July 2008

Determining the accuracy of item parameter standard error of estimates in BILOG-MG 3

Michael Toland
University of Nebraska at Lincoln, tolandmd@gmail.com

DETERMINING THE ACCURACY OF ITEM PARAMETER STANDARD ERROR OF ESTIMATES IN BILOG-MG 3

by

Michael D. Toland

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirement

For the Degree of Doctor of Philosophy

Major: Psychological Studies in Education

Under the Supervision of Professor Rafael J. De Ayala

Lincoln, Nebraska

July 7, 2008

DETERMINING THE ACCURACY OF ITEM PARAMETER STANDARD ERROR OF ESTIMATES IN BILOG-MG 3

Michael D. Toland, Ph.D.

University of Nebraska, 2008

Advisor: Rafael J. De Ayala

This study was conducted to determine the accuracy of item parameter standard error of estimates (SEEs) produced by BILOG-MG 3 by examining their performance under a variety of conditions. The Factors manipulated in this study were type of underlying difficulty (b) distribution, type of underlying discrimination (a) distribution, type of underlying lower asymptote (c) distribution, test length (I), type of underlying latent trait (θ) distribution, sample size (J), and the number of quadrature points.

Results showed that the accuracy of the estimated SE_b under the 1PL, 2PL, and 3PL models depended on the magnitude of the b parameter being estimated. Under the 1PL model, the accuracy of the estimated SE_b was related to the underlying b and θ distributions as well as I. The 2PL model results showed that the accuracy of the estimated SE_b was related to I, but no other factors in this study had an impact on the accuracy of estimation of SE_b under this model. For the 3PL model, results showed that the accuracy of the estimated SE_b tended to be impacted by I, while certain combinations of J, I, underlying b distribution, and underlying a distribution had consistently uniform accuracy of estimation of SE_b across the range of b parameters studied.

When considering the accuracy of the estimated SE_a, the 2PL and 3PL model results showed that the accuracy depended upon the magnitude of the *a* parameter being estimated, while an increase in *I* increased the accuracy of the estimated SE_a under the 2PL and 3PL models. Moreover, 2PL and 3PL model results showed the accuracy of the estimated SE_a was related to the underlying item *a*, *b*, and θ distributions as well as *J* and *I*, when the entire range of *a* parameters was considered.

The accuracy of the estimated SE_c under the 3PL model was independent of the magnitude of the item *c* parameter being estimated and unaffected by any combination of factors studied. The implications and limitations of these results are discussed.

This Dissertation is dedicated to

my wife

Lea,

my children

Rylan and Annabelle,

and my Dad

David Toland.

Acknowledgments

Finally, thirty-two and a half years in the making, I get to thank those who have supported me upon this endeavor. First, I would like to thank my wife and son, Lea and Rylan, for putting up with me for so many years while I completed my degree. The two of you kept me honest to my dissertation and pushed me to get it done. I love you both with all my heart! I would also like to thank Annabelle, my daughter. I had hoped to finish before you were born; you have apparently won the first marathon, but at least I finished before baby number three!

I also want to thank my advisor, Rafael De Ayala, for his invaluable feedback throughout my dissertation process. I would like to thank my committee members, Charles Ansorge, Roger Bruning, and Ruth Heaton who believed in me. I want to thank my friends and colleagues James Peugh, James Bovaird, and Kevin Kupzyk who listened to me process my snags and accomplishments aloud.

Finally, I want to thank my Dad, sisters, brother, and my late mother for believing in me. Each of you supported me for so long and now you get to celebrate with me as well. Although my initials M.D. Toland don't mean medical doctor, you can now officially call me Dr. Toland!

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter One

Introduction

A common method for estimating a population mean in statistics is to draw a random sample and compute the sample mean. However, a sample mean will not provide a perfect estimate of a population mean. The sample mean will vary from sample to sample with each sample mean underestimating or overestimating the true population mean. Some sample means will fall close to the population mean, while other sample means will fall further away. In reality, the mean of all sample means will equal the population mean. That is, if a researcher repeatedly took samples of the same size and repeated this process an infinite amount of times, the mean of all the sample means would equal the population mean. By taking repeated samples and computing sample means a sampling distribution is produced. To describe the variability of the sampling distribution a standard deviation is computed. The standard deviation of the sampling distribution has a special name known as the standard error. The standard error refers to the variability of all means from sample to sample and provides a way to measure the average distance between a sample mean and a population mean. Thus, the standard error gives researchers an indication of how accurate their sample data represents their intended population (Agresti & Finlay, 1997). In general, the standard error plays a pivotal role in allowing researchers to compute confidence intervals and conduct statistical significance tests.

Similarly, in testing there is variability in test scores, or ability estimates, along with variability for each item or question on a particular test or assessment. In particular,

tests developed using Item Response Theory (IRT) models give an ability estimate for each examinee along with a standard error of ability for each ability estimate. Also, each item on a test is described by one or more item parameters (e.g., difficulty, discrimination, etc.) and each item parameter has its own item parameter standard error of estimate (SEE). For instance, an item can be described by its item difficulty parameter estimate, with the item difficulty parameter estimate having its own item difficulty parameter SEE. In IRT, the SEE of an item parameter is a measure of the precision of an item parameter estimate (Thissen & Wainer, 1982), with a smaller SEE indicating greater precision. For tests developed using IRT methods the process of determining or estimating the parameters of items is known as item calibration. Item calibration provides a reference for interpreting items and test results. Item calibration is accomplished by administering a test of *J* items to *I* examinees. Then, statistical estimation procedures found in IRT are applied to item responses (e.g., 0, 1) to determine item parameter estimates and SEEs (Baker, 2001).

More importantly, SEEs derived for test items are used in many practical applications involving IRT (Drasgow, 1989). One use of IRT item parameter SEEs is in the area of differential item functioning (DIF) (Lord, 1980; Oshima, Raju, & Nanda, 2006; Smith, 1996; Wright & Stone, 1979). Testing for DIF allows researchers to investigate whether performance on any test item differs for certain groups of examinees (e.g., males-females). The main idea behind DIF is that if we match two different groups of examinees on a construct of interest, then the probability of endorsing an item should be the same for both groups of examinees. That is, DIF is present when equally able

examinees, from different groups, do not have the same probabilities of responding to an item (Hambleton, Swaminathan, & Rogers, 1991; Holland & Wainer, 1993; Lord, 1980). For example if we match males and females on statistics ability, then the probability of responding correctly to an item should be the same for males and females. However, if we find males with the same statistics ability as females had a greater probability of responding correctly to an item than females, then the item would be identified as functioning differently across gender. This means the statistics item is not only measuring statistics ability, but also measuring a second unrelated factor known as gender.

Item parameter SEEs are utilized by researchers testing for item parameter drift (IPD) (Veerkamp & Glas, 2000). An item exhibits IPD when the characteristic(s) or parameter(s) describing an item have changed after several administrations of a particular item. In other words, IPD is the differential change in item parameters over subsequent test administrations (Goldstein, 1983; Veerkamp & Glas, 2000; Wells, Subkoviak, & Serlin, 2002). Essentially, exposed items may become easier and less discriminating after multiple administrations. Checking for IPD is especially important in testing because items become exposed to numerous examinees after time. This means items are at risk of being administered to examinees at more than desirable levels (Veerkamp & Glas, 2000). One consequence of IPD is that prior item parameter estimates for drifting items may no longer accurately characterize items, with the end result being ability estimates based on items showing IPD that no longer measure the intended construct (Wells et al., 2002). Interestingly, testing for IPD has much in common with DIF methods in that both make a distinction between groups of examinees. When testing for IPD a distinction is made

between a calibration phase and a computerized adaptive testing (CAT) phase to determine if item parameters have changed between the calibration and CAT phase. CAT is a way of administering a test, usually via a computer, where items are chosen that are maximally informative for each examinee. Among other items with acceptable discriminating power, an item is typically chosen for administration so an examinee has about a 50 percent probability of answering an item correctly. In CAT, a new temporary estimate of examinee ability is estimated after each subsequently administered item, and then another item is administered based on the temporary ability estimate. To summarize, the CAT sequence starts with an item of average difficulty in the population from which the examinee is selected. Then, depending on how the examinee responds to that item, a second easier or more difficult item is administered. This process continues until an examinee's ability estimate is within some predetermined level of measurement error around the ability estimate (du Toit, 2003; Meijer & Nering, 1999; van der Linden & Glas, 2000; Wainer et al., 1990).

Researchers' examining the effect mode of administration (e.g., CAT versus paper administration) has on item parameter estimates use item parameter SEEs (e.g., see Stone & Lunz, 1994). To test for a mode effect the difference between the item parameter estimates from the two modes is divided by the pooled standard error from the two modes, which creates a standardized difference score, which is then compared to some criterion (e.g., |2|). A mode effect is concluded when an item's test statistic exceeds this criterion.

In test development there are various criteria for determining whether or not an item should be retained in a test; one criterion for not retaining an item is when an item's difficulty SEE is equal to or greater than a predetermined value. For example, El-Korashy (1995) considered excluding items, along with other criteria (i.e., item infit statistics, distribution of items along the ability continuum, and item content), that had item difficulty SEEs exceeding one standard deviation of the item difficulty estimates. In other words, items were retained if their item difficulty SEE was less than one. The advantage of this approach, in conjunction with other criteria, is that it reduces the likelihood that a poorly estimated item is retained within a test. El-Korashy (1995) was the only study found to have considered the size of an item's parameter SEE for inclusion in a test.

As described above, some IRT applications depend on item parameter SEEs, and obtaining accurate item parameter SEEs is a critical concern. However, procedures that use these estimates may arrive at erroneous conclusions (e.g., Type-I error, Type-II error), if the item parameter SEEs are inaccurate (Lord, 1980; Wang & Chen, 2005). Consequently, a small number of simulation studies have considered the accuracy of item parameter SEEs. For instance, recent research by Wang and Chen (2005) found the accuracy of item parameter SEEs produced by the WINSTEPS program (Linacre, 2001) for the Rasch model (Rasch, 1960) and the rating scale model to be accurate under varying test lengths and examinee sample sizes. In Wang and Chen (2005), accuracy was defined by the ratio of the average parameter estimate standard error variance (i.e., the average of the item difficulty parameter SEEs) over sampling variance of the item parameter estimates (i.e., the variance of the difficulty parameter estimates). However,

their study and only one other like it (see Drasgow, 1989) have been limited by the IRT estimation program and IRT model(s) considered. More details about these two studies and their results will be discussed in Chapter Two.

Given the array of IRT applications that are utilizing item parameter SEEs and limited research, there is an apparent need to examine the accuracy of standard errors (SEs) produced for item parameter estimates. One reason to examine the accuracy of item parameter SEEs is that not all test developers utilize the same item parameter estimation program. For instance, previous research has not looked at the accuracy of SEs of item parameter estimates produced by the IRT program BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003), which happens to be one of the most popular IRT programs for dichotomously scored items (e.g., correct-incorrect, agree-disagree). Also, examining the accuracy of item parameter SEEs would reduce any uncertainty researchers have about statistics or procedures that are dependent upon item parameter SEEs. The goal of this study was to add to the literature by extending our understanding of the accuracy of item parameter SEEs; specifically, those produced by the IRT program BILOG-MG 3. Potentially, results from this study are useful in providing researchers with the means to make a decision about the accuracy of item parameter SEEs produced by BILOG-MG 3 which may be otherwise unknown.

Chapter Two

A Review of the Literature

This chapter provides a review of the literature on IRT item parameter SEEs. Included in this review is an overview of IRT and three models used for analyzing dichotomously scored items. A detailed discussion of IRT techniques that use item parameter SEEs and previous research on item parameter SEEs are discussed as well. This is followed by a description of the item parameter estimation procedure used in BILOG-MG 3 and an outline of previous research involving BILOG. Then this chapter concludes with a description of the purpose of the present study.

Overview of IRT

IRT is a modern test theory approach or family of probabilistic models that expresses the relationship between item characteristics (e.g., difficulty, discrimination, etc.) and ability characteristics to the probability of endorsing an item or getting an item correct. As the name suggests, IRT models ability or test performance at the item level rather than at the test level. In the realm of IRT there are numerous mathematical models that can be used to estimate person or ability parameters (e.g., depression, anxiety, aptitude) and item parameters (Hambleton et al., 1991; van der Linden & Hambleton, 1997). Specifically, IRT models have been developed for item responses scored either dichotomously (i.e., have two response categories, for example right-wrong, yes-no, true-false, agree-disagree) or polytomously (i.e., several response categories are possible, for example Likert-type items) (Hambleton & Jones, 1993; Harvey & Hammer, 1999).

IRT models have traditionally been used by testing programs for test development, CAT, test equating, item analysis, and the development of item banks. Testing programs that use IRT have an interest in IRT because it does not have the limitations of Classical Test Theory (CTT). Unlike CTT, IRT provides item and test characteristics that are not dependent upon the ability level of examinees responding to items and ability estimates are not item or test dependent. This means item parameter estimates stay the same regardless of the group tested (sample-free item parameters) and examinee parameter estimates stay the same regardless of the characteristics of the test administered (test-free ability parameters). This special characteristic of IRT models is known as the invariance property and is considered the cornerstone of IRT (Embretson & Reise, 2000; Hambleton et al., 1991; Lord, 1980).

In addition to the invariance property a set of assumptions are made when specifying IRT models. The first major assumption relates to appropriate dimensionality. This means the correct number of underlying trait estimates or abilities is being used to explain person estimates or person performance. For the IRT models considered in this study a single ability is assumed to account for person performance. In other words, a single ability is measured by the set of items on a test and is often referred to as the assumption of unidimensionality. To sum up, the unidimensionality assumption means we are measuring a single ability and by measuring a single ability we can order our examinees on a meaningful continuum (Embretson & Reise, 2000; Hambleton et al. 1991; Lord, 1980).

Another assumption related to undimensionality is the assumption of local independence. Local independence means the response to any item is independent to a response made to any other item, while controlling for ability level or person performance. Simply put, the only factor impacting an examinee's responses to a set of test items are the abilities specified in the IRT model. Therefore, the local independence assumption makes it possible to use the multiplication rule to multiply each individual item probability (i.e., the probability of a correct or incorrect response to an item) to determine the probability that a given response pattern would occur, conditional on a specific examinee's ability level (Embretson & Reise, 2000; Hambleton et al., 1991; Lord, 1980).

Besides unidimensionality and local independence an assumption is made about functional form. The functional form assumption states that the data follow the function specified by the IRT model. Stated differently, the functional form assumption means the relationship between ability and the probability of a correct response to a particular item can be explained by the IRT model under consideration (Embretson & Reise, 2000; Hambleton et al., 1991; Lord, 1980).

IRT Models for Dichotomous Responses

Although there are a number of different IRT models, this study focused on IRT models for dichotomous responses. The three most well known IRT models for dichotomous responses are the one-parameter logistic (1PL) (Rasch, 1960) model, the two-parameter logistic (2PL) (Birnbaum, 1968) model, and three-parameter logistic (3PL) (Birnbaum, 1968) model. Note that the 1PL model is sometimes referred to as the

Rasch model (Rasch, 1960) and the 2PL and 3PL models were formally called the Birnbaum models (Lord, 1980). The models are so called because of the number of item parameters each model contains. The 3PL model is the most general model and can be described by the mathematical expression (Lord, 1980)

$$P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + e^{-Da_j(\theta_i - b_j)}}. \quad (1)$$

Here $P_j(\theta_i)$ is the probability that a randomly chosen examinee with ability value θ_i will answer item j correctly. The relationship between a correct item response and ability can be modeled using a logistic (S-shaped) function known as an item response function (IRF). This function specifies that as the level of the ability increases, the probability of a correct answer (or endorsement) on an item will increase. The values a_j, b_j, and c_j are parameters characterizing item j, e is the mathematical constant 2.71828 …, and D is a scaling factor which transforms $P_j(\theta_i)$ onto the metric of the normal ogive when D = 1.702 (Hambleton et al., 1991; Lord, 1980). When D is used the models are said to be in the normal metric with ability values typically ranging from -3 to +3 (Baker, 2001).

The c_j parameter indicates the probability that an examinee lacking in ability (e.g., $\theta = -\infty$) or with very low ability will respond correctly to an item. This parameter is called the pseudo-chance level parameter and corresponds to the lower asymptote of the IRF. Theoretically, this parameter can range from 0 to 1. In practice, c_j can take on values that are different than the value that would result from random guessing on a multiple choice test (du Toit, 2003; Embretson & Reise, 2000; Hambleton et al., 1991).

The parameter b_j is a location parameter and determines the location of the IRF on the ability continuum. The parameter b_j is called the item difficulty parameter and is also referred to as the item threshold. Items with smaller values of b_j are easier; those with larger values of b_j are more difficult (du Toit, 2003). When the ability values of a group of examinees are transformed to have a mean of 0 and standard deviation of 1, the values of b_j varies typically from -3 to +3 (Baker, 2001). When $c_j = 0$, b_j corresponds to the point (of inflection) on the ability continuum where the probability of a correct response is 0.50. However, when $c_j > 0$, b_j corresponds to the point on the ability continuum where the probability of a correct response is halfway between c_j and 1.0 (i.e., $(1 + c_j)/2$) rather than 0.50. It is important to note that in IRT models, item difficulties may be directly compared to ability levels since they are on the same metric (Baker, 2001; du Toit, 2003).

