1}{1 + e^{-a_i(\theta - b_i)}} \right). \tag{10}
\]

Again, \( b \) is the difficulty parameter, \( a \) is the discrimination parameter, \( c \) is the guessing parameter, and \( \theta \) is the ability level. The guessing parameter, \( c \), is the probability of getting the item correct purely by guessing. This is also known at the lower asymptote of the ICC. It isn’t a function of the ability level, so a side effect of using the guessing parameter is that it changes the definition of the difficulty parameter.

In the 1PL and 2PL models, the difficulty parameter, \( b_j \), is the point on the ability scale at which the probability of a correct response is .5. The guessing parameter, \( c_j \), now becomes the lower limit of the ICC instead of 0. What that results in is the item difficulty parameter becoming the point on the ability scale

\[
\text{where } P_j(\theta) = c_j + (1 - c_j)(0.5) = \frac{1 + c_j}{2}. \text{ The guessing parameter defines the lower possible value of the probability of the correct response. The item difficulty parameter, } b_j, \text{ becomes the point on the ability scale where the probability of a correct response is halfway between 1.0 and the floor. However, the item discrimination parameter, } a_j, \text{ is still able to be interpreted in the same fashion: proportional to the slope of the ICC at the point where the examinee’s ability, } \theta, \text{ is equal to the item difficulty parameter, } b_j. \text{ With the 3PL model, the ICC’s slope is now } \frac{a_j(1 - c_j)}{4}.\]
Scale Scores

Because IRT estimation methods do not rely on the percent-correct scores to estimate the examinee’s ability as does CTT, it is important to explain some alternative methods and their benefits. According to du Toit (2003), scale scores, which are transformation of the raw scores, have several advantages such as having standard errors that are more accurate, they are similar when test items are deleted or added to the assessment, and the item locations are on the same scale. Unlike CTT corrections for guessing, scale scores have the ability to offer extra robust and flexible modification. In addition, scale scores are able to take discriminating powers into account in order to give optimal weight to each individual test item. There are three most widely used types of scale scores and they are maximum likelihood estimation, Bayes estimation, and the Bayes modal estimation. These methods are discussed in the following sections.

Maximum Likelihood Estimation

Within a data set, the maximum likelihood estimate (MLE) allows inferences to be made regarding the parameters of the probability distribution. MLE assumes independence of the examinees. Regarding the scale score of an examinee, the MLE is represented by θ, ability, which maximizes the following log likelihood:

$$\log L_i(\theta) = \sum_{j=1}^{n} \left\{ x_{ij} \log P_j(\theta) + (1 - x_{ij}) \log [1 - P_j(\theta)] \right\}. \quad (11)$$

In Equation 11, the response function for the item $j$ for examinee $\theta_i$ is symbolized by $P_j(\theta)$. Therefore, the implicit likelihood function is:

$$\left( \frac{\partial \log L_i(\theta)}{\partial \theta} \right) = \sum_{j=1}^{n} \frac{x_{ij} - P_j(\theta)}{P_j(\theta) [1 - P_j(\theta)]} \left( \frac{\partial P_j(\theta)}{\partial \theta} \right) = 0. \quad (12)$$
The MLE for ability is symbolically shown as $\hat{\theta}$. The reciprocal square root of the information at $\hat{\theta}$ is the standard error of the MLE of $\hat{\theta}$:

$$SE(\hat{\theta}) = \sqrt{1/I(\hat{\theta})}.$$  \hspace{1cm} (13)

A key disadvantage of MLE is that it is not successful when an examinee achieves a perfect score or when a score of a perfect zero is obtained.

**Bayes Estimation**

Bayes estimation works best in estimating ability when there is prior information available concerning the examinees (Hambleton & Swaminathan, 1985). The foundation for Bayes estimation revolves around marginal and conditional probability, which pertains to Bayes’s theorem:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$  \hspace{1cm} (14)

According to Hambleton and Swaminathan (1985), the estimation of ability can now reform Equation 14 so that $A$ can now be represented as $\theta_a$ and $B$ is now the collection of the observed responses on $n$ items, $u$. So Equation 14 can now be modified to:

$$P(\theta_a|x) = \frac{P(x|\theta_a)P(\theta_a)}{P(x)}.$$  \hspace{1cm} (15)

Because $\theta_a$ is a continuous variable, Equation 15 can be considered as a density function, meaning that we can rewrite it to be:

$$f(\theta_a|x) = \frac{f(x|\theta_a)f(\theta_a)}{f(x)}.$$  \hspace{1cm} (16)
In Equation 16, for a set of given responses from an examinee, \( f(x) \) is a constant, \( f(x|\theta_a) \) is the posterior density of \( \theta_a \) and the prior of \( \theta_a \) is \( f(\theta_a) \).

Because \( f(x|\theta_a) \) can also be described as a likelihood function of all of the observations, we can again rewrite Equation 16 as:

\[
\begin{align*}
  f(\theta_a|x) \propto L(\theta_a|x) f(\theta_a) .
\end{align*}
\] (17)

The likelihood function is expressed as \( L(\theta_a|x) \), \( f(\theta_a) \) is the prior and the posterior is \( f(\theta_a|x) \).

The Bayes procedure assumes that the prior distribution for \( \theta_a \) is normal with a mean of zero and a unit variance (Hambleton & Swaminathan, 1985). It is logical to presume that the ability parameters of the priors are distributed independently. The distribution of the posterior then resembles:

\[
\begin{align*}
  f(x|\theta_1,\theta_2,\ldots,\theta_N) \propto L(\theta_1,\theta_2,\ldots,\theta_N|x) f(\theta_1,\theta_2,\ldots,\theta_N) .
\end{align*}
\] (18)

Although Equation 18 contains the information about abilities, the point estimates of the parameters for the abilities are still needed. However, given the observed response set of \( x_i \), the Bayes estimate is actually the mean of the posterior distribution of \( \theta \).

Through the use of the Gaussian quadrature, we are able to accurately approximate this:

\[
\begin{align*}
  \bar{\theta}_i \approx \sum_{k=1}^{a} x_k P(x_i|X_k)A(X_k) \overline{\sum_{k=1}^{a} P(x_i|X_k)A(X_k)} .
\end{align*}
\] (19)
Another name for Equation 19 is the expected a posteriori (EAP) estimate which is a function of the response pattern for \( x_i \). A way to estimate its accuracy is with the posterior standard deviation (PSD):

\[
PSD(\hat{\theta}) \approx \frac{\sum_{k=1}^{q} (X_k - \bar{x}_i)^2 P(x_i|X_k)A(X_k)}{\sum_{k=1}^{q} P(x_i|X_k)A(X_k)}.
\]

An advantage of the EAP estimator is that, when compared to other estimators, it has a small average error in the population (du Toit, 2003). While it is biased toward the mean of the population, it is a small bias of only within ±3σ of the mean when the PSD is also small. While the sample mean of the EAP is an unbiased estimator of the latent population mean, the sample standard deviation is usually smaller than the standard deviation of the latent population. This would only be a problem if the examinees were to take different or alternate forms of an assessment.