The parameter a_j is the item discriminating power and is called the item discrimination parameter. This parameter is proportional to the slope of the IRF at the point b_j on the ability continuum (du Toit, 2003; Lord, 1980). Items with higher a_j values are useful for differentiating examinees into different ability levels in the vicinity of the item difficulty than items with smaller a_j values. Theoretically, a_j can range from $-\infty$ to $+\infty$, but the usual range for a_j is between 0 and 2 (Baker, 2001; Hambleton et al., 1991).

Constraining $c_j = 0$ for all items results in the 2PL model while constraining both $c_j = 0$ and $a_j = 1$ for all items results in the 1PL model or more specifically the Rasch model. To summarize, the 3PL model allows each item to differ in terms of their difficulty, discrimination, and pseudo-chance level parameters. The 2PL model is the same as the 3PL model except it assumes all items have a pseudo-chance level parameter

set equal to zero. The 1PL model allows for items to differ in terms of their difficulty parameter, but all items on an instrument are assumed to have a common discrimination parameter along with a lower asymptote set to zero. In general, the 3PL model can be considered a more general form of the 2PL and 1PL (Rasch) models where the other two models can be considered models nested within the 3PL model (Hambleton et al., 1991).

Five hypothetical IRFs are shown in Figure 1. Item 1 represents an item with parameters $b = 0$, $a = 2$, and $c = 0$; item 2 represents an item with parameters $b = 1$, $a = 2$, and $c = 0$; item 3 represents an item with parameters $b = 1$, $a = 1$, and $c = 0$; item 4 represents an item with parameters $b = 0.5$, $a = 1$, and $c = 0.2$; item 5 represents an item with parameters $b = 1$, $a = 0.75$, and $c = 0.1$. Items 1 and 2 are two sample IRFs that conform to the 1PL model. Notice how items 1 and 2 only differ by their location on the ability continuum. When comparing item 2 to item 3 one can see that they have the same difficulty parameter ($b = 1$), but the IRFs for these two items cross. This means each item has different discriminating power. Together, the IRFs for items 1 through 3 exhibit items that would conform to the 2PL model. The IRFs for items 4 and 5 exhibit two items that vary in location on the ability continuum, level of discrimination power, and lower asymptotes. Collectively, all five items demonstrate items that conform to the 3PL model.

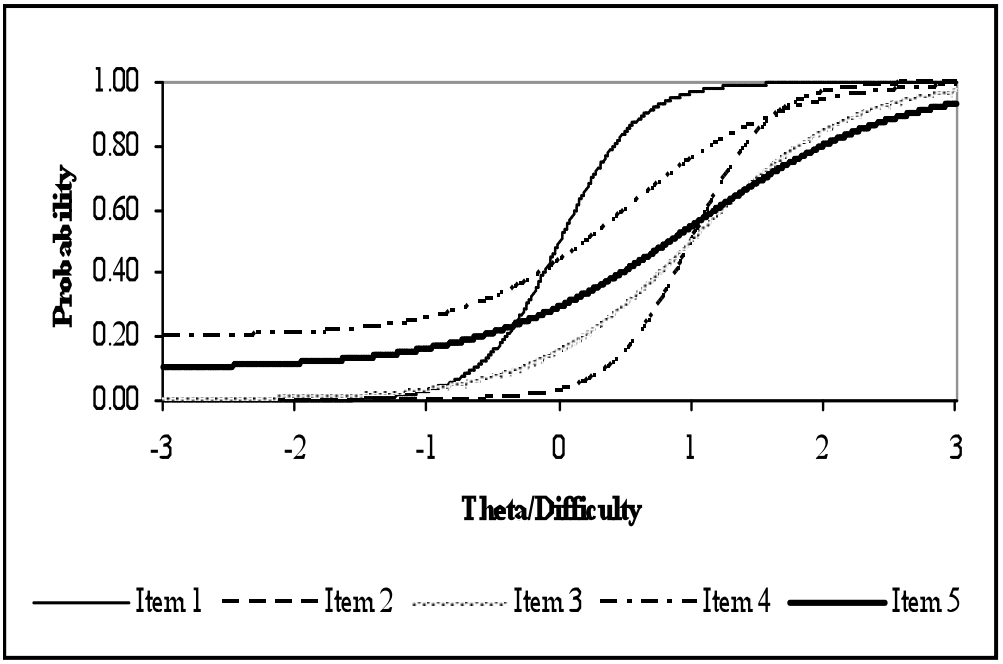

Figure 1. Item response functions for five hypothetical items. The vertical axis represents the probability of a correct response, while the horizontal axis represents the underlying construct continuum.

Applications of IRT Item Parameter SEEs

As highlighted in Chapter One there are various procedures that utilize item parameter SEEs in the area of DIF and IPD for dichotomous IRT models. Some of these procedures include Lord's Chi-square test (Lord, 1980, p. 219-223; see also Hambleton et al., 1991, p. 110-112), the separate calibration *t*-test approach (Wright & Stone, 1979; Smith, 1996), the item parameter replication (IPR) procedure (Oshima et al., 2006), and the cumulative sum (CUSUM) procedure (Veerkamp & Glas, 2000). The first three procedures are used for finding DIF items, while the CUSUM procedure is used for detecting IPD items.

Lord's Chi-square test involves computing separate calibrations for each group. Using the separate item calibrations along with item parameter SEEs the test statistic is constructed and is defined as (Lord, 1980)

$$\chi_j^2 = (\hat{b}_{j1} - \hat{b}_{j2})^2 \Big/ \hat{\sigma}_{j1}^2 + \hat{\sigma}_{j2}^2 . \quad (2)$$

Here $\hat{b}_{j1}$ is the difficulty of item *j* in the calibration based on group 1, $\hat{b}_{j2}$ is the difficulty of item *j* in the calibration based on group 2, $\hat{\sigma}_{j1}$ is the item difficulty SEE for $\hat{b}_{j1}$, and $\hat{\sigma}_{j2}$ is the item difficulty SEE for $\hat{b}_{j2}$. Since only one parameter is being compared, b_j, the degrees of freedom for this test would be 1. Thus, a researcher would compare the test statistic to a Chi-square critical value with 1 degree of freedom to consider whether or not to reject the null hypothesis $b_{j1} = b_{j2}$. Consequently, another test statistic could be computed to test the null hypothesis $a_{j1} = a_{j2}$, however, it is preferable to test both hypotheses simultaneously.

The test statistic is more formally defined as (Lord, 1980)

$$\chi_j^2 = (\hat{a}_{diff}, \hat{b}_{diff})' \hat{\sum}_j^{-1} (\hat{a}_{diff}, \hat{b}_{diff}), \quad (3)$$

where $\hat{a}_{diff} = \hat{a}_{j1} - \hat{a}_{j2}$, $\hat{b}_{diff} = \hat{b}_{j1} - \hat{b}_{j2}$, and $\hat{\sum}_j^{-1}$ is the inverse matrix, sometimes called the reciprocal matrix, of the variance-covariance matrix of the differences between parameter estimates. Since parameter estimates for group one are independent of

parameter estimates for group two, the variance-covariance matrix can be written as (Hambleton et al., 1991) $\hat{\sum}_j = \hat{\sum}_{j1} + \hat{\sum}_{j2}$, where $\hat{\sum}_{j1}$ is the variance-covariance matrix for the parameters in group one, and similarly for $\hat{\sum}_{j2}$. Note that the diagonal elements of the variance-covariance matrix represent item parameter variance estimates and the square-root of each diagonal estimate is the standard error of the item parameter estimate. The test statistic is asymptotically distributed with *k* degrees of freedom and in the case of the 2PL model *k* would equal 2 for the two item parameters being compared (Lord, 1980).

The separate calibration *t*-test approach (Wright & Stone, 1979) computes separate calibrations for the same items based on the groups of interest. Given the pairs of item calibrations and the accompanying item parameter SEEs, a *t*-test is constructed and is defined as (Wright & Stone, 1979)

$$t_j = \hat{b}_{j1} - \hat{b}_{j2} \Big/ \sqrt{\hat{\sigma}_{j1}^2 + \hat{\sigma}_{j2}^2}, \tag{4}$$

where $\hat{b}_{j1}$, $\hat{b}_{j2}$ $\hat{\sigma}_{j1}$, and $\hat{\sigma}_{j2}$ are defined as before. Typically, the *t*-test is compared to a criterion of ± 2 and if it falls above or below this criterion DIF is indicated for an item. Some recent applications or simulation studies involving the separate calibrations *t*-test can be found in Smith (1996), Smith and Suh (2003), and Arnould (2006). Note that this test statistic has also been utilized by researchers examining the effect of mode of

administration (e.g., CAT versus paper administration) on item parameter estimates (see Stone & Lunz, 1994).

The item parameter replication (IPR) method developed by Oshima et al. (2006) uses a Monte Carlo technique involving nine major steps for testing noncompensatory DIF (NCDIF) within the differential functioning of items and tests (DFIT) framework (see Raju, van der Linden, & Fleer, 1995). Note that NCDIF assumes all other items on the test except the item being examined have no DIF, which is the same assumption most other IRT based DIF indices assume (e.g., Lord's Chi-square test). Thus, other DIF tests may be considered comparable in the sense that both provide similar information about DIF (Oshima et al., 2006; Raju et al., 1995). The following steps for the IPR method come from Oshima et al. (2006).

In the IPR procedure the first step is to compute the item parameter estimates from the focal group (e.g., females), which are represented in a column vector called $\hat{M}_j$. In the case of the 3PL model, $\hat{M}_j$ would be a column vector consisting of $(\hat{b}_j, \hat{a}_j, \hat{c}_j)$ for each item. In addition, an item parameter variance-covariance matrix, $\hat{V}_j$, is computed for each item. Using $\hat{V}_j$, the estimated item parameter intercorrelations can be derived and represented in a correlation matrix, $\hat{R}_j$. Assuming $\hat{R}_j$ is positive definite (i.e., all eigenvalues of the $\hat{R}_j$ are positive), $\hat{R}_j$ can be expressed as the product of a triangular matrix, $\hat{T}_j$, and its transpose, $\hat{T}_j'$. In the context of a 3PL model, $\hat{T}_j$ can be expressed as (Oshima et al., 2006)

$$\hat{T}_i = \begin{bmatrix} 1 & \hat{r}_{b_j a_j} & \hat{r}_{b_j c_j} \\ 0 & \sqrt{1-\hat{r}^2_{b_j a_j}} & \dfrac{\hat{r}_{a_j c_j} - \hat{r}_{b_j a_j}\hat{r}_{b_j c_j}}{\sqrt{1-\hat{r}^2_{b_j a_j}}} \\ 0 & 0 & \sqrt{1-\left[\hat{r}^2_{b_j c_j} + \dfrac{(\hat{r}_{a_j c_j} - \hat{r}_{b_j a_j}\hat{r}_{b_j c_j})^2}{(1-\hat{r}^2_{b_j a_j})}\right]} \end{bmatrix}. \quad (5)$$

Second, let $k = 3$ for the 3PL model. Now, let $\hat{X}_{1j}$ and $\hat{X}_{2j}$ each represent a column vector with k elements, with each k element drawn at random from $N \sim (0,1)$. Third, create two new Z column vectors such that $\hat{Z}_{1j} = \hat{T}_j' \hat{X}_{1j}$ and $\hat{Z}_{2j} = \hat{T}_j' \hat{X}_{2j}$. Fourth, transform each Z column vector into a Y vector where $\hat{Y}_{1j} = \hat{D}_j^{1/2} \hat{Z}_{1j} + \hat{M}_j$ and $\hat{Y}_{2j} = \hat{D}_j^{1/2} Z_{2j} + \hat{M}_j$. Here, $\hat{D}_j$ is a diagonal matrix consisting of diagonal elements (variances) from $\hat{V}_j$ and off diagonal elements consisting of zeros. It is important to note that $\hat{D}_j^{1/2}$ is a diagonal matrix consisting of item parameter SEEs in the main diagonal of the matrix. Fifth, column vectors $\hat{Y}_{1j}$ and $\hat{Y}_{2j}$ now represent item parameter estimates from two populations (e.g., females and males) with identical item parameters. In other words, $\hat{Y}_{1j}$ and $\hat{Y}_{2j}$ represent expectations under the null hypothesis or no NCDIF hypothesis. Thus, an NCDIF$_j$ index can be created from $\hat{Y}_{1j}$ and $\hat{Y}_{2j}$, along with estimates of θ for the focal group (e.g., females). As discussed in Raju et al. (1995) the NCDIF

index is defined as $\text{NCDIF}_j = E_F[P_{jF}(\theta_i) - P_{jR}(\theta_i)]^2$, where $P_{jF}(\hat{\theta}_i)$ is the probability of a correct response for examinee *i* at a given θ using item parameter estimates from the focal group, while $P_{jR}(\hat{\theta}_i)$ is the probability of a correct response for examinee *i* at a given θ using item parameter estimates from the reference group. For example, if DIF were being tested between females and males, NCDIF_j would represent the difference in probability scores on item *j* for the same examinee, first treated as a member of the female group, and then treated as member of the male group. The sixth step is to replicate steps 1 through 5 a large number of times (e.g., 10,000). The seventh step is to rank order the replications from the previous step to find the desired percentile ranks (e.g., 95th) and establish the cutoff value for the desired alpha level (e.g., 0.05). The next step is to compare the initial DIF value obtained for item *j* to the cutoff value established in the seventh step. The final step is to repeat this process for all items on a test, hence potentially resulting in a different cutoff criterion for each item (Oshima et al., 2006).

The CUSUM procedure (Veerkamp & Glas, 2000) allows a researcher to conduct a one-tailed hypothesis test to determine whether an item has become easier after each subsequent CAT administration relative to the initial item estimation phase. So, at each CAT administration when the items are re-estimated, the sum of the standardized difference between the difficulty parameters is added to the sum of the previous time periods. The function used in the CUSUM procedure is (Veerkamp & Glas, 2000)

$$S_j(k) = \max\left\{S_j(k-1) + \frac{\hat{b}_j^0 - \hat{b}_j^k}{\sqrt{\sigma^2(\hat{b}_j^0) + \sigma^2(\hat{b}_j^k)}} - d, 0\right\}, \quad (6)$$

where $S_j(k)$ is the cumulative sum for item j at CAT administration k or re-estimation point k, $\hat{b}_j^0$ is the initial estimation of the item difficulty, $\hat{b}_j^k$ is the re-estimate of the item difficulty at time k, $\sigma(\hat{b}_j^0)$ is the difficulty standard error estimate based on the initial estimation, $\sigma(\hat{b}_j^k)$ is the re-estimate of the difficulty standard error at time k, and d is the smallest amount of IPD worth noting or effect size. The CUSUM procedure or chart starts with $S_j(0) = 0$, and the null hypothesis is rejected once $S_j(k) > h$, where h is some constant threshold value. Note that the procedure described above is limited to the 1PL model, but a CUSUM procedure is available for the 3PL model (see Veerkamp & Glas, 2000, DeMars, 2004).

Research Examining Item Parameter SEEs

Research on SEs of IRT item parameters for dichotomous responses can be separated into two categories: (a) papers looking at analytic based SEs or its consistency with empirical SEs derived from a single data set, and (b) simulation studies looking at the accuracy of SEs from item parameter estimates. Research on analytically derived item parameter SEs for the 1PL, 2PL, and 3PL models began with Thissen and Wainer (1982). Then, Li and Lissitz (2004) took their method one step further by examining the consistency between analytic based SEs and empirical SEs. Simulation based research examining the accuracy of item parameter SEs can be traced back to work done by Drasgow (1989) and Wang and Chen (2005).

Analytic standard errors. In Thissen and Wainer's (1982) paper they showed how to compute analytic/asymptotic SEs for any set of item parameters and sample size, with

no data required (i.e., examinees' responses are not necessary), for three commonly used IRT models for dichotomous responses (i.e., 1PL, 2PL, and 3PL models). To use the analytic method three key assumptions are made: (a) the IRT model is appropriate for the data, (b) the examinee's underlying ability distribution is known, and (c) the maximum likelihood estimation method is chosen for item calibration. However, the first two assumptions are unrealistic. Thus, analytic item parameter SEs can be treated as lower limits or a best case scenario for actual item parameter SEs. In addition to the formulas used for deriving item parameter SEs, the paper provides tables and figures that can aid in the determination of the number of examinees needed to yield a desired precision in item parameter estimates. From the tables and figures provided some general conclusions can be drawn about item parameter SEs for the three IRT models when maximum likelihood estimation is used. One, item difficulty SEs become larger as more extreme difficulty parameters (e.g., $b = -3$ or $b = 3$) are estimated under the 1PL, 2PL and 3PL models. Two, the 2PL model is adequate in the range $-2 \leq b \leq 2$, but SEs become larger at the extremes. Three, the 3PL model provides the worst estimate of item parameter SEs relative to the 1PL and 2PL, but difficulty standard errors are adequate only in the middle of the test (e.g., $-1 \leq b \leq 1$). Four, item difficulty SEs for very easy items grow exponentially large under the 3PL model. Five, as sample size goes up, the size of the item difficulty SE goes down in size for each of the IRT models considered in this paper. However, if the *c* parameter cannot be assumed to be homogeneous for all items, the previous statement does not necessarily hold true unless extremely large samples can be used.