**Bayes Modal Estimation**

Although this is comparable to the previously mentioned Bayes estimation, it differs when it comes to the size of the average error, as the Bayes Modal estimate contains a larger value of the average error (du Toit, 2003). The Bayes Modal estimate is also referred to as the maximum a posteriori (MAP). The value of the ability, \( \theta \), maximizes the following:

\[
P(\theta|X_i) = \sum_{j=1}^{n} \left[ x_{ij} \log_e P_i(\theta) + (1 - x_{ij}) \log_e [1 - P_i(\theta)] \right] + \log_e g(\theta).
\]
In Equation 21, \( g(\theta) \) expresses the function of the density pertaining to a continuous population distribution of ability, \( \theta \). The following is the likelihood equation for Equation 21:

\[
\sum_{j=1}^{n} \frac{x_j - P_j(\theta)}{P_j(\theta)[1 - P_j(\theta)]} \frac{\partial P_j(\theta)}{\partial \theta} + \frac{\partial \log g(\theta)}{\partial \theta} = 0 .
\] (22)

It is possible to estimate the PSD, the posterior standard deviation, of the MAP estimate, \( \theta \):

\[
PSD(\theta) = \sqrt{\mid I(\theta) \mid} .
\] (23)

While the MAP estimator is usually biased toward the population mean, it does exist for all response patterns, similar to the EAP estimator (du Toit, 2003).

**Goodness-of-Fit Statistic**

Typically, Pearson’s \( \chi^2 \) goodness-of-fit statistic can be used to indicate how well the indicated IRT model fits the test or a test item:

\[
\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j},
\] (24)

where the number of categories is represented by \( k \), and \( O_j \) and \( E_j \) are the observed and expected frequencies, respectively.

For a reasonably long test, the given \( \chi^2 \) may not work well (du Toit, 2003).

Because the test in the current study is long, the examinees can be allocated to intervals on the range of ability, \( \theta \). This process can be done based on their estimated ability value of \( \theta \). The EAP estimate is used in this case and the estimates of the \( \theta \)’s are then rescaled in order to ensure that the same distribution variance is equal to that of the latent
distribution. Within each interval, the number of examinees who correctly responded to an item can be simply tallied from their item scores. From this point, a $\chi^2$ statistic can be used as a comparison for the frequencies of both correct and incorrect responses:

$$G_j^2 = 2 \sum_{h=1}^{n_g} \left[ r_{hj} \log_e \frac{r_{hj}}{N_h P_j(\theta_h)} + (N_h - r_{hj}) \log_e \frac{N_h - r_{hj}}{N_h[1 - P_j(\theta_h)]} \right],$$

(25)

where $n_g$ is equal to the number of intervals, $r_{hj}$ represents the observed frequency of the correct response to item $j$ in interval $h$, $N_h$ is the number of examinees allocated to the specific interval, and $P_j(\theta_h)$ is the value of the fitted response function for the item $j$ at $\theta_h$. It should be noted that $\theta_h$ represents the average ability of the examinees that are in interval $h$.

**Placement Tests**

The majority, if not all, of the research concerning the issue of domain score estimation focuses on the estimation of these scores on assessments whose purpose is to provide an indication of how a certain student has performed in a given subject area. These scores are currently being used to help define domains and levels of achievement as well as areas of growth in achievement and an easy way to report district-level and school-level scores on the National Assessment of Educational Progress (NAEP) (Pommerich, 2006). Placement tests, although not too different from assessments like the NAEP, attempt to simply show what an examinee does or does not know in order to place them into a specific course. It is not at all uncommon these days for incoming freshman to be required to take a mathematics placement exam for their new college or university. These tests become a way to advise new students into the best fitting mathematics course
(Schoen, Cebulla, & Winsor, 2001). The majority of these tests is primarily multiple-choice and focuses almost exclusively on algebraic concepts and skills. Although the current study does not focus on this issue, research has yet to investigate whether students from different high school curriculums have similar success rates in the courses which they are placed into based on the results of their score on the university mathematics placement exam (Schoen et al., 2001).
CHAPTER 2

THE ESTIMATION OF DOMAIN SCORES THROUGH IRT METHODS
ON A MATHEMATICS PLACEMENT TEST

The term domain score developed out of the concept of a criterion-referenced measure, which was initially coined in 1963 by Robert Glaser (Berk, 1980c). The term, criterion-referenced measure, as it is known today developed in the early 1960s. There were a few researchers before this time who gave this concept a foundation. Shaycoft (1979) added that in the late 1930's, Flanagan’s research described what would be a criterion-referenced test today. In the 1950s, Flanagan accurately distinguished between and defined the concepts of norms and standards. These are both the basis for norm-referenced measures and for criterion-referenced measures, respectively. Despite the clear distinctions made between the two types of measures, in the 1960s and 1970s some experts in the measurement field assumed that a criterion-referenced test could be solved with norm-referenced methods (Popham, 1978). While criterion-referenced and norm-referenced measures can be used in the same testing situation, it is important to remember that if this is done, some of the usefulness of the results will be reduced.

By the 1970s, the definition of a criterion-referenced test was expanded as were similar terms (Berk, 1980c). The use of several terms in the place of criterion-referenced, such as domain-referenced and mastery test started to appear. While the terms
may have been changing, the definition of a criterion-referenced measure remained as
being not as a measure to compare one student to others or to give rank to students, but as
a measure to provide information about what an individual student had learned.

As stated previously, a criterion-referenced measure was used interchangeably
with the term domain-referenced measure. While these two terms have been said to be
one and the same, some of the research has been shown to offer a distinction between the
two. While these two concepts are similar, the aim of a domain-referenced measure is to
evaluate how well the measure reflects the status of an individual on the domain (Berk,
1980b). Popham (1978) explained that because a criterion-referenced test usually consists
of a well-defined domain, some researchers chose to refer to that type of form as domain-
referred. Although there is minimal dispute over the issue of interchangeable terms for
a criterion-referenced measure, it is important to remember that all of those terms
encompass the concept of criterion-referenced measurement.

One of the two main principals for the use of criterion-referenced tests is to
estimate domain scores (Hambleton, 1980a; Berk, 1980a). The domain score was defined
as the true proportion of items that an examinee can answer correctly from the population
of items. Initially, it was argued that using item statistics to select the items for the
domain would undermine the character of the test and could possibly lessen the accuracy
of the interpretation of the domain score (Berk, 1980d). Shaycoft (1979) wrote that the
overall score on a domain-referenced measure has criterion-referenced meaning. The
overall score on a domain-referenced test shows the proportion of the domain that the
individual does know, or has mastered. This is indicative of what Berk mentions above,
in the sense that the overall score is equivalent to his definition of a domain score. Martuza (1977) explained that a domain-referenced test consists of a representative content domain that is well-defined. He goes on further to say that the percent-correct score which the individual receives on a domain-referenced measure is a true estimate of their status in regards to the domain. In order to have a clearly specified domain, it is important to have a specified objective for the domain that is being studied, regarding the content that is being covered by the domain as well as the population that will be tested (Hambleton, 1980b).