To expand upon Thissen and Wainer's (1982) research, Li and Lissitz (2004) examined the consistency between the analytically expected asymptotic standard errors (AEA-SEs) of maximum likelihood and empirically determined standard errors of marginal maximum likelihood estimates (MMLE)/Bayesian item estimates (EMB-SEs), which is a replication based approach, for three IRT models (2PL, 3PL, and generalized partial credit model). Specifically, Li and Lissitiz (2004) treated the item parameters from the Algebra End-of-Course Assessment (Educational Testing Service, 1998) as the true population parameters, which consisted of 24 multiple-choice items, eight short-response dichotomously-scored items, and 10 constructed response items (3 three-category items, 3 four-category items, and 4 five-category items). Using this test as their population (N = 6,426) the authors sampled 1,290 examinees' responses for the 42-item length test and repeated this process for a total of 50 data replications. To calculate the EMB-SEs the following steps were taken: (a) generate a test dataset; (b) simultaneously fit the three models to the item responses and calibrate item parameter estimates using the MMLE/Bayesian estimation method found in PARSCALE (Muraki & Bock, 1996); (c) transform the estimated item parameters to the metric of the true item parameters; (d) repeat the previous steps 50 times; and (e) calculate the *BIAS* and root mean squared error (*RMSE*) for each item parameter estimate. In this study *BIAS* and *RMSE* were defined as

$$BIAS(\xi_j) = \frac{1}{50}\sum_{j=1}^{50}(\hat{\xi}_j^r - \xi_j) \tag{7}$$

and

$$RMSE(\zeta_j) = \sqrt{\frac{1}{50}\sum_{j=1}^{50}(\hat{\zeta}_j^r - \zeta_j)^2}\ , \tag{8}$$

where ζ_j was the true parameter for item *j*, $\hat{\zeta}_j^r$ was the estimated item parameter for item *j*, and *r* represented the data replication number. From these calculations EMB-SE estimate for an item was defined as

$$SE(\hat{\zeta}_j) = \sqrt{RMSE(\zeta_j)^2 - BIAS(\zeta_j)^2}\ . \tag{9}$$

Using the same set of 42 item parameter estimates, the estimated posterior distribution of abilities reported in the PARSCALE output to define the latent distribution of abilities, and a sample size of 1,290, the AEA-SEs were calculated. To test for the precision of SEEs between the two methods dependent samples *t*-tests were performed. In addition, Pearson correlation coefficients were calculated between the two measures along with correlations between *BIAS* and AEA-SE, and *BIAS* and EMB-SE.

Overall, results indicated that the AEA and EMB methods produced very similar SEEs of item parameters for the three IRT models examined, except the correlations of SEEs between these two approaches was slightly lower under the 3PL model. Specifically, correlations between the AEA-SEs and EMB-SEs under the 3PL model were 0.90, 0.89, and 0.91 for the parameters *a*, *b*, and *c*, while correlations between these two approaches under the 2PL model were 0.97 for both *a* and *b* parameters and 0.97,

0.93, 0.94, 0.99, and 0.99 for the *a* parameter and category parameters b_{j2}, b_{j3}, b_{j4}, and b_{j5} under the generalized partial credit model.

Simulation studies. As stated previously, two simulation studies have looked at the accuracy of IRT item parameter SEEs for dichotomous models. Drasgow's (1989) simulation study investigated the accuracy of one approach to estimating item parameters and standard errors of MMLE for the 2PL model. The factors manipulated in this simulation study were test length (5, 10, 15, and 25) and number of examinees (200, 300, 500, and 1,000). The item parameters used in this simulation study consisted mostly of difficulty parameters around -1.5 with discrimination parameters ranging from 0.40 to 1.80. Note that the item difficulty distribution did not match the mean of the θ distribution. Item responses were generated according to the 2PL model. Drasgow (1989) used 10 data replications for each of the four levels of number of examinees and four test lengths to generate independent response vectors. A computer program was written by Drasgow (1989) to estimate item parameters and their corresponding SEEs for the 2PL model. To assess the accuracy of SEEs by the MMLE method, estimated standard errors were compared to observed standard errors. Observed standard error was defined as

$$\sqrt{Var(\hat{b}_j)} = \sqrt{\frac{1}{9}\sum_{r=1}^{10}(\hat{b}_j^r - \bar{b}_j)^2}\,, \tag{10}$$

where $\hat{b}_j^r$ is the difficulty parameter estimate for item *j* in the *r*th replication and $\bar{b}_j$ is the mean difficulty parameter estimate over replications. The same formula used for the

difficulty parameter was also used for the discrimination parameter by substituting *a* for *b* in the formula. Consequently, item parameter estimates from Drasgow's computer program and item parameter estimates from the LOGIST computer program (Wingersky, Barton, & Lord, 1982) were used in the above formulas to compute observed standard errors. Note that the LOGIST computer program was used to provide a frame of reference and that the program was modified so that LOGIST estimates were as close as possible to providing joint maximum likelihood estimates (JMLE). However, JMLEs were only provided for the 15- and 25-item tests. Estimated standard errors were defined as the square roots of the average (over replications) sampling variances obtained from the Fletcher-Powell weight matrix for MMLEs, while estimated standard errors for the JMLEs were computed by taking the square roots of the average sampling variances obtained from formulas given by Lord (1980, p. 191) for JMLE. Estimation accuracy was evaluated at the item level across replications and not averaged across all items.

Overall, results showed that estimated item parameter standard errors obtained from the Fletcher-Powell weight matrix for MMLEs were in close agreement with observed standard errors. Also, the estimated standard errors from the Fletcher-Powell weight matrix for MMLEs were much more accurate than those obtained from the JMLE method. Specifically, Drasgow (1989) concluded that when item parameters are typical of those found on attitude scales or moderately easy tests, as few as 200 examinees and 5 items are needed for reasonably small item parameter standard errors under the 2PL model. Drasgow (1989) also added that larger item parameter SEEs are associated with large item parameters when using MMLE.

In a recent simulation study Wang and Chen (2005) examined the accuracy of item parameter estimates, item parameter SEEs, and item fit statistics produced by the JMLE method in the WINSTEPS program (Linacre, 2001) for the Rasch model and the rating scale model. In this study the researchers manipulated three independent variables: (a) IRT model (the Rasch model and the rating scale model), (b) test length, and (c) number of examinees (100, 200, 400, 600, 800, 1,000, 1,500, and 2,000). Test lengths for the Rasch model were set to 10, 20, 40, and 60 items, while test lengths for the rating scale model were set to 5, 10, and 20 items, with five response categories in each item. Under the Rasch model item difficulties were generated from $N(0,1)$. Item difficulties under the rating scale model were set at -1, -0.5, 0, 0.5 and 1 for the 5-item test and repeated twice for the 10-item test and repeated four times for the 20-item test. Note that the mean ability was set equal to the mean item difficulty for both models. For the rating scale model the researchers focused on 5-point scales only. Therefore, the four intersection or step parameters were set at -2, -0.7, 0.7, and 2 logits. All simulees (i.e., ability estimates) were generated from $N(0,1)$, with 500 replications made under each condition. All simulated data sets were calibrated using WINSTEPS with default options. To assess the accuracy of item parameter SEEs a ratio of the average error variance estimate over the sampling variance was computed for each item. The average error variance was defined as

$$AEV(\hat{\xi}) = \sum_{r=1}^{500} SE(\hat{\xi}_r)^2 / 500, \tag{11}$$

where $SE(\hat{\xi}_r)$ was the standard error of estimate of parameter ξ in the *r*th replication, while the sampling variance was defined as

$$SV(\hat{\xi}) = \sum_{r=1}^{500} (\hat{\xi}_r - \bar{\hat{\xi}})^2 / 499, \quad (12)$$

where $\bar{\hat{\xi}}$ was the mean of the estimates over replications. Two overall conclusions regarding item parameter SEEs were drawn from this simulation study. One, WINSTEPS did not substantially underestimate or overestimate the item difficulty parameter SEEs under the Rasch model for any of the 32 conditions. Two, results under the rating scale model indicated that item parameter SEEs of the overall difficulties and intersection/step parameters were underestimated by about 10 to 40 percent.

Estimation of Item Parameters and Standard Errors in BILOG-MG 3

The estimation of item parameters in BILOG-MG 3 uses an approach efficient for short and long tests called MMLE (Bock & Aitken, 1981; Harwell & Baker, 1991; Harwell, Baker, & Zwarts, 1988; Mislevy, 1986), which was developed by Bock and Aitkin (1981) and extended by Mislevy (1986) to include prior probability distributions for both ability and item parameters. In general, BILOG-MG 3 is a program for multiple group analysis of dichotomously scored data with the 1PL, 2PL, and 3PL models.

The approach used in BILOG-MG 3 for estimating item parameters and standard errors is described in the following sections. In order to understand the estimation process used in BILOG-MG 3 some underlying processes and terminology must be explained.

Prior ability distribution. To estimate item parameters in BILOG-MG 3 an approach is invoked where examinees represent a random sample from an assumed prior population ability distribution $g(\theta|\tau)$, where τ is the vector containing the parameters, μ_θ and σ_θ, of the examinee population ability distribution. In this approach ability is removed from the estimation process and item parameters are estimated in the marginal distribution. In essence, estimation of item parameters is not dependent upon estimation of each examinee's ability estimate, but is dependent on the ability distribution specified a priori. The specification of the prior ability distribution is based on a researcher's knowledge of the distribution of ability for the test and examinees of interest. By invoking this approach an assumption is made that the prior ability distribution is the same for all examinees (Baker & Kim, 2004; du Toit, 2003). The prior ability distribution is important in the item estimation process because an incorrect specification could potentially lead to inaccurate item parameter estimates and standard errors (i.e., the true ability distribution does not match the prior ability distribution). Note that BILOG-MG 3 also provides the option of concurrently estimating the population ability distribution along with the item parameters instead of specifying a *fixed* prior ability distribution (du Toit, 2003). The basic idea behind this latter approach is that once the test has been administered observational data is collected (i.e., examinees responses to each item that are scored 0, 1) on each examinee and based on these data the prior distribution is modified to incorporate observational data about each examinee. The modified distribution is now called the posterior distribution (Harwell et al., 1988).

Gaussian quadrature. Before going on, it is important to point out that the MMLE procedure used in IRT applications for estimating item parameters is usually presented in integral form, however, integration is difficult to evaluate by a computer (Harwell & Baker, 1991). As a result, the MMLE method used in BILOG-MG 3 for estimating item parameters makes use of numerical integration (quadrature), which is better known as Gaussian quadrature, for approximating the integral (Baker & Kim, 2004). In BILOG-MG 3, a simple histogram technique is used to make Gaussian quadrature work. As described above, this is done by making the assumption that examinees are randomly sampled from some continuous ability distribution in the population. Typically, a standard normal prior ability distribution, $g(\theta|\tau)$, is assumed with q equally spaced standard-normal histograms used over the ability range -4 to +4 (Harwell & Baker, 1991). This means the continuous ability distribution can be approximated by using a discrete ability distribution consisting of q histograms over this range and can be more closely approximated by including more histograms. Each histogram will have a midpoint, which is known as a quadrature point (node), X_q (q = 1, 2, …, Q). Each quadrature point will have an associated weight, $A(X_q)$, that reflects the height of the function (i.e., probability of occurrence), $g(\theta|\tau)$, around X_q. The quadrature weight is found by multiplying the width of each rectangular histogram by its height. That is, the probability density at X_q multiplied by $(X_q - X_{q+1})$ gives $A(X_q)$ (Baker & Kim, 2004; Harwell & Baker, 1991; Harwell et al., 1988).

Artificial data. The use of Gaussian quadrature entails item parameters that are not estimated directly from the individual examinee data but rather from artificial data at

each of the q quadrature points. The artificial data at each quadrature point consists of the expected (conditional) number of examinees, $\bar{n}_{jq}$, and the expected (conditional) number of correct responses, $\bar{r}_{jq}$, responding to item j at each quadrature point (X_q) (Baker & Harwell, 2004). Here $\bar{n}_{jq}$ and $\bar{r}_{jq}$ are defined as (Baker & Kim, 2004)

$$\bar{n}_{jq} = \sum_{i}^{I} P(X_q \mid Y_i, \varepsilon, \tau) = \sum_{i}^{I} \left[\frac{L(X_q)A(X_q)}{\sum_{q}^{Q} L(X_q)A(X_q)} \right] \tag{13}$$

and

$$\bar{r}_{jq} = \sum_{i}^{I} y_{ji} P(X_q \mid Y_i, \varepsilon, \tau) = \sum_{i}^{I} \left[\frac{y_{ji} L(X_q)A(X_q)}{\sum_{q}^{Q} L(X_q)A(X_q)} \right], \tag{14}$$

where

$$L(X_q) = \prod_{j}^{J} P_j(X_q)^{y_{ji}} Q_j(X_q)^{1-y_{ji}} \tag{15}$$

which is the quadrature form of the likelihood of Y_i conditional on $\theta_i = X_q$ and the item parameters

$P_j(X_q)$ comes from the IRT model (e.g., 3PL) using X_q instead of θ_i and

$Q_j(X_q) = 1 - P_j(X_q)$

$i = 1, \ldots, I$ (where I equals the number of examinees)

$j = 1, \ldots, J$ (where J equals the number of items)

y_{ji} is the response (i.e., 0, 1) to item j by examinee i

$q = 1, 2, \ldots, Q$ (recall Q equals the number of quadrature points)

Y_i is a vector of item responses of the ith examinee to the J items

ε is a vector of item parameters

τ is the vector containing the parameters of the examinee population ability distribution.

Concretely, Equation 13 is the expectation (probability) of each examinee having an ability X_q for all values of X_q. Then the $\overline{n}_{jq}$ are found by summing these probabilities separately for each X_q. In sum, a separate expected number of correct responses and number of examinees responding to item j is computed at each quadrature point. These artificial data are then used in BILOG-MG 3 to estimate item parameters (Baker & Kim, 2004; Harwell & Baker, 1991).

The MMLE estimation equations in BILOG-MG 3. The MMLE estimation equations, written in Gaussian quadrature form, for item parameters used in BILOG-MG 3 for the 3PL model are (Baker & Kim, 2004)

$$L_1(\alpha_j) = e^{\alpha_j}(1 - c_j)\sum_{q}^{Q}[\overline{r}_{jq} - \overline{n}_{jq}P_j(X_q)]w_{jq}(X_q - b_j) \qquad (16)$$

$$L_2(b_j) = -e^{\alpha_j}(1-c_j)\sum_{q}^{Q}[\bar{r}_{jq} - \bar{n}_{jq}P_j(X_q)]w_{jq} \tag{17}$$

$$L_3(c_j) = (1-c_j)^{-1}\sum_{q}^{Q}\frac{[\bar{r}_{jq} - \bar{n}_{jq}P_j(X_q)]}{P_j(X_q)} \tag{18}$$

where

α_j = initial value for item j discrimination parameter

b_j = initial value for item j difficulty parameter

c_j = initial value for item j pseudo-chance level parameter

$w_{jq} = P_j^*(X_q)Q_j^*(X_q)/P_j(X_q)Q_j(X_q)$

defined also (Baker & Kim, 2004)

$$P_j^*(X_q) = \frac{e^{a_j(X_q - b_j)}}{1+e^{a_j(X_q - bj)}},\ Q_j^*(X_q) = 1 - P_j^*(X_q). \tag{19}$$

To solve Equations 16 through 18 they are each set equal to 0 and the item parameter estimates for a single item are estimated simultaneously by the Fisher scoring-for-parameters method within the context of an EM algorithm (Baker & Kim, 2004; du Toit, 2003; Mislevy, 1986). However, Equations 16 through 18 do not contain the Bayesian components pertaining to the prior distributions imposed on the item parameters as implemented in BILOG-MG 3 (Baker & Kim, 2004). Before elaborating

on the full item parameter estimation equations used in BILOG-MG 3, the prior component used in BILOG-MG 3 and their function in the estimation process will be discussed in the following sections.

Priors used in estimating item parameters in BILOG-MG 3. In BILOG-MG 3 a prior component is imposed on each item parameter during the estimation of item parameters. The term prior comes from Bayesian statistics, often referred to as the prior probability distribution, and provides information about a variable in the absence of data. Essentially, Bayesian statistics is based on the idea that each parameter of interest has its own distribution, whereas most typically view parameters as fixed characteristics of the population. The function of the prior distribution in Bayesian statistical inference is for a researcher to specify their assumption about the distribution of the parameter(s) of interest (Baker & Kim, 2004).

In the IRT literature, many authors have advocated that priors be used in estimating item parameters so reasonable or identifiable parameter estimates may be found (Harwell & Baker, 1991; Mislevy, 1986; Swamminathan & Gifford, 1985). As a result, prior distributions and their hyper parameters (e.g., μ and σ of the distribution) are utilized in BILOG-MG 3 in estimating item parameters along with their respective standard errors (Baker & Kim, 2004). By imposing prior distributions on the items BILOG-MG 3 is utilizing a Bayesian approach and the MMLE approach in BILOG-MG 3 is then referred to by others as the marginalized Bayesian item parameter estimation procedure (Baker & Kim, 2004; Harwell & Baker, 1991). However, it is easier to consider the marginalized Bayesian model as an extension of MMLE (Baker & Kim,

2004). To keep things simple, only the prior distributions imposed on the item parameters in BILOG-MG 3 are discussed.

In BILOG-MG 3 the default prior discrimination (*a*) distribution is believed to be lognormal over the range 0 to ∞ (Baker & Kim, 2004). As Mislevy (1986) describes, the rationale for this prior distribution is that most IRT applications have a_j that are greater than 0, suggesting a positively skewed distribution like the lognormal distribution. Accordingly, BILOG-MG 3 implements the transformation $\alpha_j = \log a_j$ to produce a normal distribution for each α_j with probability density function that is proportional to $e^{-(\alpha_j - \mu_j)/2\sigma_\alpha^2}$ with default $\mu_\alpha = 0$ and $\sigma_\alpha = 0.5$, which result in $\mu_a = 1.13$ and $\sigma_a = 0.6$ (Mislevy, 1986; du Toit, 2003). As will become more apparent in the next section, this convenient transformation is employed because it keeps the metric of the discrimination parameter the same in both components of the model estimation equation (Harwell & Baker, 1991, p. 384), which consists of a likelihood component, refer to Equations 16 through 18, and a prior component (Baker & Kim, 2004).

Since α_j is normally distributed, the prior component used in the item discrimination equation in BILOG-MG 3 is $\lambda = -(\alpha_j - \mu_\alpha)/\sigma_\alpha^2$ (Baker & Kim, 2004; see Mislevy, 1986, for details on how this prior component is derived). To keep in line with the marginalized Bayesian model utilized in BILOG-MG 3, this prior component is appended to the likelihood component to produce the two components of the marginalized Bayesian item parameter estimation equation (Baker & Kim, 2004).

Similarly for the *bs*, a normal prior distribution can be requested with $\mu_b = 0$ and $\sigma_b = 2$ (Zimowski et al., 2003). This prior distribution is selected because the distribution

of *bs* in IRT applications typically follow a normal distribution and vary between -4 to +4 (Harwell & Baker, 1991). By inspection of the prior component used for the item discrimination parameter, the prior component for the difficulty parameter is $\delta = -(b_j - \mu_b)/\sigma_b^2$ (Baker & Kim, 2004).

For the *c*s a prior Beta distribution is assumed with parameters ALPHA = 5 and BETA = 17. These parameters are defined as ALPHA = *mp* + 1 and BETA = *mp* + 1, where *p* is the mean of the Beta distribution and *m* is an a priori weight of 20 observations of respondents who are marking randomly (Zimowski et al., 2003). The use of a Beta prior distribution for the *c* parameters pertains to interpreting *p* as the mean probability of a correct response for an examinee with low ability. In this case $p = 1/k$, where *k* is the number of response options. By default *k* is 5 in BILOG-MG 3, so $p = .2$. The central idea behind ALPHA and BETA values is to find values that give a desired *p* value (Baker & Kim, 2004; Harwell & Baker, 1991). The prior component utilized in BILOG-MG 3 for estimating the pseudo-chance level parameter is $\eta = [(ALPHA - 2)/c_j - (BETA - 2)/(1 - c_j)]$ (Baker & Kim, 2004; see Mislevy, 1986 for details on how this prior component is derived).

The function of priors on item parameters in BILOG-MG 3. Prior components on the item parameters are utilized so parameter estimates can be constrained from taking on deviant (unreasonable) values in some data sets (Baker & Kim, 2004). Therefore, if a prior component for a parameter provides useful information, then the appending term should affect the item parameter estimation process. The role of a prior distribution in the item estimation process for an item parameter depends on how much the item parameter

estimate “shrinks” towards the mean of the item parameter prior distribution and the size of the item parameter prior distribution variance (Novick & Jackson, 1974). Essentially, the closer the item parameter estimate is to its prior distribution mean, the less the prior distribution affects the item parameter estimate, assuming other things are equal. The prior distribution variance also influences the amount of contribution a prior distribution has because a smaller standard deviation can make the prior component have a larger impact on item parameter estimation (Baker & Kim, 2004; Harwell & Baker, 1991).

It is important to mention that the choice of priors does not have a strong impact on item parameter estimates when *N* is large, but for smaller sample sizes priors play an important role and item parameter estimates will tend to drift toward the mean of the prior distribution (Rupp, 2003, pg. 376). As a result, users often use the default prior distribution values provided in BILOG-MG 3 (Harwell & Baker, 1991; Rupp, 2003) and the default priors provide reasonable estimates that work well across a variety of disciplines (Harwell & Janosky, 1991; Rupp, 2003). However, Mislevy and Stocking (1989) suggest users should understand the default values when using BILOG or in this case BILOG-MG 3.

Item parameter estimation equations in BILOG-MG 3. The marginalized Bayesian item parameter estimation equations, written in Gaussian quadrature form, are (Baker & Kim, 2004)

$$L_1(\alpha_j) = e^{\alpha_j}(1-c_j)\sum_{q}^{Q}[\bar{r}_{jq} - \bar{n}_{jq}P_j(X_q)]w_{jq}(X_q - b_j) + \lambda \qquad (20)$$

$$L_2(b_j) = -e^{\alpha_j}(1-c_j)\sum_{q}^{Q}[\bar{r}_{jq} - \bar{n}_{jq}P_j(X_q)]w_{jq} + \delta \tag{21}$$

$$L_3(c_j) = (1-c_j)^{-1}\sum_{q}^{Q}\frac{[\bar{r}_{jq} - \bar{n}_{jq}P_j(X_q)]}{P_j(X_q)} + \eta\,. \tag{22}$$

The first part of Equations 20 through 22 each consist of the marginalized likelihood component for each item parameter in Gaussian quadrature form, while the latter part of each equation appends the prior component. The prior component allows us to examine the effect of a prior distribution on estimating an item parameter. Prior distributions are important because they supplement the information found in the sample data; as a result, if the prior distribution is informative, the second component (the prior component) should have an effect on the item parameter estimation process (Baker & Kim, 2004).

The Fisher scoring-for-parameters method. To solve Equations 20 through 22 they are each set equal to 0 and the item parameter estimates for a single item are estimated simultaneously by the Fisher scoring-for-parameters method within the context of an EM algorithm (Baker & Kim, 2004; Mislevy, 1986). Because item parameter estimates for a particular item do not depend on the parameters of other items, the estimation process continues one item at a time (Baker & Kim, 2004).

The Fisher scoring equations to be solved iteratively are (Baker & Kim, 2004)

$$\begin{bmatrix} \hat{a}_j \\ \hat{b}_j \\ \hat{c}_j \end{bmatrix}_{(t+1)} = \begin{bmatrix} \hat{a}_j \\ \hat{b}_j \\ \hat{c}_j \end{bmatrix}_{(t)} - \begin{bmatrix} \Lambda_{11} \Lambda_{12} \Lambda_{13} \\ \Lambda_{21} \Lambda_{22} \Lambda_{23} \\ \Lambda_{31} \Lambda_{32} \Lambda_{33} \end{bmatrix}^{-1}_{(t)} x \begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix}_{(t)} \tag{23}$$

where

$t = 1 \,..\, T$

$$\Lambda_{11} = -(e^{\alpha_j})^2 \sum_q^Q \overline{n}_{jq} (X_q - b_j) \left[\frac{P_j(X_q) - c_j}{1 - c_j} \right]^2 \frac{Q_j(X_q)}{P_j(X_q)} - \frac{1}{\sigma_\alpha^2} \tag{24}$$

$$\Lambda_{22} = -(e^{\alpha_j})^2 \sum_q^Q \overline{n}_{jq} \left[\frac{P_j(X_q) - c_j}{1 - c_j} \right]^2 \frac{Q_j(X_q)}{P_j(X_q)} - \frac{1}{\sigma_\beta^2} \tag{25}$$

$$\Lambda_{33} = -\sum_q^Q \frac{\overline{n}_{jq}}{(1 - c_j)^2} \frac{Q_j(X_q)}{P_j(X_q)} - \frac{\alpha_\beta - 2}{c_j^2} \frac{\beta_\beta - 2}{(1 - c_j)^2} \tag{26}$$

$$\Lambda_{12} = \Lambda_{21} = (e^{\alpha_j})^2 \sum_q^Q \overline{n}_{jq} (X_q - b_j) \left[\frac{P_j(X_q) - c_j}{1 - c_j} \right]^2 \frac{Q_j(X_q)}{P_j(X_q)} \tag{27}$$

$$\Lambda_{13} = \Lambda_{31} = -(e^{\alpha_j})^2 \sum_q^Q \overline{n}_{jq} (X_q - b_j) \frac{P_j(X_q) - c_j}{1 - c_j} \frac{Q_j(X_q)}{P_j(X_q)} \tag{28}$$

$$\Lambda_{23} = \Lambda_{32} = e^{\alpha_j} \sum_{q}^{Q} \overline{n}_{jq} \frac{P_j(X_q) - c_j}{(1 - c_j)^2} \frac{Q_j(X_q)}{P_j(X_q)}. \quad (29)$$

The iterative solution of Equation 23 is known as the Fisher-scoring method for item parameters (Baker & Kim, 2004). The matrix in equation 23 is known as the Fisher-scoring information matrix. By taking the inverse of the information matrix the variance-covariance matrix of item parameter estimates is derived and the square-root of the main diagonals of this matrix produce the asymptotic standard errors of the item parameter estimates (Baker & Kim, 2004).

Summary of the BILOG-MG 3 approach for estimating item parameters and standard errors. To solve the item parameter estimation equations (i.e., Equations 20 through 22) the so-called EM algorithm and Fisher-scoring methods are used (du Toit, 2003). "In general, the EM algorithm is an iterative procedure for finding maximum likelihood estimates of parameters of probability models in the presence of unobserved random variables" (Baker & Kim, 2004, p. 169). The E stands for expectation and M stands for maximization. Conceptually, the (iterative) method of obtaining item parameter estimates begins with provisional estimates of the item parameters and successfully updating it through a series of E steps and M cycles until our item parameter equations are all essentially 0 or close enough to zero based on a convergence criterion (Baker & Kim, 2004). More concretely, the method of estimating item parameters in BILOG-MG 3 can be summarized in three steps (Baker & Kim, 2004, p. 171; Harwell et al., 1988, p. 255):

1. The E-step:

    a) Use Equation 15 and provisional values of the item parameter estimates to compute the likelihood that each examinee's vector of item responses to the *J* items at each of the *q* quadrature points.

    b) Use Equation 15 and the quadrature weights $A(X_q)$ at each of the *q* quadrature points to calculate the posterior probability that the ability of the *i*th examinee is X_q.

    b) Calculate $\bar{n}_{jq}$ and $\bar{r}_{jq}$ for each item at each of the *q* ability (quadrature) points.

2. The M-step: Solve the marginal Bayesian item parameter estimation Equations of 20 through 22 treating the artificial data, $\bar{n}_{jq}$ and $\bar{r}_{jq}$, as the complete data (or as constants). Since Equations 20 through 22 are nonlinear in the parameters, a series of Fisher-scoring steps (iterations) (sometimes referred to as the Newton-Gauss method or Newton-Raphson procedure, Baker & Kim, 2004, p. 40; see also Harwell et al., 1988), Equation 23, for parameters is used within the M-step of the EM algorithm to obtain the item parameter estimates (Baker & Kim, 2004) and SEEs. This means that within each Fisher-scoring iteration an adjustment (improvement) is made to the item parameter estimate. This continues until a minimum change in a parameter estimate between iterations is met or a convergence criterion is met (Baker & Kim, 2004). The BILOG-MG 3 default number of Newton-Gauss (Fisher-scoring) iterations during the M-step is set at $T = 2$ and the convergence criterion within the M-step is .01 (du Toit, 2003).

3. Repeat steps 1 and 2 until the item parameter estimates are unchanged from the previous EM cycle or the item estimation process has converged at some criterion. If convergence has not occurred at the end of an EM cycle, the latest parameter estimate values are used as available starting values in the next E- and M-steps (Baker & Kim, 2004). In BILOG-MG 3 the default maximum number of EM cycles is 20 with a .01 convergence criterion for the entire EM cycle (du Toit, 2003). Upon attaining overall convergence the item parameter SEEs are found by inverting the information matrix in the final Fisher-scoring solution (Baker & Kim, 2004).

It is important to note that before each E-step of the item parameter estimation process in BILOG-MG 3, adjusted quadrature weights are computed and an undocumented algorithm is used to normalize the histogram so that the following constraints are met: $\sum_{q}^{Q} A(X_q)X_q = 0$, $\sum_{q}^{Q} A(X_q)X_q^2 = 1$, and $\sum_{q}^{Q} A(X_q) = 1$ (Harwell et al., 1988). It is also important to point out that a complete description of all internal workings of BILOG-MG 3 has not been documented in great detail. As such, the procedure discussed is based mostly in part on the BILOG-MG 3 manual (du Toit, 2003), Baker and Kim (2004), Harwell et al. (1988), and Harwell and Baker (1991). However, the X_q values remain the same throughout both the E-step and M-step of the estimation process (Baker & Kim, 2004).

Variables that may influence item parameter SEEs in BILOG-MG 3. Because the estimation technique used in BILOG-MG 3 uses Gaussian quadrature methods, the number of quadrature points used in the estimation process, as seen in Equations 20 through 22 and 24 through 28, may impact item parameter SEEs. Inspection of Equations

13 and 14 shows that the artificial data are taken over the number of examinees (*I*), while Equation 15 shows the likelihood is taken over the number of items (*J*). This means the number of items and examinees may each play a role in the item parameter SEEs. Additionally, inspection of Equations 20 through 22 and 24 through 25 show the values of the hyper parameters for the prior *a*, *b*, and *c* distributions may affect the item parameter SEEs. It can also be seen by inspection of Equations 24 through 28 that other parameter estimates for an item (e.g., c_j) play a role in the estimation of item parameter standard errors for the same item parameter. It is important to point out that the number of iterations, *T*, utilized during the Fisher-scoring procedure, number of EM cycles, and convergence criterion for the entire EM cycle may each impact the estimation of item parameter standard errors.

Previous Research Involving BILOG or BILOG-MG

Table 1 below provides a summary of research involving the program BILOG or BILOG-MG. As Table 1 shows numerous simulation studies have assessed the accuracy of item parameter estimates produced by BILOG. Most of the research involving BILOG has primarily focused on the accuracy of item parameter estimates produced by the MMLE procedure under the 2PL and/or 3PL model and how these estimates compare to those produced by other estimation programs under varying sample sizes and test lengths. Also, some of the articles have considered the impact a prior distribution on the *a* parameter (i.e., varying the variance of the *a* parameter prior distribution) would have on item parameter estimates. The results of all these articles provide a bright outlook on the performance of BILOG, as the generating item parameters were successfully recovered in

most articles. Unfortunately, none of these articles have considered the accuracy of standard errors of item parameters produced in BILOG or BILOG-MG.