While there may be some confusion between the terms criterion-referenced and domain-referenced, it would be important to bear in mind that criterion-referenced measurement was initially developed and out of this concept grew several others that all relate to what is known as criterion-referenced. Some of the past research has drawn similarities and differences between these two fundamental concepts in educational measurement. The significance is that a domain-referenced measure is an extension of criterion-referenced measurement. They are similar in nature, but criterion-referenced should be looked at as the way in which an individual’s score is interpreted while domain-referenced simply refers to showing the relationship between a given set of items and the domain from which they originate.

As Pommerich et al. (1999) suggest in their research, a domain score shows the percentage of points that are achieved on an instrument. The stated percentage gives the percentage on a given domain. A domain-referenced test grew out of the concept of a criterion-referenced test. Some researchers use the terms interchangeably, while others
have attempted to distinguish the two from each other, showing that a domain-referenced assessment stems from the idea of a criterion-referenced measurement. In other words, a domain-referenced assessment is a subcategory of a criterion-referenced test. It may be more specific, as it attempts to focus on a single domain, or perhaps multiple specific domains, whereas a criterion-referenced measurement may be composed of several domain-referenced assessments in its effort to analyze a student’s ability in a given domain.

The initial purpose of a domain-referenced instrument was to use a test of skill as an explanation of the status of the test-taker regarding the set of well-defined domain items or behaviors (Berk, 1980b). The score on the domain-referenced instrument has been referred to as the domain score, which expresses what the examinee does know on a particular ability (Schulz et al., 1999). Pommerich (2006) explained that domain scores are typically used to show performance within a domain of items that represent the comprehension and abilities necessary to display mastery of the content.

Being able to comprehend the true meaning of students’ scores on an assessment is of great importance for parents, teachers, and school administrators. Domain scores allow the involved parties to gain an understanding as to what the student does and does not know. However a problem can arise when attempting to translate a students’ numerical achievement score to a written statement. It is possible that some of the meaning of the assessment score could be lost or not fully understood as item statistics may be difficult for the typical person to understand (Schulz, Lee, & Mullen, 2005). To facilitate a better understanding of achievement performance, domain scores are currently
used as a communication tool between the school system and its students and parents (Pommerich, 2006). A domain score indicates how a student performs on certain skills within a domain and whether or not a student is mastering the content that is being tested.

While domain scores are easy to understand because they are usually expressed as the expected percent correct for the domain items, as Pommerich et al. (1999) explain, there are a few different ways in which they are expressed. For example, if a group of students is being studied, then the domain score would be expressed as the expected average percent correct for the domain items. Domain scores can also be expressed in relation to the types of questions that appear on the assessment. For example, if only a single student is taking an assessment that is comprised of only multiple-choice items, then a domain score is the percent of items correctly answered in the domain. But if a group of students is strictly taking a multiple-choice item assessment, then the expression of a domain score changes to the average percent correct. There are also certain conditions for an assessment that may be a mixture of multiple-choice items and open-ended items. If the assessment includes open-ended items and it is being taken by only one student, then a domain score changes its definition to the proportion of total possible points that the student can receive. Again, if a group is taking an assessment that includes open-ended items, then the domain score can be expressed as the average proportion of total possible points in the group. Domain scores are preferred over the customary percent-correct scores because they allow for a much simpler interpretation and these scores allow for a quicker process of performance evaluation (Pommerich et al., 1999).
The intention of creating a test that revolves around testing a certain domain in order to look at an examinee’s performance on the given domain developed out of the criterion-referenced test. The basic purpose of a criterion-referenced test is to create a connection between the items on the test and an established, well-defined domain. It is important to know that in order to create a well-defined domain, one must include the skills and abilities that the examinee is to obtain as well as the way in which they are to obtain it (Pommerich et al., 1999). These scores can also be used as norm-referenced scores. When a specific domain is being studied and this domain is similar to part of a curriculum being used by a school system, a domain score obtained from the test could be used to summarize any norm-referenced abilities.

An important concept of domain scores is that they indicate what students know and what they do not know, or what they have learned and what they did not learn. A domain score is not intended to differentiate between the skills of a higher-level student and a lower-level student. In addition to aiding students, parents, and school personnel, domain scores have also had an impact on national assessments. Pommerich (2006) mentioned that they are currently being used in correspondence with the National Assessment of Educational Progress (NAEP) by showing progress in achievement as well as aiding in the reporting of school-level and district-level scores.

One of the most efficient ways to estimate a domain score is through the use of item response theory (IRT). It has been argued that IRT estimation of domain scores provides more accuracy in estimating the scores rather than using the observed percent
correct score for a single student or the observed average percent correct score for a group of students (Pommerich, 2006). Through the use of IRT estimation, one is able to take two factors into consideration. First, is the students’ performance in the domain area on the test and the second is the item characteristics within the domain. When compared to a percentage of items that were correct on a test, domain scores that have been estimated through the practice of IRT have been seen to be a more accurate performance predictor (Schulz et al., 1999). However, Bock et al. (1997) explained that an item sampling concept is missing in IRT, but that it is possible to convert the IRT scale score received on a test to an estimated domain score. They suggested that this can be done through a few certain conditions. One of those being that there must be a bank of test items that is considered to be a probability sample of the domain. Another condition is that there is an item response model that fits the items in the bank well. The parameters of the well-fitted models must have been estimated from large samples in a specific population. Finally, it is important that a previous test should have already been created that was composed of the items from the test bank. These test items should cover all of the aspects of the given domain.

While there can be numerous variables that affect the accuracy of a domain score estimate, there are a few that have been discussed repeatedly in the literature. Pommerich (2006) discussed the issues of missing data, small sample size, and a small number of items in the domain. Missing data is seen as an obstacle when working with real data. The problems with missing data are only furthered when the students taking the assessment do not see this as a problem. Missing data can lead to results that are
inaccurate because a form could be missing, a question was not answered, or a student carelessly marked responses without considering what was being asked in the question. There are numerous ways in which researchers handle the case of missing data, but it depends on the researcher’s preference. Whichever way it is dealt with, missing data is a large limitation of real data studies.

Small sample size as well as having a small number of domain items can render inaccurate estimates. Pommerich (2006) ascertained that although small domain sizes were present in her study, the results seemed to be satisfactory, but in hopes of gaining more accurate domain score estimates, there is a need for more items in the domain that is presented. Although there were somewhat accurate results found in the research study by Pommerich, previous research indicated that an assessment with less than 20 domain items would not be able to produce accurate estimates of the domain score (Pommerich et al., 1999). These researchers also advise against using the method of IRT estimation for extremely small numbers of domain items as it will not be an accurate estimation of the students’ ability because it is only based on such a limited number of items.

When it comes to the issue of group size, if domain scores are estimated when the size of the groups is small, then these group-level estimates may misrepresent what the true estimate may be. It is suggested that a sample size of at least 50 students would be needed to produce accurate results (Pommerich, 2006).
CHAPTER 3

PROCEDURE

Instrumentation

The current study used data from a mathematics placement test from a large southeastern university administered during the year 2004. The test was required of all freshman students who planned to attend the university. It was not required, however, for incoming transfer students or for students who received a high score on their AP Calculus exam during high school. This test was taken during summer orientation week. The test was comprised of multiple-choice items only.