Table 1

Summary of BILOG and BILOG-MG Articles

Article	Purpose of study	Design	Findings
Yen, W. M. (1987)	Compared BILOG and LOGIST CPU time, item parameter estimates, item characteristic functions (ICF), trait estimates, and true scores under the 3PL model.	Program (BILOG and LOGIST) by test length (one 10-item test, four 20-item test, and four 40-item test) by θ distribution (normal and nonnormal distributions) for an N = 1,000 under the 3PL model.	Results indicated that BILOG generally produced more accurate item parameter estimates. BILOG also produced more accurate ICF for the 10-item test, but both programs provided similar accuracy in ICF under the 20- and 40-item tests.
Baker, F. B. (1990)	Examined the equating of BILOG results to an underlying metric for the 2PL model for three different datasets under seven varying specifications for the prior discrimination distribution.	Three sets of item response data were generated for a 45-item test and 500 simulees under the 2PL model. Dataset 1, 2, and 3 had the following generating parameters ($\theta_\mu = 0$, $\sigma_\theta = 1$, $a_{min} = 1$, $a_{max} = 2$, $b = 0$, $\sigma_b = .8$), ($\theta_\mu = -.5$, $\sigma_\theta = 1.5$, $a_{min} = .5$, $a_{max} = 1.5$, $b = .5$, $\sigma_b = .8$), and ($\theta_\mu = .5$, $\sigma_\theta = .75$, $a_{min} = .3$, $a_{max} = .7$, $b = -.5$, $\sigma_b = .8$), respectively. Each dataset also had seven different specified item discrimination priors (no prior; default prior $\mu = 0$, $\sigma = .5$, no Float option; default prior with Float option; prior $\mu = 0$, $\sigma = .75$, no Float option; prior $\mu = 0$, $\sigma = .75$, with Float option; prior $\mu = 0$, $\sigma =$	The results indicated that item parameters were recovered accurately in BILOG. Also, the estimated mean difficulty and θ parameters were not impacted by the prior discrimination distribution characteristics. Moreover, the results showed that BILOG preserved the underlying θ distribution variance when it was small, but standardized the variance when the

		.25, no Float option; prior $\mu = 0$, $\sigma = .25$, with Float option.	underlying θ distribution had a large variance.
Lim, R. G., & Drasgow, F. (1990)	Compared MMLE (with no prior distributions) and Bayes model estimation when assessing DIF under a 2PL model for two sample sizes in BILOG.	Sample size (250 and 750) by estimation (MMLE with no priors or Bayes model estimation with priors) for a 20-item test under the 2PL model.	Results for both estimation methods were similar for $n = 750$, but MMLE (with no priors) showed slightly less estimation error than Bayes model estimates for $n = 250$.
Seong, T. J. (1990)	Examined the impact type of prior θ distribution, underlying θ distribution, number of examinees, and number of quadrature points had on item and θ estimates in the MMLE procedure used in BILOG.	Type of prior θ distribution (normal, positively-, and negatively-skewed) by underlying θ distribution (normal, positively-, and negatively-skewed) by number of examinees (100 and 1,000) by number of quadrature points (10 and 20) for a 45-item test under the Two-parameter normal ogive IRT model.	Results indicated item parameters were more accurately estimated when the two θ distributions matched and number of examinees was large. Also, the number of quadrature points improved the accuracy of item parameter estimates, but only when the two θ distributions matched and the number of examinees was large.
Harwell, M. R., & Janosky, J. E. (1991)	Examined the efficiency of BILOG to recover item parameters under varying prior variances for the *a* parameter, sample size, and test length for the 2PL model.	Number of examinees (75, 100, 150, 250, 500, and 1,000) by number of items (15 and 25) by variance for the prior distributions of *a* (no prior, $.75^2$, $.5^2$, $.25^2$, and $.1^2$ in a lognormal metric) for the 2PL model.	Results suggested that for samples of 250 or more the effect of prior variances is minimized, the prior variance plays a major role for smaller samples and shorter tests (i.e., 15 items) in the accuracy of the *a* parameter estimate. Thus,

			researchers should not rely on the BILOG default prior variance of $.5^2$ for the a parameter under small samples (i.e., $n < 250$) and short tests (i.e., 15 items).
Cohen, A. S., Kim, S., & Subkoviak, M. J. (1991)	Compared the influence of prior distributions on the detection of DIF in BILOG and LOGIST for two DIF methods.	Program (BILOG without priors, BILOG with priors and FLOAT option, BILOG with priors and without FLOAT option, and LOGIST) for 4 datasets (1,000 per group; 200 per group; 1,000 for group A and 200 for group B; 200 for group A and 1,000 for group B) for a 50-item test under the 2PL model.	Results indicated that item parameter estimates varied less when priors were used than when they were not used. Also, the identification of DIF was related to program and to some extent to type of dataset.
Abdel-fattah, A. A. (1994)	Examined the accuracy of item parameter estimation procedures for the 3PL model under varying sample sizes, test lengths, and underlying θ distributions.	Estimation procedure (joint maximum likelihood in LOGIST, MMLE and marginal Bayesian procedures in BILOG) by sample size (250 and 1,000) by underlying θ distribution (normal, truncated normal, and Beta) by test length (20 and 60) for the 3PL model.	Results indicated that the marginal Bayesian procedure in BILOG produced accurate item parameter estimates when the underlying θ distribution was normal or truncated normal, sample size was small, and test length was short.
Patsula, L. N., & Gessaroli, M. E. (1995)	Compared the effects test lengths and sample sizes have on the 3PL model item and ability parameter estimates obtained from BILOG and TESTGRAF.	Test length (20 and 40 items) by sample size (100, 250, 500, and 1,000) by program (BILOG and TESTGRAF) under the 3PL model and assuming the underlying θ distribution was normal.	Results indicated TESTGRAPH and BILOG provided about the same level of accuracy in item parameter estimates under most conditions. However,

			TESTGRAF was more accurate than BILOG in estimating the *c* parameter at both test lengths. Also, both programs were more accurate as sample sizes increased, but TESTGRAF was more accurate in estimating *a* and *c* parameters at all sample sizes.
Carlson, R. D. & Locklin, R. H. (1995)	Compared BILOG and MicroCat item and ability parameter estimates, and item fit statistics for the 1PL (Rasch), 2PL and 3PL models.	Program (BILOG and MicroCat) by type of IRT model (1PL, 2PL, and 3PL) by data matrix (complete and incomplete) for a 72-item mathematics test for an N = 1,000.	Both programs showed nearly identical results for *b* parameter estimates for both types of data matrices under the 1PL (Rasch) model. For the 2PL and 3PL models both programs showed close agreement for item parameter estimates using the incomplete data matrix, while strong, but weaker, agreement was found for the complete data matrix.
Parshall, C. G., Kromrey, J. D., & Chason, W. M. (1996)	Examined the impact sample size has on item parameter estimates for six IRT models in BILOG.	Sample size (100, 250, 500, and 1,000) by IRT model (1PL, 2PL, 3PL, 3PL with a restricted prior *a* distribution, 2PL with a restricted prior *a* distribution, and 3PL with restricted prior *a* distribution and common *c* parameter) for a 40-item test with simulees' θs assumed to follow a	Results indicated that using a more informative prior variance on the *a* parameter improved the fit and stability of parameter estimates relative to models with the same number of

		standard normal distribution and generated under a 3PL model.	parameters and no imposed *a* prior distribution, primarily for smaller samples.
Parshall, C. G., Kromrey, J. D., Chason, W. M., & Yi, Q. (1997)	Examined the impact sample size has on item parameter estimates for six IRT models in BILOG.	Sample size (100, 250, 500, and 1,000) by IRT model (1PL, 2PL, 3PL, 2PL with a restricted prior *a* distribution, 3PL with a restricted prior *a* distribution, and 3PL with restricted prior *a* distribution and common *c* parameter) for an 80-item test with simulees' θs assumed to follow a standard normal distribution and generated under a 6 dimensional model.	Results indicated that the additional constraints to the models (e.g., a 2PL with restricted prior *a* distribution) improved stability, but decreased both fit and accuracy, in comparison to the unconstrained models.
Baker, F. B. (1998)	Compared the item parameter recovery characteristics of a Gibb's sampling approach to the estimation approach in BILOG for varying sample sizes, test lengths, and underlying θ distribution for the Two-parameter normal ogive model.	Method (Gibb's sampling and BILOG) by sample size (30, 60, 120, and 500) by test length (10, 20, 30, and 50) for two underlying θ distributions (standard normal and normal with $\mu_\theta = .25$ and $\sigma_\theta = .83$) under the Two-parameter normal ogive model.	Results showed that a test of 50 items and 500 examinees yielded excellent item parameter recovery by BILOG. Also, BILOG's ability to recover item parameters was superior to Gibb's sampling approach under small samples and short tests.
Ban, J-C., Hanson, B. A, Wang, T., Qing, Y., & Harris, D. J. (2001)	Compared and evaluated five online pretest item calibration methods in computerized adaptive testing with respect to item parameter recovery under three sample sizes.	Method (MMLE with one EM-cycle, MMLE with multiple EM-cycles (MEM), Stocking's Method A, Stocking Method's B, and BILOG/Prior method) by sample size (300, 1,000, and 3,000) for a 30-item fixed-length adaptive test.	The MEM method provided the smallest total error in pretest item parameter calibrations under all sample size conditions, while the other methods produced results similar to MEM under the 3,000

			sample size, but the BILOG/Prior method produced the largest total error in pretest item calibrations under the 300 and 1,000 sample size conditions.
Kirisci, L., Hsu, T., & Yu, L. (2001)	Examined the effects of test dimensionality, underlying θ distribution, and IRT program on the accuracy of item and person parameter estimates under the 3PL model.	Dimensionality (Unidimensionality and three-dimensional) by underlying θ distribution (normal, positively skewed, and platykurtic) by IRT program (BILOG, MULTILOG, and XCALIBRE). Data were generated using a multidimensional compensatory 3PL model for 1,000 examinees on a 40-item test.	Overall, BILOG produced the most accurate item parameter estimates and the effect of multidimensionality on the estimation of item parameters was minimal for BILOG.
Sass, D. A., Schmitt, T. A., & Walker, C. M. (2004)	Examined the effect skewed θ distributions, sample size, test length, and estimation method have on item and ability parameter estimates under the 2PL model in BILOG-MG.	Underlying θ distribution (standard normal, skew = 1, and skew = 2) by sample size (500 and 1,000) by test length (20 and 40 items) by six estimation methods under the 2PL model.	Results indicated item parameter estimates were less precise under skewed distributions and differed a little under small sample sizes. However, results in general suggested item parameter estimates are relatively robust to skewed distributions.

Purpose Statement

Researchers using standard errors of item parameter estimates need to know if their test statistics using item parameter SEEs calibrated from IRT computer programs (e.g., BILOG-MG 3) are accurate. Existing research indicates item parameter SEEs for the Rasch (1PL) model and 2PL model are accurate under short test lengths (e.g., 5, 10, and 20 items; Drasgow, 1986, p. 85) and small to moderate sample sizes (i.e., 100 … 2,000 examinees) when using JMLE as found in WINSTEPS. However, none of the aforementioned studies have examined the accuracy of item parameter SEEs produced in BILOG-MG 3. Further, none of the studies reviewed have considered the impact of different underlying item parameter distributions, underlying ability distributions, and number of quadrature points would have on estimated standard errors. The overarching goal of this study was to identify the effect of test length, sample size, number of quadrature points, underlying item parameter(s) distribution(s), and underlying θ distribution(s) on the accuracy of item parameter SEEs for the three IRT models found in BILOG-MG 3. A rationale for each variable and levels is presented in Chapter Three.

Research Question and Hypotheses

The primary research question under consideration in this study was: Does the accuracy of item parameter SEEs produced in BILOG-MG 3 vary by underlying item parameter(s) distributions, underlying θ distribution, test length, sample size, and number of quadrature points under the 3PL, 2PL, and 1PL models? This research question was based on an inspection of the item parameter standard error estimation equations utilized in BILOG-MG 3 and from Hambleton et al. (1993), Li and Lissitz (2004, p. 91-95), and Thissen and Wainer (1982) who indicated that the shape of the underlying ability

distribution and sample size are factors affecting the size of standard errors of item parameter estimates.

As discussed in the review of literature, the following hypotheses were proposed:

1. As sample size increases, smaller item parameter SEEs result (Li & Lissitz, 2004; Thissen & Wainer, 1982) and would consequently lead to more accurate SEEs.
2. Since the number of quadrature points play a role in the estimation process (e.g., see Equations 15 through 20), it was expected that increasing the number of quadrature points would improve the accuracy of item parameter SEEs.
3. As Harwell et al. (1988, p. 247) pointed out, the item parameter estimates for a particular item do not depend on estimates of other items because the estimation process estimates item parameters and standard errors independently of other items. However, previous research (see Kirisci et al., 2001; Sass, et al., 2004) has shown that increasing test length can improve item parameter estimates. Given this information, it was predicted that test length would not have an impact on the accuracy of item parameter SEEs.
4. The item parameter SEEs would be more accurate when more of the underlying item parameter and ability distributions were similar to the prior item parameter and ability distributions specified in BILOG-MG 3, than when the underlying distributions and prior distributions were not similar.

Chapter Three

Method

Independent Variables

The following seven independent variable factors were crossed in this study: type of underlying difficulty (*b*) distribution, type of underlying discrimination (*a*) distribution, type of underlying lower asymptote (*c*) distribution, test length, type of underlying latent trait (θ) distribution, sample size, and number of quadrature points.

Underlying difficulty distribution. The underlying difficulty distribution was varied because it allowed an examination of the accuracy of the item parameter SEEs when the underlying difficulty distribution matched or did not match the prior underlying latent trait distribution assumed in BILOG-MG 3. A $N(0,1)$ distribution was selected as one level for the underlying difficulty distribution because it matched the underlying latent trait distribution assumed in BILOG-MG 3 and is typical of a difficulty distribution seen in practice (Harwell & Baker, 1991, p. 378). A second level for the underlying difficulty distribution, $U(-3,3)$, was selected because it did not match the underlying latent trait distribution assumed in BILOG-MG 3, the uniform distribution is typical for simulation studies (Kirisci et al., 2001), and $-3 \leq b \leq 3$ is the typical range of difficulty values seen in practice (Baker, 2001).

Underlying discrimination distribution. The type of underlying discrimination distribution was varied because it was expected to better inform users on the accuracy of item parameter SEEs when the underlying discrimination distribution matched or did not match the prior discrimination distribution assumed in BILOG-MG 3. As mentioned in Chapter Two, the variance for a prior distribution plays a crucial role in parameter

estimation; prior distributions with larger variances are less informative than those with smaller variances (Harwell & Janosky, 1991). Thus, the first underlying discrimination (*a*) distribution was varied to follow a lognormal distribution with $\mu_\alpha = 0$ and $\sigma^2_\alpha = .25$, $a \sim LN(0,.25)$, which resulted in $\mu_a = e^{0+.5(.25)} = 1.13$ and $\sigma^2_a = e^{2(0)+2(.25)} - e^{2(0)+.25} = .36$. Note that μ_α and σ^2_α are the respective scale and shape parameters used to determine the form of the underlying distribution. This first type of underlying discrimination distribution was chosen because it mimicked the prior discrimination distribution assumed in BILOG-MG 3. The second type of underlying discrimination distribution, $a \sim LN(0,.36)$, was chosen to not match the prior discrimination distribution used in BILOG-MG 3 but to reflect an underlying discrimination distribution that was more realistic (i.e., had more variability). A common $a = 1$ for all items was also used because this restriction along with the additional restriction of $c = 0$ (described below) enabled one to examine the effect the aforementioned factors had under the 1PL model.

Underlying lower asymptote distribution. The type of underlying lower asymptote distribution was varied because it informs users on the accuracy of item parameter SEEs when the underlying lower asymptote distribution matched or did not match the default prior lower asymptote distribution assumed in BILOG-MG 3. Three levels were chosen for the underlying lower asymptote distribution: $c \sim BETA4(5,17,0,1)$ and another four parameter Beta distribution, $c \sim BETA4(9,33,0,1)$, and fixed $c = 0$ for all items. In the four parameter Beta distribution the first two values represented the two shape parameters, α and β, while the last two values, l and u, represented the lower and upper limit of the distribution. The first distribution was chosen to match the default prior *c* distribution assumed in BILOG-MG 3. The second distribution reflected a four parameter Beta

distribution with less variability and a lower mean than the default Beta distribution, but maintained a realistic underlying *c* distribution that may be seen in practice. In addition, fixing $c = 0$ for all items was chosen because this restriction enabled us to examine the effect that the aforementioned factors had under the 2PL model.

Test length. The test lengths examined in this study were: $J = 50$ and $J = 10$. A 50-item test was chosen to represent a long test (i.e., more than 20 items; du Toit, 2003, p. 603) and was longer than the average test length based on a review of research which applied the 3PL model. A 10-item test was selected because it represented a short test (i.e., 11 to 20 items; du Toit, 2003, p. 603; 5 or 10 items; Drasgow, 1986, p. 85) and was shorter than the test lengths that might be seen when measuring some attitudinal constructs or student behaviors, where only a few items may be administered to an examinee. A survey of selected empirical studies (DeMars, 2003; El-Korashy, 1995; Obiekwa, 2001; Richichi, 1996, Wightman & De Champlain, 1994) that applied the 3PL model between 1980 and 2005 (selected from the Eric Education from First Search database using the keywords, subject phrases, or combinations such as Item Response Theory, Latent Trait Theory, Calibration) had a mean and median test length of 39 and 34, respectively ($SE_{mean} = 6$, $SD = 17$, $n = 8$). The test lengths for the calibrations conducted in these eight studies were 25, 25, 25, 25, 42, 47, 53, and 70. Note that some studies conducted calibrations for multiple samples. Mislevy and Stocking (1989) have suggested MMLE methods, as utilized in BILOG-MG 3, should produce dependable item parameter estimates, even for short tests reliant upon the accuracy of the (IRT) model. Moreover, Cohen et al. (1991) suggested that BILOG-MG should produce accurate item parameter estimates for short tests.

Underlying latent trait distribution. The type of underlying latent trait (θ) distribution was varied in this study because standard text books on IRT have noted that it has an impact on the SEE of item parameters (e.g., Embretson & Reise, 2000, p. 195; Hambleton & Swaminathan, 1985). It has also been shown that characteristics of the prior θ distribution affect item parameter estimates, and the correct specification of the prior θ distribution produce MMLE item parameter estimates that are consistent (Harwell et al., 1988). Thus, varying the θ distribution allowed us to consider the accuracy of item parameter SEEs when the prior latent trait distribution specified in BILOG-MG 3 matched or did not match the underlying θ distribution. The two underlying θ distributions were selected so one mimicked the default features found in BILOG-MG 3 and another corresponded to underlying θ distributions not assumed in BILOG-MG 3. To test this effect an underlying θ distribution that was $N(0,1)$, which matched the default assumed in BILOG-MG 3, was compared to estimates produced from a positively-skewed underlying θ distribution that did not match the BILOG-MG 3 default. The second level for the underlying θ distribution, a positively-skewed distribution, $\theta \sim \chi^2(5)$ standardized to have a mean of -.5, was chosen because not all underlying latent trait distributions are normally distributed in educational applications of IRT (Seong, 1990). It is important to note that the positively-skewed distribution will be referred to as $\theta \sim \chi^2$ for the remainder of the study.