The test was initially designed to investigate the basic mathematical understanding of the students as well as to evaluate how prepared the students were for college-level mathematics courses. The main purpose of the placement test was to place students into specific university mathematics courses. While the test items were not initially in a domain set, domains were eventually identified from the mathematics department at the university. There was a consensus among experts that the specific domains utilized on this placement test were algebra, trigonometry, and geometry. There were three geometry items, five trigonometry items, and 18 algebra items. While trigonometry can be considered as both algebra and geometry, for the purposes of the present study, it was included as a domain by itself.
The test contained 26 multiple-choice questions. The total score for each individual on the placement test ranged from 1 to 26. Each test item contained five answer choices. If an item was skipped or not answered, it received a value of 6 and was marked incorrect. The responses were coded as zero for incorrect and one for correct. Each student was placed into a math class based on their total score on this placement test. If a student scored between 12 and 15, they were placed into a pre-calculus course. A score of 16 or greater placed a student into calculus. Any score below an 11 placed the student into a mathematical modeling class.

Sample

The data for the 2004 university mathematics placement test consisted of 3,694 students from various ability levels and backgrounds. The test was comprised of three geometry items, five trigonometry items, and 18 algebra items. The mean of the 26 test items was 13.76 with a standard deviation of 4.42. Classical item statistics, such as means, standard deviations, item difficulty values, and item discrimination values, can be found in Table 1. It can be seen from Table 1 that one of the items had a negative item discrimination value. This item was in the algebra domain. Item 14 was considered to be a bad test item and was therefore dropped. This improved the reliability of the test from .748 to .759.

The average difficulty of the 18 algebra items was .56 with a standard deviation of .46; while the average difficulty and standard deviation of the 17 algebra items without item number 14 was .58 and .46, respectively. The reliability of all 18 algebra items was .66, but when item 14 was deleted, the reliability of the 17 algebra items was improved.
The domain of algebra was selected as the exhaustive domain because it contained the majority of the test items. From it, subsets of six and 12 items were sampled. This process was replicated 30 times for each subset. All of the random selection of item subsets can be found in Appendices A and B.

Table 1.

*Classical Item Statistics for the 26 Test Items*

<table>
<thead>
<tr>
<th>Item No.</th>
<th>Item Difficulty</th>
<th>Standard Deviation</th>
<th>Item Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.831</td>
<td>.375</td>
<td>.221</td>
</tr>
<tr>
<td>2</td>
<td>.334</td>
<td>.472</td>
<td>.238</td>
</tr>
<tr>
<td>3</td>
<td>.633</td>
<td>.482</td>
<td>.376</td>
</tr>
<tr>
<td>4</td>
<td>.726</td>
<td>.446</td>
<td>.378</td>
</tr>
<tr>
<td>5</td>
<td>.759</td>
<td>.428</td>
<td>.213</td>
</tr>
<tr>
<td>6</td>
<td>.593</td>
<td>.491</td>
<td>.244</td>
</tr>
<tr>
<td>7</td>
<td>.596</td>
<td>.491</td>
<td>.286</td>
</tr>
<tr>
<td>8</td>
<td>.551</td>
<td>.497</td>
<td>.326</td>
</tr>
<tr>
<td>9</td>
<td>.332</td>
<td>.471</td>
<td>.281</td>
</tr>
<tr>
<td>10</td>
<td>.459</td>
<td>.498</td>
<td>.282</td>
</tr>
<tr>
<td>11</td>
<td>.413</td>
<td>.492</td>
<td>.318</td>
</tr>
<tr>
<td>12</td>
<td>.843</td>
<td>.364</td>
<td>.263</td>
</tr>
<tr>
<td>13</td>
<td>.683</td>
<td>.465</td>
<td>.293</td>
</tr>
<tr>
<td>14</td>
<td>.175</td>
<td>.380</td>
<td>-.075</td>
</tr>
<tr>
<td>15</td>
<td>.742</td>
<td>.437</td>
<td>.296</td>
</tr>
<tr>
<td>16</td>
<td>.864</td>
<td>.343</td>
<td>.180</td>
</tr>
<tr>
<td>17</td>
<td>.669</td>
<td>.470</td>
<td>.347</td>
</tr>
<tr>
<td>18</td>
<td>.522</td>
<td>.500</td>
<td>.384</td>
</tr>
<tr>
<td>19</td>
<td>.333</td>
<td>.472</td>
<td>.291</td>
</tr>
<tr>
<td>20</td>
<td>.351</td>
<td>.477</td>
<td>.338</td>
</tr>
<tr>
<td>21</td>
<td>.445</td>
<td>.497</td>
<td>.306</td>
</tr>
<tr>
<td>22</td>
<td>.557</td>
<td>.497</td>
<td>.338</td>
</tr>
<tr>
<td>23</td>
<td>.365</td>
<td>.481</td>
<td>.324</td>
</tr>
<tr>
<td>24</td>
<td>.334</td>
<td>.472</td>
<td>.089</td>
</tr>
<tr>
<td>25</td>
<td>.481</td>
<td>.500</td>
<td>.334</td>
</tr>
<tr>
<td>26</td>
<td>.171</td>
<td>.377</td>
<td>.214</td>
</tr>
</tbody>
</table>
Computer Program

BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2002) is a widely used statistical software program that allows for a wide range of IRT applications to real-world testing problems. The latest edition to this program allows the program to now be used as a Windows program where the syntax can be created using menu options. The interface goes in the same order as the way in which syntax would be entered. This program assumes binary data, whether it is multiple-choice or short answer. With its ability to handle large amounts of data as well as multiple groups, BILOG-MG has become a versatile program in the field of psychometrics.

Another feature of this program is its capability for differential item functioning (DIF) and the detection and correction for parameter trends over time (DRIFT). It also allows for the equating of equivalent and non-equivalent groups. It permits two-stage testing, vertical equating, and the estimation of group means, standard deviations, and latent distributions. BILOG-MG also utilizes the maximum likelihood, Bayes, and Bayes modal estimation methods.

BILOG-MG enables researchers to work with very large data sets and an unlimited number of items and examinee’s (du Toit, 2003). It also has the ability to include subtests with only one run of the program. BILOG-MG can also consider “variant items,” which may be items that were used on tests in order to estimate item statistics, but they would not be included with the examinee scores. Lastly, with this program, it is possible to extend the concept and fundamentals of IRT to multiple groups.
CHAPTER 4

RESULTS

Model Selection

The initial step of the data analyses was to select a model that best fit the data. As previously stated in Chapter 1, the three types of models in IRT are the one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL). In order to establish which model was the best fit for the data, the $-2 \log$ likelihood values were studied for the 17 multiple-choice items. This value shows the fit value, with smaller values demonstrating a better fit. For the 1PL, 2PL, and 3PL models, the $-2 \log$ likelihood values were 73549.7020, 73411.7827, and 73330.1345 respectively. While the 3PL produced a lower likelihood, indicating that it would be the best model to fit the data, both the 2PL and 3PL models were considered for this paper. Both results are presented. This was performed because the 2PL and 3PL models fit the data better than the 1PL model. Analyzing the data with the 2PL and 3PL models also provides an example of how the IRT models can be used and the type of results that can be attained.