Sample size. Sample size was selected because it has been shown to have an important effect on the accuracy of item parameter estimation (see Seong, 1990) and the minimum sample size needed to provide accurate item parameter estimates was a primary concern during calibration. The two sample sizes investigated were: $I = 500$ and $I =$

4,000. Rupp (2003) recommended an $I = 500$ as a minimum guideline for reaching stable parameter estimates for tests consisting of 15 to 50 items for the 3PL model and Cohen et al. (1991) have suggested that BILOG-MG should produce accurate item parameter estimates for small samples; however, Cohen et al. (1991) did not define what they meant by small samples. Additionally, research has shown that samples of 500 are just below the minimum sample size recommended for the 3PL model (Hulin, Lissak, & Drasgow, 1982).

However, Thissen and Wainer (1982) suggest larger samples are needed to better estimate item parameters and reduce the magnitude of item parameter standard errors. Thus, $I = 4{,}000$ was selected to reflect a large sample size. This larger sample size is within the range of sample sizes that applied researchers use with the 3PL model. As described previously, a survey of selected studies which applied the 3PL model had a mean and median sample size of 4,647 and 263, respectively ($SE_{\text{mean}} = 4{,}337$, $SD = 12{,}265$, $n = 8$). The sample sizes for the calibrations conducted in these eight studies were 230, 240, 247, 255, 270, 433, 500, and 35,000.

Number of quadrature points. The number of quadrature points was varied to inform us on whether or not there is a gain in the accuracy of the estimation of item parameter SEEs beyond the number of quadrature points used as the default in BILOG-MG 3. The default used in BILOG-MG 3 is 15 quadrature points. 60 quadrature points was also used during item parameter standard error estimation because more quadrature points provided a better approximation to a continuous distribution (i.e., fewer gaps in the distribution) and improved the accuracy of item parameter estimates (Seong, 1990).

Data Generation and Calibrations

The first step in the data generation process was to generate population *a*, *b*, and *c* item parameters for the 50-item and 10-item length tests. Using SAS, a macro program was written to generate two sets of 50 *b* item parameters from a *N*(0,1) or *U*(-3,3) distribution, two sets of 50 *a* item parameters from a *LN*(0,.25) or *LN*(0,.36) distribution, and two sets of 50 *c* item parameters from a *BETA4*(5,17,0,1) or *BETA4*(9,33,0,1) distribution. Next, two sets of 10 *b* item parameters were generated from a *N*(0,1) or *U*(-3,3) distribution, two sets of 10 *a* item parameters were generated from a *LN*(0,.25) or *LN*(0,.36) distribution, and two sets of 10 *c* item parameters were generated from a *BETA4*(5,17,0,1) or *BETA4*(9,33,0,1) distribution. All item *a*, *b*, and *c* parameters were randomly and independently generated. A detailed summary of the item parameter generating distributions and sampled parameters is found in Appendix A. A listing of the sampled item parameters for conditions with 50- and 10-item length tests are provided in Appendix B and C, respectively.

Once item parameters for the various test lengths had been created, item response data was generated. A modified version of a SAS macro program written by Whittaker, Fitzpatrick, Williams, and Dodd (2003) was used to generate θ values for simulees from the appropriate underlying θ distribution and item responses for simulees based on the 3PL model. Appendix D provides a modified version of the Whittaker et al. (2003) SAS macro program that was used to generate four of the conditions in this study. To generate the item responses, a simulee was randomly assigned an ability value from a given underlying θ distribution (*N*(0,1) or χ^2). Using the defined item parameters for a specified test length and the simulee's ability value, the probability of answering an item correct

was computed according to the 3PL model. This probability was compared to a random number sampled from a uniform distribution with domain (0,1). A simulee's response was considered correct (1) when the probability exceeded or was equal to the random number; otherwise, the simulee's response was scored incorrect (0). This process was repeated for every simulee and every item.

Then, all simulated datasets were calibrated with BILOG-MG 3. In running BILOG-MG 3 all default options were used except when manipulations to default features were needed for testing a particular independent variable in this study (i.e., changing the number of quadrature points). In addition, the default ridge constant of RIDGE = (2, 0.1, 0.01) was changed to RIDGE = (2, .01, 0.2) on the BILOG CALIB line, but this was only done for the 3PL model calibrations. This modification to the ridge constant was done to combat the excessively high number of nonconverging datasets exhibited during preliminary 3PL model calibrations, which occurred from the algorithm getting stuck and bouncing in the Newton phase. BILOG-MG 3 was selected given its frequent use within IRT. Furthermore, this recent version of BILOG-MG 3 has not been evaluated with the various combinations of the independent variables used in this study. Sample 1PL, 2PL, and 3PL model calibration input files for BILOG-MG 3 are provided in Appendices E, F, and G, respectively.

To summarize, data for this study were simulated for two test lengths (50, 10), two sample sizes (500, 4,000), two underlying θ distributions ($\theta \sim N(0,1)$, $\theta \sim \chi^2$), two underlying difficulty (*b)* distributions, ($b \sim N(0,1)$ or $b \sim U(-3,3)$), three underlying discrimination (*a)* distributions ($a \sim LN(0,.25)$, $a \sim LN(0,.36)$, $a = 1$), three underlying lower asymptote (*c)* distributions ($c \sim BETA4(5,17,0,1)$, $c \sim BETA4(9,33,0,1)$, $c = 0$), and

two number of quadrature points levels (15, 60). However, the conditions resulted in a partially factorial design because certain combinations of the manipulated independent variable conditions did not lend themselves to meaningful IRT models, and were subsequently ignored (i.e., generating item responses for a 2PL model when $a = 1$ and c varies). This simulation study had 128 conditions under the 3PL model, 64 conditions under the 2PL model, and 32 conditions under the 1PL model, for a total of 224 conditions. Appendix H presents all the levels of the conditions simulated in this study. One thousand datasets were generated for each of the 224 conditions, with a pair of "unique" seeds (starting values to begin the random number generators) used for each condition. The first seed was used for the random number generator when selecting a simulee's θ value, while the second seed was used for the random number generator when selecting a random uniform value to compare a simluee's response probability against.

Data Analysis

After all BILOG-MG 3 runs had completed the convergence rates and percentages of omitted items were recorded. The next step was to examine the accuracy of the item parameter SEEs produced by BILOG-MG 3 by calculating the average estimate of bias (AEBias) and root mean square error (RMSE) for each item within each condition

$$\text{AEBias}(\hat{\xi}_j) = \frac{1}{R}\sum_{r=1}^{R} SE(\hat{\xi}_{jr}) - SE_{emperical}(\hat{\xi}_j) \tag{30}$$

and

$$\text{RMSE}(\hat{\xi}_j) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(SE(\hat{\xi}_{jr}) - SE_{emperical}(\hat{\xi}_j))^2} \quad (31)$$

where

j and r, respectively, denote items and replications,

$SE(\hat{\xi}_{jr})$ is an item parameter's standard error estimate in the rth replication,

R is the number of replications (1,000 in this case), and

$SE_{emperical}(\hat{\xi}_j)$ is defined as

$$SE_{emperical}(\hat{\xi}_j) = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\xi}_{jr} - \bar{\hat{\xi}}_j)^2} \quad (32)$$

where

$\hat{\xi}_{jr}$ is an item parameter's estimate in the rth replication, and

$\bar{\hat{\xi}}_j$ is the mean of the item parameter estimates over replications for parameter ξ_j.

Note that 1,000 replications per condition was selected to provide more stable analysis of results and it is also greater than the number of replications typically found in parameter estimation studies. For conditions that did not achieve convergence R was less than 1,000, but this did not happen for many conditions as described in the results section.

AEBias measured the magnitude and direction of bias for a particular estimated item parameter standard error relative to the corresponding item parameter standard error,

as measured by $SE_{emperical}$. RMSE measured the average unknown discrepancy between an item parameter standard error, as measured by $SE_{emperical}$, and the corresponding estimated item parameter standard error. These two measures, AEBias and RMSE, represented the dependent variables in this study. Estimation accuracy was evaluated at the item level across replications and was not averaged across all items.

Chapter Four

Results

Convergence and Omitted Items

Convergence rates for the 3PL, 2PL, and 1PL models were 98.05%, 98.96%, and 99.98%, respectively, with an overall convergence rate of 98.6%. Generally, for the 3PL model calibrations, nonconvergence was high (i.e., > 5%) for *J* of 50, *I* of 500, underlying θ distributed *N*(0,1), and underlying *b* distributed *U*(-3,3) conditions. Nonconvergence was also high for 3PL conditions based on *J* of 10, *I* of 4,000, underlying θ distributed χ^2, *b* distributed *U*(-3,3), *a* distributed *LN*(0,.36), and *c* distributed *Beta4*(5,17,0,1); and those conditions also based on *J* of 10, *I* of 4,000, underlying θ distributed χ^2, *b* distributed *U*(-3,3), *a* distributed *LN*(0,.25), *c* distributed *Beta4*(5,17,0,1), and 60 quadrature points. For the 2PL model calibrations, nonconvergence was high for *J* of 10, *I* of 500, underlying θ distributed χ^2, and underlying *b* distributed *U*(-3,3) conditions. Also, under the 2PL model, nonconvergence was high for *J* of 10, *I* of 500, underlying θ distributed χ^2, *b* distributed *U*(-3,3), and *a* distributed *LN*(0,.36). Nonconvergence was not high for any particular condition under the 1PL model. A detailed summary of the percentage of nonconvergence within conditions for the 3PL, 2PL, and 1PL models, respectively, is provided in Appendices I, J, and K.

Although nonconvergence was not a problem, a small number of replications within conditions did have one less item estimated during the calibration. An item was omitted from the BILOG-MG 3 calibration process when its item biserial correlation was less than the program's criterion (i.e., biserial correlations less than -.15). Items were only

omitted from the 2PL and 3PL model calibrations. These omitted items generally came from a 50-item length test consisting of low discrimination values (i.e., $a = .064$ or $a = .208$ or $a = .264$) with the exception of one item having a discrimination value of 1.054, $b = .177$, and $c = .231$. Omitted items under the 10-item length test came from 3PL model calibrations and consisted of the more difficult items (i.e., $b = 2.169$ or $b = 2.605$) for this test length, but *a* and *c* parameters were not unreasonably low or high. A summary of items omitted by condition are provided in Appendix L.

Gap Analysis

After the removal of nonconverging datasets and items that were omitted from the BILOG-MG 3 calibration process a gap analysis was performed. A gap analysis was performed because upon inspection of plots showing RMSE as a function of parameter values it became noticeably clear that there were exceptionally large RMSE values (i.e., > 1 but < 15), which tended to influence the overall trend presented in the plots. Notably, this gap in the plots trend only occurred under the 3PL model. As such, all 3PL model conditions were further scrutinized to identify which item(s) and replication(s) within each condition had potentially influential item difficulty parameter SEEs. This was done by inspecting plots of item difficulty parameter SEEs as a function of replications for each item within each condition. A particular item's replication was considered for removal if the following two conditions were met. One, a particular item difficulty parameter SEE displayed a gap between its estimate and other item difficulty SEEs. Two, the same item replication difficulty SEE was twice the size of its difficulty parameter estimate. Consequently, all item replications within a condition displaying both a gap and having SEE twice the size of an item's parameter estimate were removed from all future

analyses. Items removed from the gap analysis all came from data sets generated from the 3PL model with *I* of 500 and, with the exception of one item, all items consisted of above average positive *b*s (i.e., $b > 1.296$), low to moderate *a*s (i.e., $.3 \leq a \leq 1.462$), and low to high *c* parameters (i.e., $.098 \leq c \leq .435$). In addition, all *b* parameter estimates were at least $b = 6.19$ or greater and had *b* parameter SEE ranging from 13.27 to 450.63. A summary of the items removed from the gap analysis is provided in Appendix M.

RMSE and Bias as a Function of Parameter Values

RMSE standard error of difficulty results. Figure 2 contains plots of the relationship between the RMSE standard error of *b* (SE_b) and the *b* parameter for 15 quadrature points under the 1PL model conditions. The patterns under the 60 quadrature points 1PL model conditions can be inferred from these plots because they mimicked what was observed for 15 quadrature points. In general, the accuracy of estimation of SE_b was not a function of *b* for conditions having an *I* of 4,000 or conditions based on underlying *b* and θ distributed $N(0,1)$ with an *I* of 500 (Figures 2a and 2c). For the remaining conditions based on *I* of 500 (Figures 2b and 2d), the accuracy of estimation of SE_b was a function of *b*. Specifically, the RMSE SE_b increased a little in magnitude for extreme *b*s (i.e., *b*s in both tails of the distribution) for conditions based on underlying *b* distributed $U(-3,3)$ (Figure 2d) and conditions based on underlying θ distributed χ^2 and underlying *b* distributed $N(0,1)$ (Figure 2b).

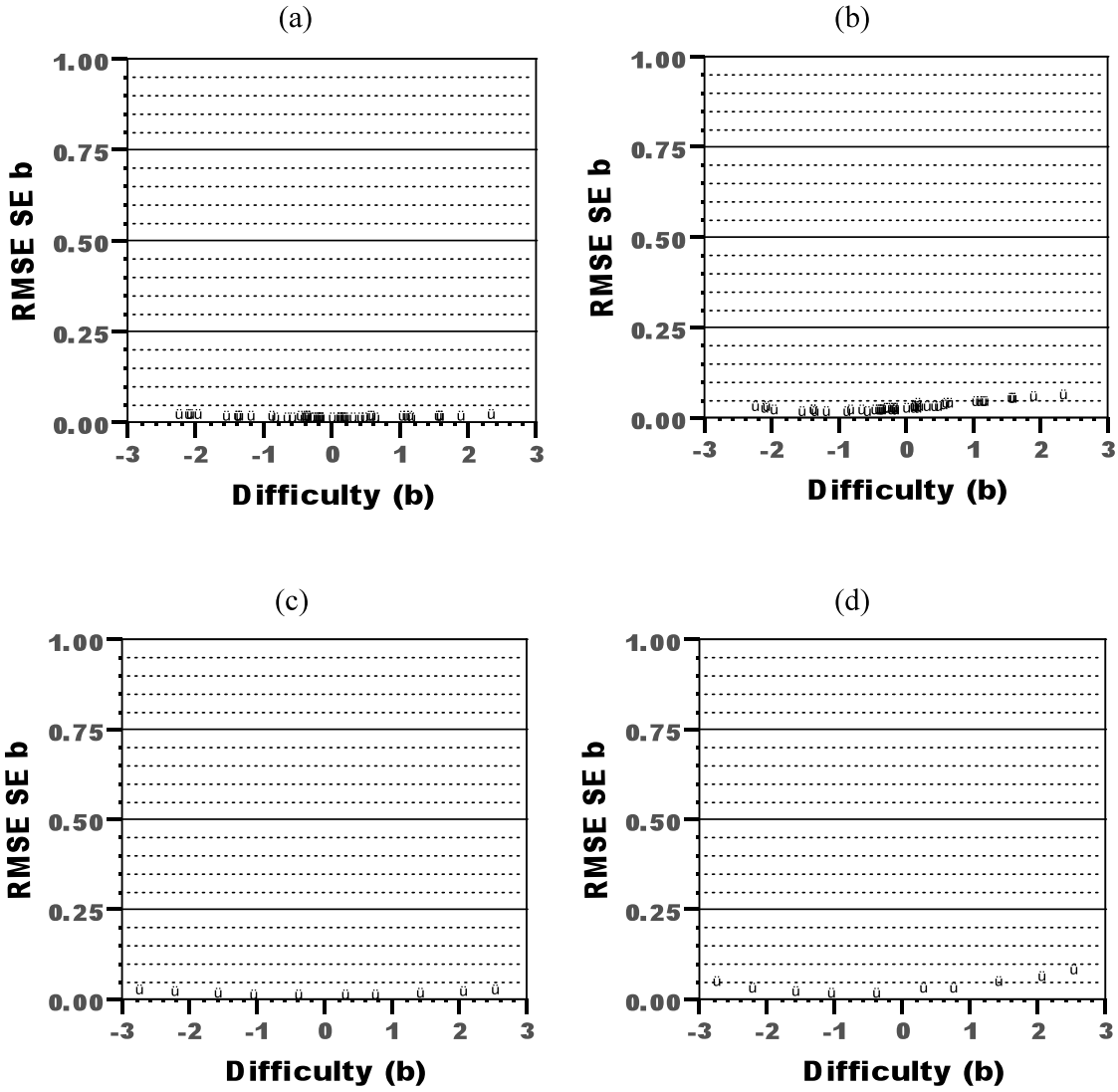

Figure 2. Relationship between the RMSE standard error of b and the item difficulty parameter under the 1PL model. All plots come from 15 quadrature points. Plots a and b are based on J of 50, $b \sim N(0,1)$, while plots c and d are based on J of 10 and $b \sim U(-3,3)$. Also, plots a and c are based on I of 4,000 and $\theta \sim N(0,1)$, while plots b and d are based on I of 500 and $\theta \sim \chi^{2}$.