Parameter Estimation

The parameters of the 17 algebra items were calibrated with the 2PL and 3PL models. The parameters for the 3PL were estimated with normal priors (0, 0.5) on the log slope distributions. The parameter estimates for both models can be found in Table 2. As can be seen from these results, there was much variation in these parameter estimates.
from both models. The values of the 2PL estimates are considerably different than the 3PL estimates.

Table 2.

<table>
<thead>
<tr>
<th>Item</th>
<th>2PL</th>
<th>3PL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_i$</td>
<td>$b_i$</td>
</tr>
<tr>
<td>1</td>
<td>0.767</td>
<td>-2.319</td>
</tr>
<tr>
<td>2</td>
<td>0.626</td>
<td>1.199</td>
</tr>
<tr>
<td>4</td>
<td>1.312</td>
<td>-0.986</td>
</tr>
<tr>
<td>5</td>
<td>0.664</td>
<td>-1.891</td>
</tr>
<tr>
<td>6</td>
<td>0.650</td>
<td>-0.640</td>
</tr>
<tr>
<td>7</td>
<td>0.762</td>
<td>-0.583</td>
</tr>
<tr>
<td>8</td>
<td>0.854</td>
<td>-0.284</td>
</tr>
<tr>
<td>9</td>
<td>0.751</td>
<td>1.047</td>
</tr>
<tr>
<td>10</td>
<td>0.737</td>
<td>0.248</td>
</tr>
<tr>
<td>11</td>
<td>0.884</td>
<td>0.461</td>
</tr>
<tr>
<td>12</td>
<td>0.933</td>
<td>-2.096</td>
</tr>
<tr>
<td>13</td>
<td>0.840</td>
<td>-1.056</td>
</tr>
<tr>
<td>16</td>
<td>0.658</td>
<td>-3.037</td>
</tr>
<tr>
<td>18</td>
<td>1.118</td>
<td>-0.106</td>
</tr>
<tr>
<td>19</td>
<td>0.754</td>
<td>1.028</td>
</tr>
<tr>
<td>20</td>
<td>0.924</td>
<td>0.785</td>
</tr>
<tr>
<td>21</td>
<td>0.823</td>
<td>0.308</td>
</tr>
</tbody>
</table>

**Domain Score Estimation**

Once the item parameter estimates were obtained for both models, the next step of the analysis was to estimate each individual’s domain score. This was performed through the use of the random subsets of items discussed in Chapter 3. The estimation of the domain scores was performed using the BILOG-MG computer software program (Zimowski et al., 2002). Examples of the commands used for the domain score estimates can be found in Appendices C and D. The scores were estimated using three different
methods of IRT estimation: Bayes, maximum likelihood, and Bayes modal estimation. The three types of estimation methods were analyzed for both the 2PL and the 3PL model. In order to compare the classical test the ory (CTT) estimates with the IRT estimates, the root mean squared errors (RMSEs) were first computed for the classical test estimate as well as the IRT estimate. These values show the difference between an observed value and a predicted value. The RMSE is calculated as

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\text{true score} - \text{estimated score})^2}{n}},
\]

(26)

where the true score is the domain score calculated from the complete set of 17 algebra items and the estimated score is the domain score obtained from using only a subset of items or one of the three IRT estimation methods.

Secondly, the ratios of the RMSE’s were then calculated as

\[
\text{ratio} = \frac{RMSE(\text{test score})}{RMSE(\text{IRT scale score})},
\]

(27)

where the test score specifies the case with the domain score using the subset of items, which was used in Equation 26 as the estimated score, and the IRT scale score indicates the domain score using one of the three IRT estimation methods, which was also used in Equation 26. Considering the term \( RMSE(\text{IRT scale score}) \), the estimated score from the previous procedure should be replaced with the domain score estimate based on one of the three alternate IRT estimation methods. Bear in mind that this process was replicated 30 times for each of the two IRT models and for each of the combinations of the two item subsets.
The process of obtaining the values for Equations 26 and 27 was completed by calculating the true score based on the entire set of 17 algebra items for each examinee (N=3,694). After this was accomplished, the estimated score was calculated based on the six or 12 sampled algebra items for each examinee. By obtaining the square-root average of the term, $RMSE(\text{test score})$, the true score and estimated score was obtained.

The ratios of the RMSE’s are shown graphically in Figures 1-12. The 2PL ratios are shown in Figures 1-6, and the 3PL ratios are shown in Figures 7-12. A summary of the findings can be found in Table 3. As it can be seen from the figures, these ratios have highly skewed distributions, which is typical of these ratios. However, it can also be seen that several of the ratios are not greater than one, showing that IRT did not have greater accuracy than the CTT scores in the majority of the comparison. The Bayes estimation of domain scores for six items for both the 2PL and 3PL models shows that IRT is more accurate than CTT score, as nearly all of the ratios are greater than one. However, for both the 2PL and 3PL models, the Bayes estimation of 12 items shows that the CTT score had great accuracy than the IRT percent-correct score. This was also the case for the maximum likelihood (ML) estimation of both six and 12 items for both models, as well as for the Bayes modal estimation of 12 items for the 2PL and 3PL models. IRT was shown to be a more accurate estimator of domain scores, in addition to the Bayes estimation of six items, for the Bayes modal estimation of six items for both models.
Table 3.

A Summary of the More Accurate Estimation Method for Each Subset of Items Used in Both IRT Models Based on the Ratios of the Root Mean Squared Errors

<table>
<thead>
<tr>
<th>Model</th>
<th>2PL 6 Items</th>
<th>2PL 12 Items</th>
<th>3PL 6 Items</th>
<th>3PL 12 Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes IRT</td>
<td>IRT</td>
<td>CTT</td>
<td>IRT</td>
<td>CTT</td>
</tr>
<tr>
<td>Bayes Modal</td>
<td>IRT</td>
<td>CTT</td>
<td>IRT</td>
<td>CTT</td>
</tr>
<tr>
<td>MLE CTT</td>
<td>CTT</td>
<td>CTT</td>
<td>CTT</td>
<td>CTT</td>
</tr>
</tbody>
</table>
Figure 1. Classical/IRT-RMS predictive error ratios: Algebra, Bayes estimation, six items, 2PL model
$\text{RMS predictive error ratio}$