Plots of the relationship between the RMSE SE_b and b parameters for 15 quadrature points under the 10-item and 50-item 2PL model conditions are presented in Figure 3. The patterns observed under the 60 quadrature points 2PL model conditions can be inferred from these plots based on 15 quadrature points as varying the number of quadrature points did not impact the observed patterns. The patterns shown in Figures 3a and 3b, respectively, represent the typical pattern seen in conditions based on I of 500, J of 50, and underlying b distributed $N(0,1)$ or $U(-3,3)$. Figure 3c represents the trend seen in conditions based on underlying b distributed $U(-3,3)$, J of 10, and I of 500. Figure 3d represents the RMSE SE_b patterns seen in conditions based on J of 10, underlying b distributed $N(0,1)$, with I of 500 conditions, and all I of 4,000 conditions. The accuracy of estimation of SE_b was a function of b. For both test lengths, the accuracy of estimation of SE_b increased as I increased (Figures 3a and 3b). Moreover, the accuracy of estimation of SE_b tended to drop for larger bs, this trend was exaggerated in almost all of the I of 500 conditions (Figures 3a, 3b, and 3c).

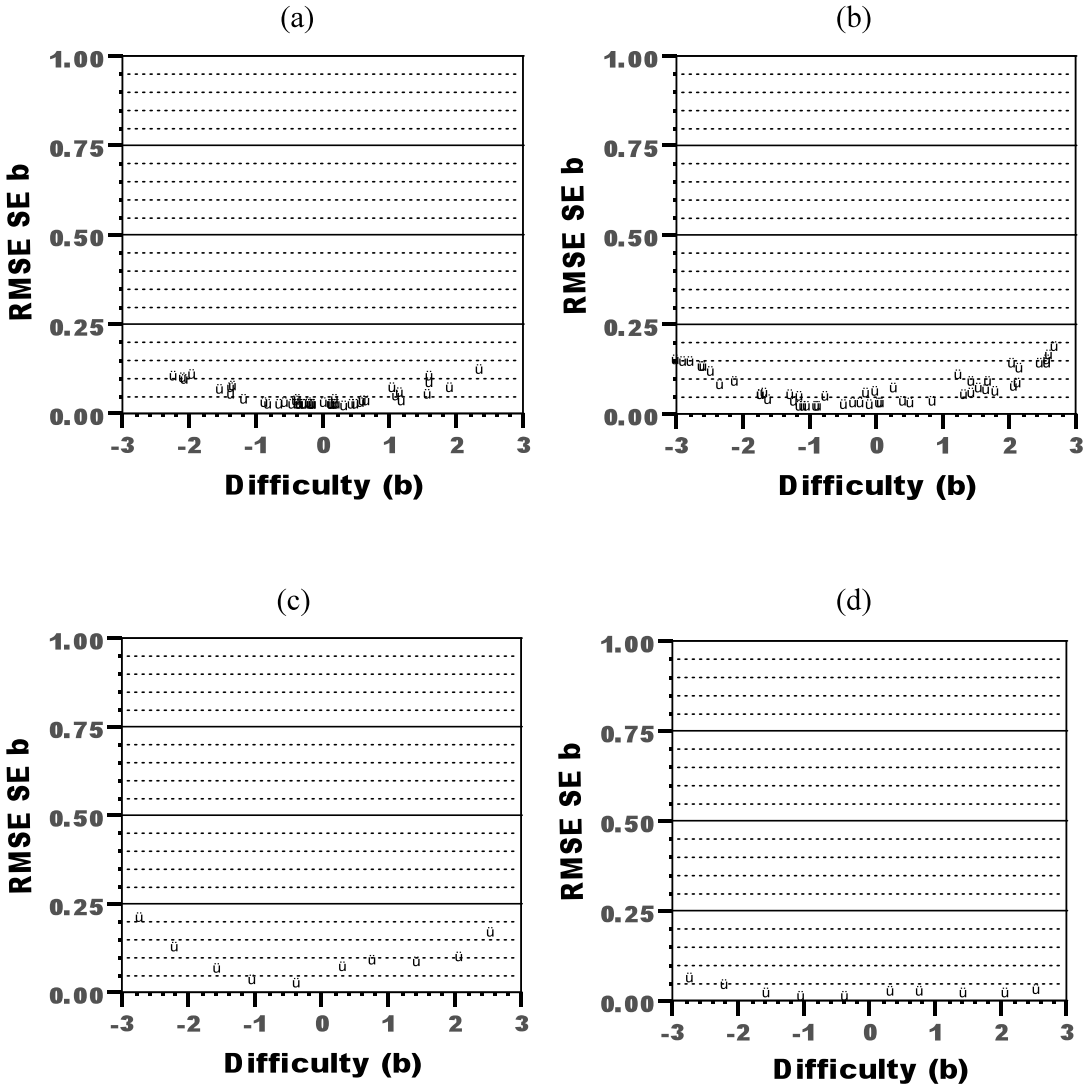

Figure 3. Relationship between the RMSE standard error of b and the item difficulty parameter under the 2PL model. All plots come from 15 quadrature points. Plots a and b are based on J of 50, while plots c and d are bsed on J of 10. Also, plots a, b, c are based on I of 500, while plot d is basd on I of 4,000. Moreover, plot a is based on $b \sim N(0,1)$, $a \sim LN(0,.25)$, and $\theta \sim N(0,1)$, while plots b, c, and d are based on $b \sim U(-3,3)$, $a \sim LN(0,.36)$, and $\theta \sim \chi^2$.

The relationship between the RMSE SE_b and b parameters for 15 quadrature points under the 3PL model conditions are presented in Figures 4 and 5. The patterns observed under the 60 quadrature points 3PL model conditions were the same as those observed for 15 quadrature points conditions. Figure 4a shows the typical trend seen for

conditions based on *J* of 50, underlying *b* distributed *U*(-3,3), underlying *a* distributed *LN*(0,.25) and *I* of 500. Figure 4b represents conditions based on *J* of 50, underlying *b* distributed *N*(0,1), and underlying *a* distributed *LN*(0,.25) as well as conditions based on *J* of 50, underlying *b* distributed *U*(-3,3), underlying *a* distributed *LN*(0,.25), and *I* of 4,000. Figure 4c is representative of the trend seen for conditions based on *J* of 50, underlying *a* distributed *LN*(0,.36) and *I* of 500, while Figure 4d represents the same conditions except *I* of 4,000. Figures 5a and 5c, respectively, represent the trends for conditions based on underlying *b* distributed *N*(0,1) and *U*(-3,3) with *I* of 500, while Figures 5b and 5d represent the trends for conditions based on underlying *b* distributed *N*(0,1) and *U*(-3,3) with *I* of 4,000. All of the plots show that the variability in RMSE SE_b tended to vary across conditions and was often unsystematic.

In general, the accuracy of estimation of SE_b was not a function of *b*s for all 3PL model conditions, but it was for conditions based on *J* of 50, *I* of 500, underlying *a* distributed *LN*(0,.25), and underlying *b* distributed *U*(-3,3) (see Figure 4a). The accuracy of estimation of SE_b was also a function of *b*s for conditions based on *J* of 10, *I* of 500, and underlying *b* distributed *U*(-3,3) (Figure 5c). In these previously mentioned conditions, the accuracy of estimation of SE_b tended to diminish for larger *b*s, creating a "*j*" shape. For conditions based on *J* of 50, underlying *a* distributed *LN*(0,.25), and *I* of 4,000 as well as those based on *I* of 500, *J* of 50, underlying *b* distributed *N*(0,1), and underlying *a* distributed *LN*(0,.25), RMSE SE_b was consistently estimated across the range of *b*s (Figure 4b). This same pattern was also seen for the remaining *J* of 10 conditions shown in Figures 5a, 5b, and 5d. In the remaining *J* of 50 conditions, RMSE SE_b had no systematic scatter across the range of *b*s (See Figures 4c and 4d).

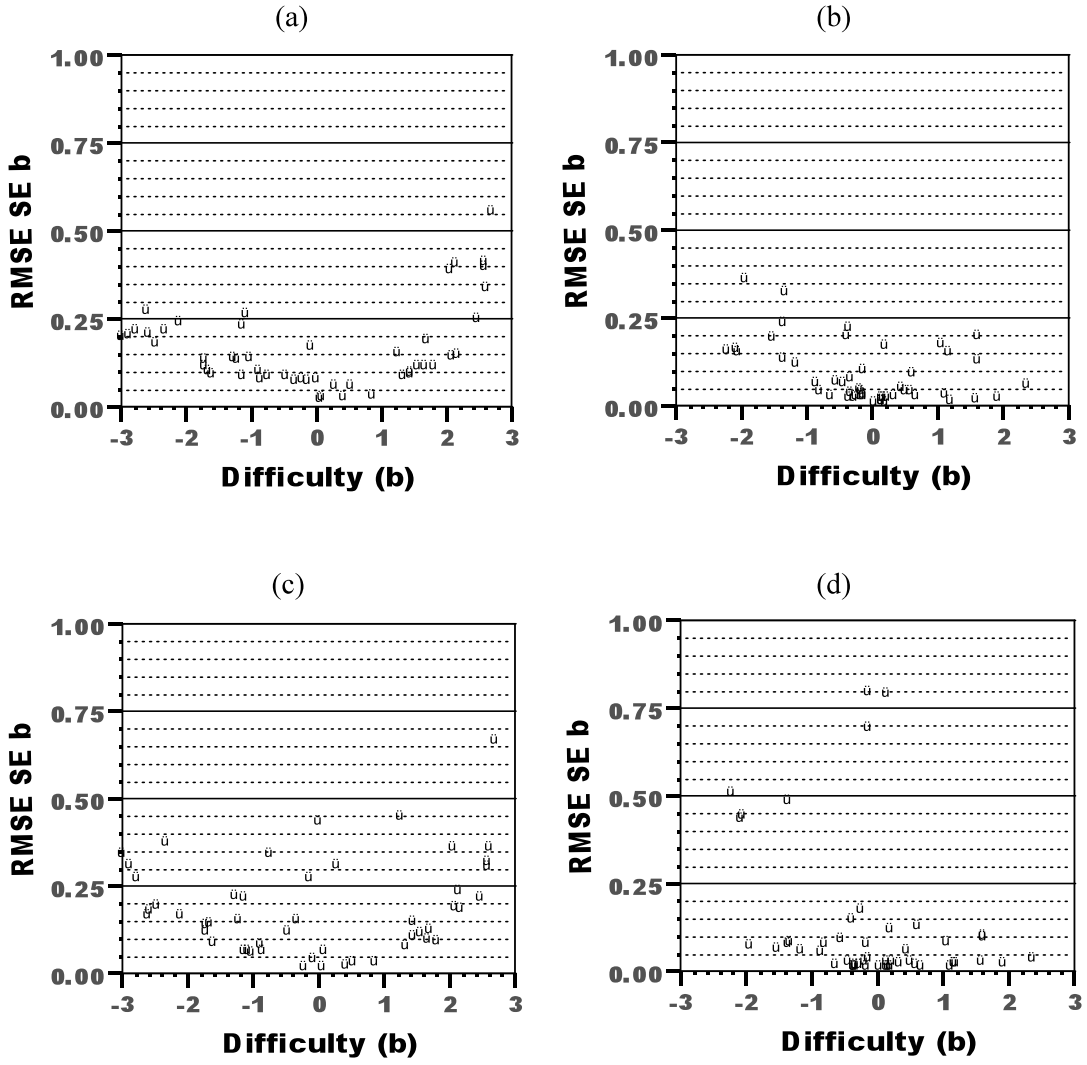

Figure 4. Relationship between the RMSE standard error of b and the item difficulty parameter under the 3PL model (J = 50). All plots come from 15 quadrature points. Plots a and c are based on $b \sim U(-3,3)$ and I of 500, while plots b and d are based on $b \sim N(0,1)$ and I of 4,000. Plots a and b are based on $a \sim LN(0,.25)$, while c and d are based on $a \sim LN(0,.36)$. Also, plots a, b, and d are based on $c \sim Beta4(5,17,0,1)$, while plot c is based on $c \sim Beta4(9,33,0,1)$. Moreover, plot b is based on $\theta \sim N(0,1)$, while plots a, c, and d are based on $\theta \sim \chi^2(5)$.

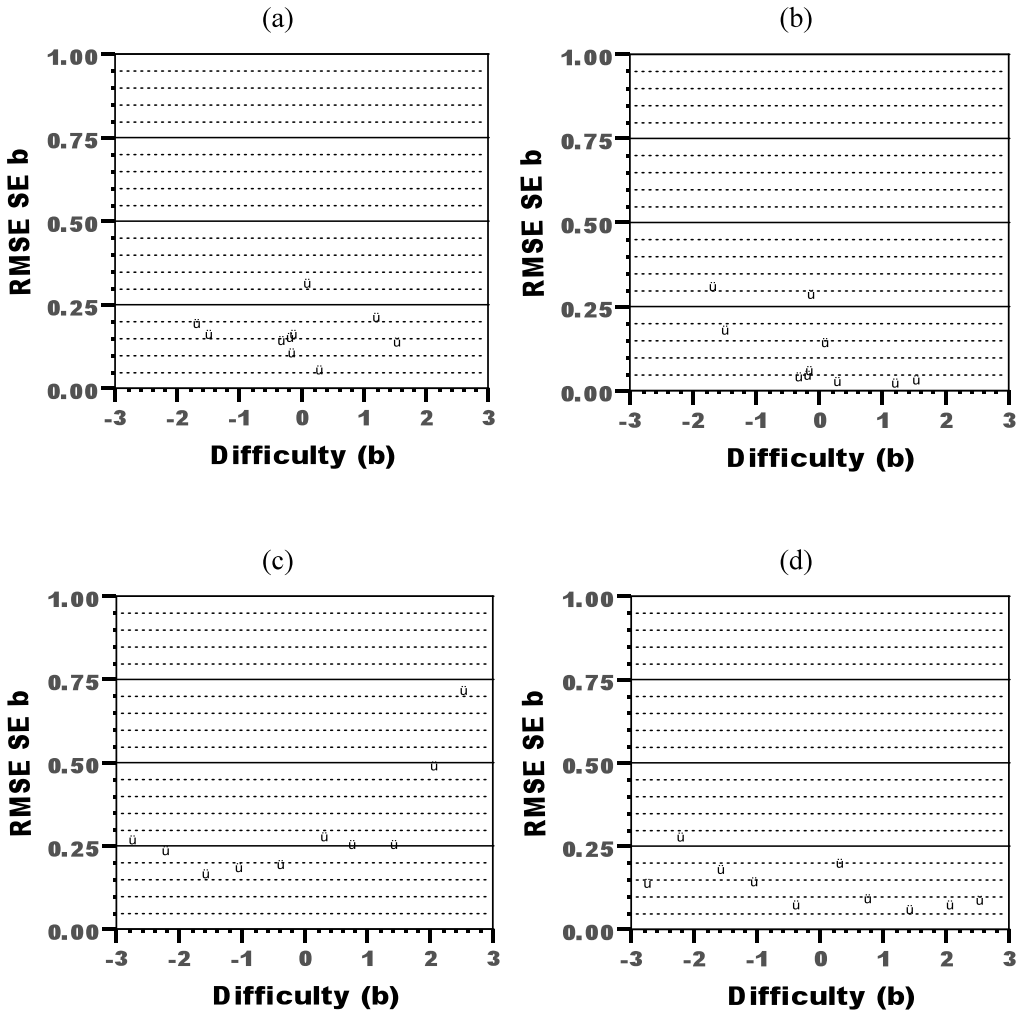

Figure 5. Relationship between the RMSE standard error of b and the item difficulty parameter under the 3PL model (J = 10). All plots come from 15 quadrature points. Plots a and c are based on $b \sim N(0,1)$ and I of 500, while plots b and d are based on $b \sim U(-3,3)$ and I of 4,000. Moreover, plot a is based on $a \sim LN(0,.25)$, $c \sim Beta4(5,17,0,1)$, and $\theta \sim N(0,1)$, while plots b, c, and d are based on $a \sim LN(0,.36)$, $c \sim Beta4(9,33,0,1)$, and $\theta \sim \chi^{2.}$

RMSE standard error of discrimination results. The relationship between the RMSE SE_a and a parameters for 15 quadrature points under the 50-item and 10-item 2PL model conditions are presented in Figures 6 and 7, respectively. The results for the 60 quadrature points 2PL model conditions can be inferred from these plots as they did not differ from those observed for the 15 quadrature points conditions. Figure 6a is typical of the trend observed in conditions based on I of 4,000 and J of 50, while Figure 6b

represents the trend observed for conditions based on *J* of 50, *I* of 500, and underlying *a* distributed *LN*(0,.25). Figures 6c and 6d are typical of the range of patterns observed in conditions based on *J* of 50, *I* of 500, and underlying *a* distributed *LN*(0,.36). Figure 7a is representative of the pattern seen in the conditions based on *I* of 4,000 and *J* of 10. However, Figure 7b is typical of the pattern seen in conditions based on *I* of 4,000, *J* of 10, underlying θ distributed χ^2, and underlying *b* distributed *U*(-3,3). It is important to note that the RMSE SE_a for the largest *a* in Figure 7b was not as exaggerated when based on underlying *a* distributed *LN*(0,.25). The remaining conditions based on *I* of 500 and *J* of 10 have patterns falling somewhere between Figures 7c and 7d. In general, accuracy of estimation of SE_a tended to improve for smaller *a*s, but diminished for larger *a*s. This trend was most evident in conditions based on *I* of 500 (Figures 6b, 6c, and 6d) and conditions based on *I* of 4,000, *J* of 10, underlying θ distributed χ^2, and underlying *b* distributed *U*(-3,3) (Figures 7b, 7c, and 7d). For conditions based on *I* of 4,000, accurate estimates of SE_a were observed across the range of item *a* parameters. However, this did not hold true for larger item *a* parameters in conditions based on *I* of 4,000, *J* of 10, underlying θ distributed χ^2, and underlying *b* distributed *U*(-3,3).