$\text{Number of trials}$

Figure 2. Classical/IRT-RMS predictive error ratios: Algebra, Bayes estimation, 12 items, 2PL model
Figure 3. Classical/IRT-RMS predictive error ratios: Algebra, Maximum likelihood estimation, six items, 2PL model
Figure 4. Classical/IRT-RMS predictive error ratios: Algebra, Maximum likelihood estimation, 12 items, 2PL model
Figure 5. Classical/IRT-RMS predictive error ratio: Algebra, Bayes modal estimation, six items, 2PL model.
Figure 6. Classical/IRT-RMS predictive error ratios: Algebra, Bayes modal estimation, 12 items, 2PL model
Figure 7. Classical/IRT-RMS predictive error ratios: Algebra, Bayes estimation, six items, 3PL model
Figure 8. Classical/IRT-RMS predictive error ratios: Algebra, Bayes estimation, 12 items, 3PL model
Figure 9. Classical/IRT-RMS predictive error ratios: Algebra, Maximum likelihood estimation, six items, 3PL model
Figure 10. Classical/IRT-RMS predictive error ratios: Algebra, Maximum likelihood estimation, 12 items, 3PL model
Figure 11. Classical/IRT-RMS predictive error ratios: Algebra, Bayes modal estimation, six items, 3PL model
Figure 12. Classical/IRT-RMS predictive error ratios: Algebra, Bayes modal estimation, 12 items, 3PL model
For the final step of the data analysis, simple scatter plots were constructed to observe the relationship between the IRT estimated domain scores and the true domain scores for the 17 algebra items. The scatter plots are presented in Figures 13-18, with the 2PL plots in Figures 13-15 and the 3PL plots in Figures 16-18. For both models, each alternate IRT estimation method was examined and compared to the true domain score result. Looking at the scatter plots, the basic patterns that were expressed in each of the histograms for the RMSE ratios for each estimation method can be seen for both models. The scatter plot for the relationship between the Bayes estimation method using the 2PL model and the true domain score shows a somewhat linear relationship indicating that the IRT estimate was similar to the true domain score. This was also reflected in the ratios found for this comparison. However, the scatter plot showing the relationship between the ML estimated domain score and the true domain score using the 2PL model has a curve in the data points, with several points at each end of the graph that suggest the IRT estimated domain score would not be the more accurate predictor of test scores using the ML estimation method.

The scatter plot showing the relationship between the true domain scores and the IRT domain score estimates using the method of Bayes modal estimation with the 2PL model shows a linear relationship between these two scores, graphically expressing that IRT is the more accurate estimation method using Bayes modal estimation.

Looking at the scatter plots of the relationship between the true domain scores and the IRT estimated domain score utilizing the 3PL model, it is clear that the results are similar to those of the 2PL model mentioned above. The scatter plot showing the
relationship between the true domain score and the IRT estimated domain score utilizing the Bayes estimation method shows a somewhat linear relationship, but it possesses several additional data points that suggest IRT may not be the more accurate estimation method when it utilizes this particular estimation method. This is similar to what was found when the RMSE ratios were computed for the random item subsets of six and 12 items. IRT was shown to be the more accurate estimation method using the Bayes estimation method for six items, but the same finding was not true when there were 12 items. Both of these findings were true for the 2PL and 3PL models.

The scatter plot of the relationship between the true domain score and the IRT estimated domain score using the ML estimation method also confirms the findings for the 2PL model, with regard to the scatter plot and the RMSE ratios. The scatter plot of the 3PL model shows that ML would not provide a more accurate estimation method when IRT is being compared with the true, or classical, score. There are several points on the plot that suggest the IRT estimation method was not accurate. Lastly, the scatter plot showing the relationship between the true domain score and the IRT estimated domain score using the Bayes modal estimation method presents a linear relationship between the two scores, suggesting that this type of estimation method was more precise when compared to the true score. This analysis can also be confirmed with the RMSE ratios for the 2PL and 3PL models. These ratios showed that the IRT estimated domain score using Bayes modal estimation was more accurate than the true test score for six items using both models. The IRT estimated domain score was not as accurate for 12 items, but there
were a couple ratios that were greater than one when the 3PL model was utilized, but not when the 2PL model was used.

Overall, the scatter plots somewhat confirm what was found through the computations of the RMSE ratios. These scatter plots show that the Bayes estimation for six items is more accurate than the true test score, but this is not so for the 12 item situation. Looking at the scatter plot of the ML estimate relationship with the true score, the pattern of the data points validate the findings of the ratios in the sense that the IRT estimated domain score was not the more accurate estimate when it was compared with the true score. This finding was accurate for both models and also for both subsets of items. Lastly, the Bayes modal estimation scatter plots again verify the findings of the RMSE ratios. Similar to the findings with the Bayes estimation method, the Bayes modal estimation method shows that it is the more accurate estimated score for both models in the subset of six items, but not with 12 items, even though the 3PL model utilizing the Bayes modal estimation method for 12 items resulted in a couple of the ratios that were over the value of one, the IRT estimated domain score using this method was not as accurate overall when compared to the classical test estimate.
Figure 13. Scatter plot of the relationship between the IRT estimated domain score using the Bayes estimation method and the true domain scores of the 17 algebra items; 2PL model
Figure 14. Scatter plot of the relationship between the IRT estimated domain score using the Maximum likelihood estimation method and the true domain scores of the 17 algebra items; 2PL model
Figure 15. Scatter plot of the relationship between the IRT estimated domain score using the Bayes modal estimation method and the true domain scores of the 17 algebra items; 2PL model
Figure 16. Scatter plot of the relationship between the IRT estimated domain score using the Bayes estimation method and the true domain scores of the 17 algebra items; 3PL model.
Figure 17. Scatter plot of the relationship between the IRT estimated domain score using the Maximum likelihood estimation method and the true domain scores of the 17 algebra items; 3PL model
Figure 18. Scatter plot of the relationship between the IRT estimated domain score using the Bayes modal estimation method and the true domain scores of the 17 algebra items; 3PL model
CHAPTER 5
SUMMARY AND DISCUSSION

Summary

The purpose of the present study was to investigate the effect of a small number of domain items on the domain score estimation. Item response theory (IRT) had previously been shown to be a more accurate estimator of domain scores than the percent-correct score (Schulz et al., 1999; Pommerich, 2006). The present study explored whether or not item response theory (IRT) was the more accurate estimation method when compared with classical test theory (CTT). Three alternative IRT estimation methods were explored and compared with the CTT estimate.

The study performed by Bock et al. (1997) explored similar ideas that were examined in the current study. They used 1,000 examinees and each examinee responded to 100 vocabulary items. The researchers then randomly sampled item subsets of 15 and 20 from the 100 items and replicated the procedure. The parameter estimates were obtained using a 2PL model. From the random replications, the RMSE ratios were obtained and the researchers compared the IRT estimated score with the CTT score. Bock et al. considered the estimation methods of Bayes and maximum likelihood (ML) and found that all of the ratios were greater than one for each estimation method and for each subset of items, signifying the better precision in IRT estimated scores rather than the
classical estimate. The researchers added that because Bayes estimation has greater
stability of the score distribution at the extremes, it is typically preferred to
ML estimation since ML estimates generally have standard errors that are typically larger
at the extreme score-values (Bock et al., 1997).

In the present study, the ratios of the root mean squared errors (RMSEs) for the
IRT models were not consistently more accurate than the CTT estimate. The results were
similar for both the 2PL and the 3PL models. IRT estimation was more accurate than
CTT when the estimation methods of Bayes and Bayes modal were used, but only for
item subsets of six test items, not 12 test items. The Bayes estimation method for six
domain items produced ratios that ranged from .92 to 1.25 for the 2PL model and .91 to
1.25 for the 3PL model. The range of ratios for the Bayes modal estimation of six items
using the 2PL model was .92 to 1.25, while the range was .91 to 1.30 for the 3PL model.
The Bayes estimation of 12 items produced ratios that ranged from .75 to .99 for the 2PL
model. The range of the ratios for the 3PL model was from .73 to .96.

In this study, CTT was the more precise estimation method when compared to
IRT estimation utilizing the ML estimation for item subsets of six and 12 items, as well
as for Bayes and Bayes modal estimation methods of item subsets of 12 items. For the
ML estimation of six items with the 2PL model, the ratios ranged from .70 to .91, but for
the ML estimation of six items with the 3PL model, the range of the ratios was from .70
to 1.14. This value over one shows that IRT was the more accurate predictor of the
domain score for one trial out of the total of 30. The range of ratios for the ML
estimation of 12 items for the 2PL model was from .67 to .84. The ratios ranged from .61 to .76 for the ML estimation of 12 items using the 3PL model. Looking at the Bayes modal estimation of 12 items using the 2PL model, the ratios ranged from .74 to .98. With the 3PL model, the Bayes modal estimation of 12 items ranged from .71 to 1.05, with two cases over the value of one, which shows that the IRT was more accurate than the classical estimate using Bayes modal estimation for two trials out of the 30 total trials. These findings were also corroborated in the graphical displays of histograms and scatter plots. The histograms provided a graphical display of the RMSE ratios, showing the instances in which IRT was superior in its estimation of the domain scores and vice versa.

The scatter plots displayed the relationship between the IRT estimated domain scores and the true domain scores for the 17 algebra items. The scatter plot information coincided with that of the histograms in regard to the relationship between the IRT estimated score and the true score, despite there being a difference in the number of test items (i.e. the scatter plots show the relationship between the two scores for 17 items while the histograms displayed the ratios that dealt with the item subsets of only six and 12 items).

The classical item statistics for all 26 test items on the math placement test yielded a low item discrimination value (-0.075) for item number 14, which lead to the item being dropped from the study. The reliability of the entire test was .748. The deletion of item 14 improved the reliability of the test to .759. The reliability of the 18 algebra items was .664, but with the removal of item 14, the reliability improved to .682.
The fit values for the 1 PL, 2PL, and 3PL models were 73549.7020, 73411.7827, and 73330.1345, respectively. There was a wide range of difficulty parameters from both models. The values of the difficulty parameter ranged from -3.037 to 1.199 for the 2PL model. For the 3PL model, the values ranged from -2.493 to 1.457.

Discussion

Previous research studies that examined the estimation methods of IRT and compared those findings to that of CTT estimates, found that IRT was consistently the more accurate estimation method. In addition, earlier studies had also examined this relationship with empirical and simulated data. While the current study used empirical data, the findings were not consistent with the results found by the previous study performed by Bock et al. (1997).

Further study is needed to examine the issues of the total number of test items, domain items, and the total number of responses or participants. While Bock et al. (1997) sampled 15 and 20 items from the 100 items and had 1,000 examinee responses to the items, the current study analyzed six and 12 items from the 17 algebra items. One of the advantages of the current study was the number of examinees, which were 3,694 as compared to 1,000 in the Bock et al. (1997) study.

Note again that, the previous study selected subsets of 15 and 20 items from the item bank of 100 items. The current study selected subsets of only six and 12 items from an item bank of 17 items. An earlier research study by Pommerich et al. (1999) utilized domain sizes of 30, 60, and 120 with item subsets of 5, 10, and 20, while another study
by Pommerich (2006) used a total of 40 algebra items and sampled totals between six and eight items. While the purpose of the present study was to examine the effects of a small number of items on the domain score estimation, the item bank may have been too small.

Pommerich et al. (1999) advised that trying to obtain an ability estimate from five or fewer items using IRT methods should be carefully considered. While the minimum number of items in this study is greater than five, it is indeed close to that minimum. Even though the selection of an item subset of six items may have been a little risky, another idea would be to investigate small item subsets like what was examined in the present research study, but a larger item bank (or a test with more items) should be attained. With only 26 total test items, and then using a domain of only 17 items from a test of 26 items, the test may have been too short in nature and therefore not provided a fair representation of the estimation of domain scores with a small number of items. The number of responses was large and beneficial in the study.

It is worthy to note that the results showed the CTT estimate was more accurate than the IRT estimate for item subsets of 12, but not six. This was found in the estimation methods of Bayes and Bayes modal estimation, not ML estimation. This is an interesting finding because it could be assumed that if IRT was more accurate for only six items, that it would maintain that accuracy for a domain subset of 12 items. This, however, was not the case in the present study.

The finding of the superiority of the Bayes and Bayes modal estimation methods to the ML estimation was not surprising. The similar finding was discovered by Bock et al. (1997). The histograms provide an excellent summary of the findings. The scatter
plots provide an additional graphical display to highlight the findings of the ratios which are displayed in the histograms. The scatter plots display the relationship between the IRT score and the CTT score of the 17 algebra items. Looking at the scatter plot of the ML estimated points for the 17 algebra items, it is clear that there are some significant differences between the IRT score and the CTT score. A majority of the points are not close to each other and this plot does not display the linear trend as does the plot for the Bayes and Bayes modal estimation methods. This observation from the scatter plot visual confirms the finding of the histograms of the ratios for ML estimation when compared to CTT. None of the ratios were over the value of one for either model and for either subset of items.

While these graphs and plots do not necessarily show the exact findings, they all provide an idea of what was found in the current study. Additionally, they coincide with each other in the sense that the scatter plots show what was generally found in the study and the histograms can confirm what the scatter plots show. The scatter plots allow for a more general understanding and idea of what the investigation will yield, while the histograms provide a more in depth and specified understanding of the study and its ideas.

Issues that should be examined for future research are the number of total items that will be used in the study. Perhaps a test or instrument that contains at least 50 items should be used, and then subsets of smaller items could be randomly selected from it. Pommerich et al. (1999) advise not to estimate observed percent-correct scores for tests
that are shorter than 20 items. The total number of responses or participants should remain to be large.

In summary, both Bayes and Bayes modal estimated domain scores were more accurate when compared to CTT domain scores for the six item subsets. CTT clearly provided the more accurate domain scores for the 12 item subsets.
REFERENCES


## APPENDICES

A. Random selection of 6 items from the 17 algebra items: 30 trials

<table>
<thead>
<tr>
<th>Trial Number</th>
<th>Randomly Selected Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 5 6 10 11 12</td>
</tr>
<tr>
<td>2</td>
<td>1 4 5 7 11 13</td>
</tr>
<tr>
<td>3</td>
<td>5 8 10 16 20 21</td>
</tr>
<tr>
<td>4</td>
<td>2 6 10 11 18 19</td>
</tr>
<tr>
<td>5</td>
<td>2 5 6 12 16 20</td>
</tr>
<tr>
<td>6</td>
<td>4 6 9 13 20 21</td>
</tr>
<tr>
<td>7</td>
<td>1 4 16 18 19 20</td>
</tr>
<tr>
<td>8</td>
<td>5 6 8 13 16 20</td>
</tr>
<tr>
<td>9</td>
<td>2 5 10 11 18 20</td>
</tr>
<tr>
<td>10</td>
<td>2 4 7 9 19 20</td>
</tr>
<tr>
<td>11</td>
<td>1 4 5 6 9 10</td>
</tr>
<tr>
<td>12</td>
<td>1 2 8 10 12 13</td>
</tr>
<tr>
<td>13</td>
<td>3 5 12 13 14 15</td>
</tr>
<tr>
<td>14</td>
<td>2 9 11 19 20 21</td>
</tr>
<tr>
<td>15</td>
<td>6 8 9 11 16 19</td>
</tr>
<tr>
<td>16</td>
<td>1 4 5 6 10 18</td>
</tr>
<tr>
<td>17</td>
<td>4 6 8 10 12 18</td>
</tr>
<tr>
<td>18</td>
<td>1 8 9 12 18 21</td>
</tr>
<tr>
<td>19</td>
<td>2 5 6 11 12 13</td>
</tr>
<tr>
<td>20</td>
<td>5 8 9 11 13 16</td>
</tr>
<tr>
<td>21</td>
<td>1 10 11 12 16 21</td>
</tr>
<tr>
<td>22</td>
<td>1 6 8 10 16 21</td>
</tr>
<tr>
<td>23</td>
<td>6 7 16 18 19 21</td>
</tr>
<tr>
<td>24</td>
<td>4 5 7 9 11 21</td>
</tr>
<tr>
<td>25</td>
<td>2 6 7 9 19 21</td>
</tr>
<tr>
<td>26</td>
<td>2 4 7 8 10 21</td>
</tr>
<tr>
<td>27</td>
<td>1 8 13 16 19 21</td>
</tr>
<tr>
<td>28</td>
<td>6 11 12 16 20 21</td>
</tr>
<tr>
<td>29</td>
<td>4 7 10 11 19 21</td>
</tr>
<tr>
<td>30</td>
<td>7 8 10 18 19 21</td>
</tr>
</tbody>
</table>
B. Random selection of 12 items from the 17 algebra items: 30 trials

<table>
<thead>
<tr>
<th>Trial Number</th>
<th>Randomly Selected Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 6 7 8 9 10 11 13 16 18 19 20</td>
</tr>
<tr>
<td>2</td>
<td>1 4 5 6 7 8 10 16 18 19 20 21</td>
</tr>
<tr>
<td>3</td>
<td>2 4 6 7 8 9 12 16 18 19 20 21</td>
</tr>
<tr>
<td>4</td>
<td>1 5 7 8 9 10 11 12 13 18 20 21</td>
</tr>
<tr>
<td>5</td>
<td>1 2 4 5 6 10 12 13 16 18 20 21</td>
</tr>
<tr>
<td>6</td>
<td>5 6 7 9 10 11 12 13 16 19 20 21</td>
</tr>
<tr>
<td>7</td>
<td>1 2 4 5 8 9 10 11 13 16 18 20</td>
</tr>
<tr>
<td>8</td>
<td>1 2 5 7 9 10 12 16 18 19 20 21</td>
</tr>
<tr>
<td>9</td>
<td>1 2 4 5 6 8 10 11 12 13 18 20</td>
</tr>
<tr>
<td>10</td>
<td>1 4 6 7 8 9 10 11 12 13 20 21</td>
</tr>
<tr>
<td>11</td>
<td>1 4 5 6 7 9 10 11 12 16 20 21</td>
</tr>
<tr>
<td>12</td>
<td>1 2 5 7 8 9 10 12 13 16 18 20</td>
</tr>
<tr>
<td>13</td>
<td>2 4 5 6 7 8 9 10 11 18 19 20</td>
</tr>
<tr>
<td>14</td>
<td>1 2 4 6 7 8 10 12 13 19 20 21</td>
</tr>
<tr>
<td>15</td>
<td>1 2 5 6 7 8 9 11 13 18 20 21</td>
</tr>
<tr>
<td>16</td>
<td>1 2 4 6 7 9 11 13 18 19 20 21</td>
</tr>
<tr>
<td>17</td>
<td>1 4 5 8 9 11 12 13 18 19 20 21</td>
</tr>
<tr>
<td>18</td>
<td>2 6 7 8 10 11 12 16 18 19 20 21</td>
</tr>
<tr>
<td>19</td>
<td>1 2 4 5 6 7 12 13 16 18 19 20</td>
</tr>
<tr>
<td>20</td>
<td>1 2 4 5 7 10 11 12 13 19 20 21</td>
</tr>
<tr>
<td>21</td>
<td>2 4 6 7 8 9 10 11 12 13 18 20</td>
</tr>
<tr>
<td>22</td>
<td>1 2 4 5 7 8 9 10 12 13 16 18</td>
</tr>
<tr>
<td>23</td>
<td>2 4 5 6 7 9 10 11 12 18 20 21</td>
</tr>
<tr>
<td>24</td>
<td>2 5 6 7 11 12 13 16 18 19 20 21</td>
</tr>
<tr>
<td>25</td>
<td>1 2 4 5 6 9 10 12 13 18 19 20</td>
</tr>
<tr>
<td>26</td>
<td>1 2 4 5 6 7 9 11 12 16 18 20 21</td>
</tr>
<tr>
<td>27</td>
<td>2 5 7 8 11 12 13 16 18 19 20 21</td>
</tr>
<tr>
<td>28</td>
<td>1 2 4 5 7 8 9 11 12 18 20 21</td>
</tr>
<tr>
<td>29</td>
<td>1 2 4 5 6 8 10 11 16 18 19 20</td>
</tr>
<tr>
<td>30</td>
<td>1 4 5 7 8 9 12 16 18 19 20 21</td>
</tr>
</tbody>
</table>
C. BILOG-MG Command for the Calibration of a 12 Item Subset from the 17 Algebra Items Using the 3_Parameter Model

>GLOBAL DFName = 'ALGEBRA.DAT',
    NPArm = 3,
    SAVE;
>SAVE SCOre = '3DS2MODAL12.SCO';
>LENGTH NITems = (12);
>INPUT NTOtal = 17,
    NIDchar = 4,
    KFName = 'ALGEBRAKEY.DAT';
>ITEMS INAme = (ALG01(1)ALG17);
>TEST1 TNAme = 'ALGEBRA',
    INUmber = (1, 3, 4, 5, 6, 7, 9, 13, 14, 15, 16, 17);
(4A1,6X,17A1)
>CALIB SE lect = (0),
    ACCel = 1.0000;
>SCORE METHod = 3,
    DOMain = 12,
    FILe = '3DS2MODAL12.PAR';
D. BILOG-MG Command for the Calibration of a 6 Item Subset from the 17 Algebra Items Using the 2_Parameter Model

>GLOBAL DFName = 'ALGEBRA.DAT',
    NPArm = 2,
    SAVE;
>SAVE SCOrere = 'DS6BAYES6.SCO';
>LENGTH NITems = (6);
>INPUT NTOtal = 17,
    NIDchar = 4,
    KFName = 'ALGEBRAKEY.DAT';
>ITEMS INAmes = (ALG01(1)ALG17);
>TEST1 TNAme = 'ALGEBRA',
    INUumber = (3, 5, 8, 12, 16, 17);
(4A1,6X,17A1)
>CALIB SELect = (0),
    ACCel = 1.0000;
>SCORE DOMain = 6,
    METHOD = 2,
    FILE = 'DS6BAYES6.PAR';