The relationship between the RMSE SE_a and *a* parameters for 15 quadrature points under the 50-item 3PL model conditions are presented in Figures 8 and 9, while 10-item 3PL model conditions are in Figure 10. The results for the 60 quadrature points 2PL model conditions did not differ from the 15 quadrature points conditions; therefore they can be inferred from these plots. All *I* of 500 and *J* of 50 conditions had trends falling somewhere between Figures 8a and 8c, while all *I* of 4,000 and *J* of 50 conditions had trends falling somewhere between Figure 8b and 8d. However, Figures 9a and 9b are

exceptions to the trends presented in Figure 8. Figure 9a represents the trend observed for the condition based on *J* of 50, underlying *b* distributed *N*(0,1), underlying *a* distributed *LN*(0,.36), underlying *c* distributed *Beta4*(9,33,0,1), underlying θ distributed χ^2, and *I* of 500, while Figure 9b represents the same condition but with an *I* of 4,000. All *I* of 4,000 and *J* of 10 conditions had trends falling somewhere between Figures 10a and 10c, while all *I* of 4,000 and *J* of 50 conditions had patterns falling somewhere between Figures 10b and 10d. In general, the RMSE SE_a plots show the accuracy of estimation of SE_a was a function of *a*, where by accuracy of estimation of SE_a tended to diminish for larger *a*s. This trend was the strongest in the conditions based on *J* of 50 and *I* of 500 (Figures 8a and 8c). For the conditions based on *J* of 50 and *I* of 4,000, (Figures 8b and 8d) this trend became more evident when fewer of the underlying *a*, *c*, and θ distributions were similar to the prior distributions used in BILOG-MG 3 and underlying *b* was distributed *U*(-3,3) (Figure 8d). This trend was also discernable in the *J* of 10 conditions (Figure 10), but this trend was only realized because one extreme *a* parameter (i.e., $a > 2.5$) in these conditions had a reduction in accuracy of SE_a. For all conditions, the accuracy of estimation of SE_a tended to improve as *I* increased (Figures 8a, 8b, 9a, 9b, 10a, and 10b).

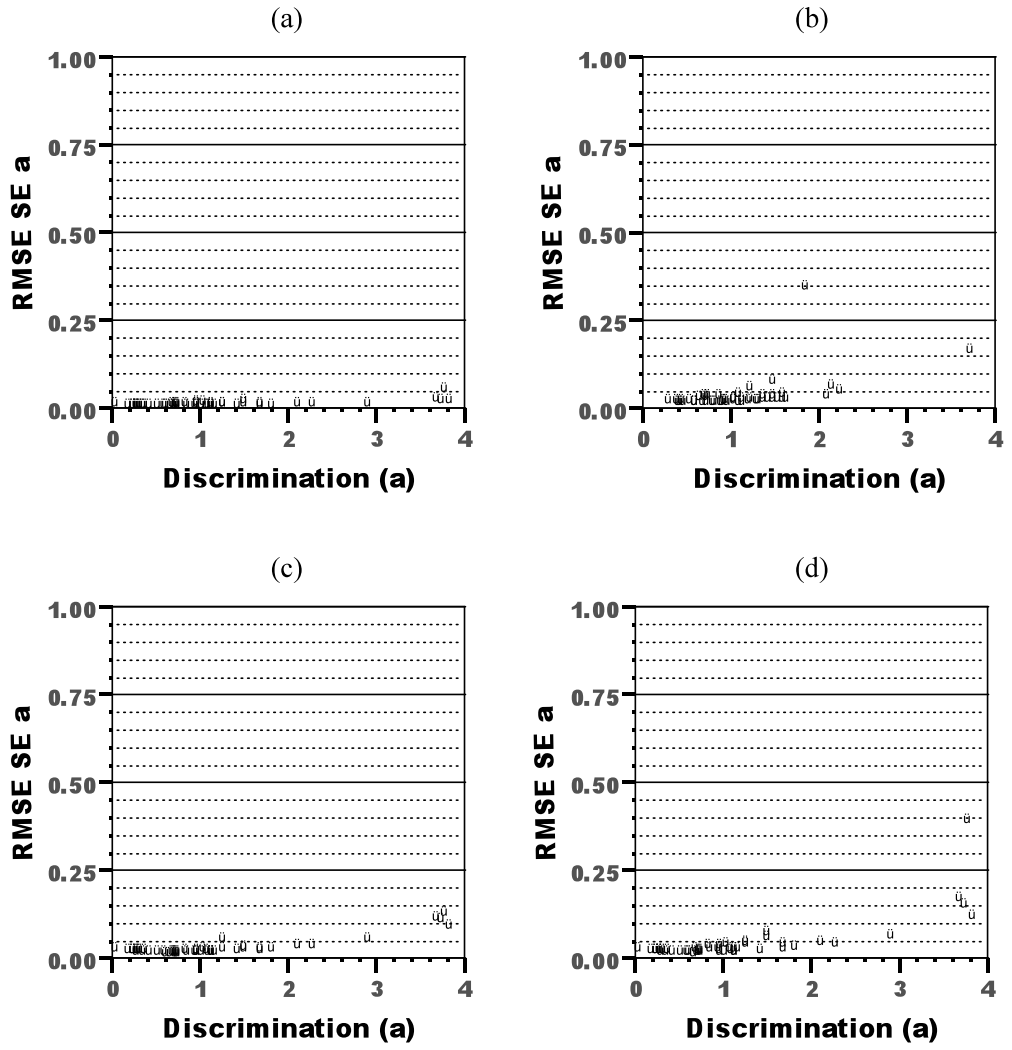

Figure 6. Relationship between the RMSE standard error of a and the item discrimination parameter under the 2PL model (J = 50). All plots come from 15 quadrature points and $\theta \sim \chi^2$. Plots a, c, and d are based on $a \sim LN(0,.36)$, while plot b is based on $a \sim LN(0,.25)$. Plots a, b, and d are based on $b \sim U(-3,3)$, while plot c is based on $b \sim N(0,1)$. Also, plot a is based on I of 4,000, while plots b, c, and d are based on I of 500.

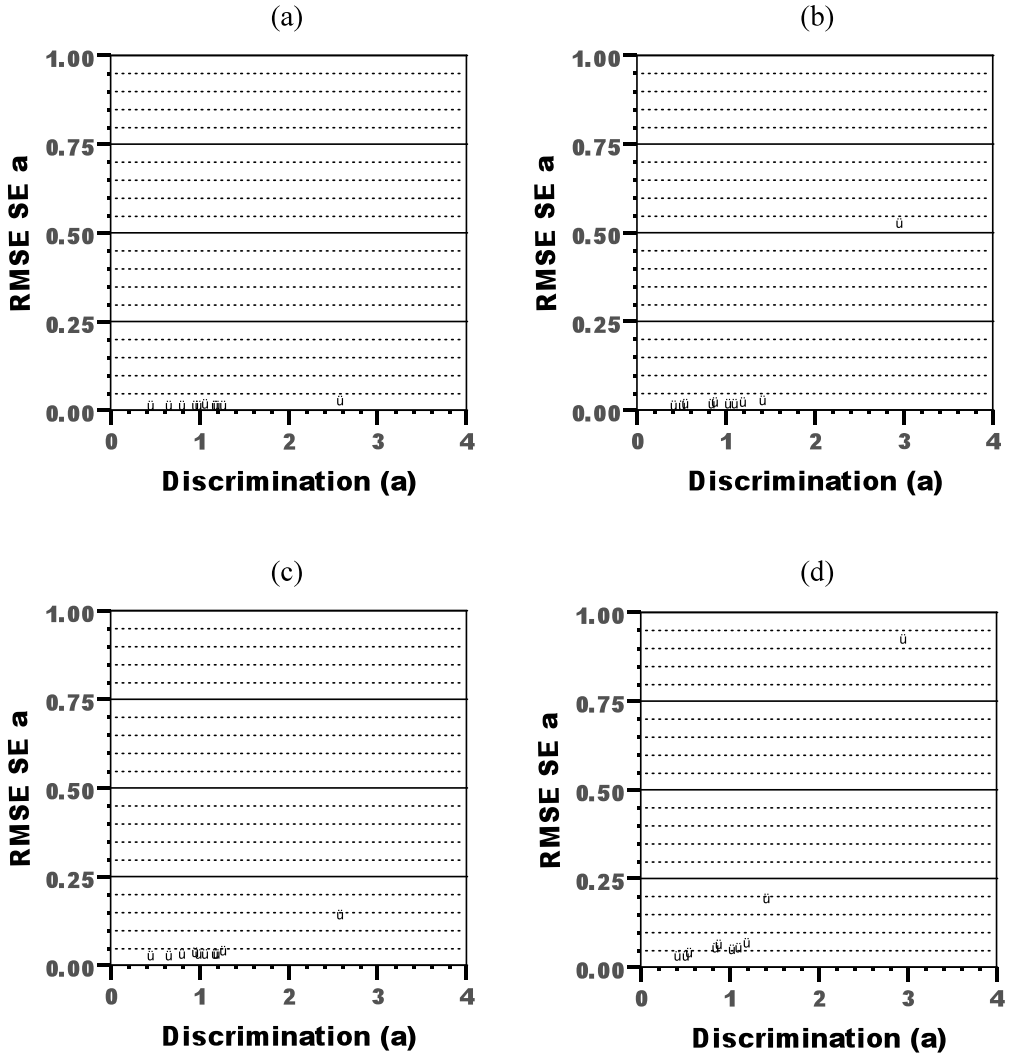

Figure 7. Relationship between the RMSE standard error of a and the item discrimination parameter under the 2PL model (J = 10). All plots come from 15 quadrature points. Plots a and c are based on $a \sim LN(0.25)$, $b \sim N(0,1)$, and $\theta \sim N(0,1)$, while plots b and d are based on $a \sim LN(0,.36)$, $b \sim U(-3,3)$, and $\theta \sim \chi^2$. Moreover, plots a and b are based on I of 4,000, while plots c and d are based on I of 500.

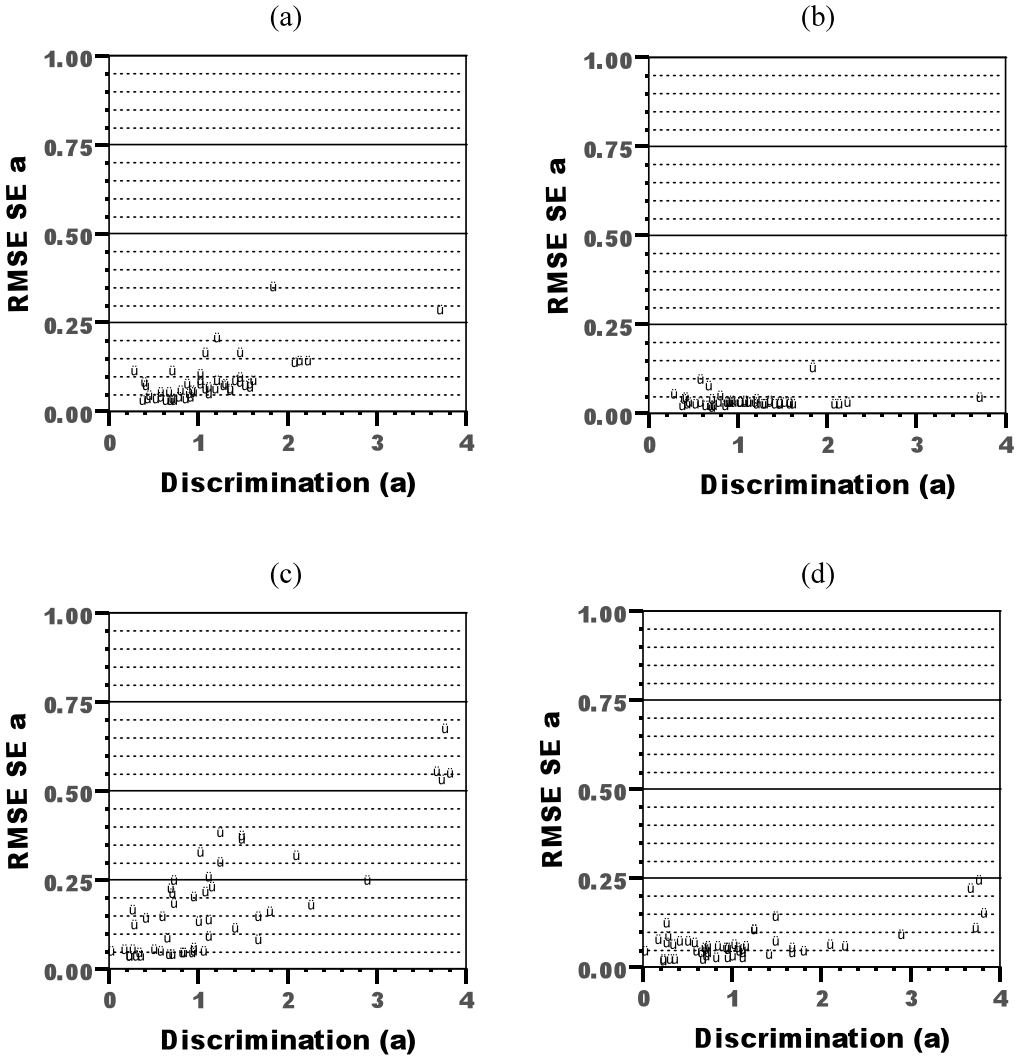

Figure 8. Relationship between the RMSE standard error of a and the item discrimination parameter under the 3PL model (J = 50). All plots come from 15 quadrature points. Plots a and b are based on $b \sim N(0,1)$, $a \sim LN(0,.25)$, $c \sim Beta4(5,17,0,1)$, and $\theta \sim N(0,1)$, while plots c and d are based on $b \sim U(-3,3)$, $a \sim (0,.36)$, $c \sim Beta4(9,33,0,1)$, and $\theta \sim \chi^2$ Also, plots a and c are based on I of 500, while plots c and d are based on I of 4,000.

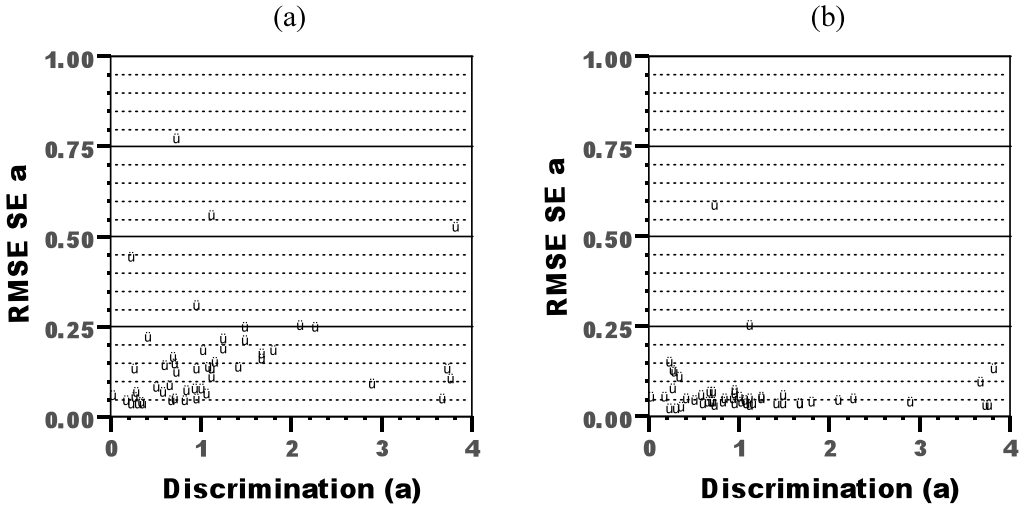

Figure 9. Relationship between the RMSE standard error of a and the item discrimination parameter under the 3PL model (more J = 50). Both plots come from 15 quadrature points, $b \sim N(0,1)$, $a \sim LN(0,.36)$, $c \sim Beta4(9,33,0,1)$, and $\theta \sim \chi^2$. Plot a is based on I of 500, while plot b is based on I of 4,000.

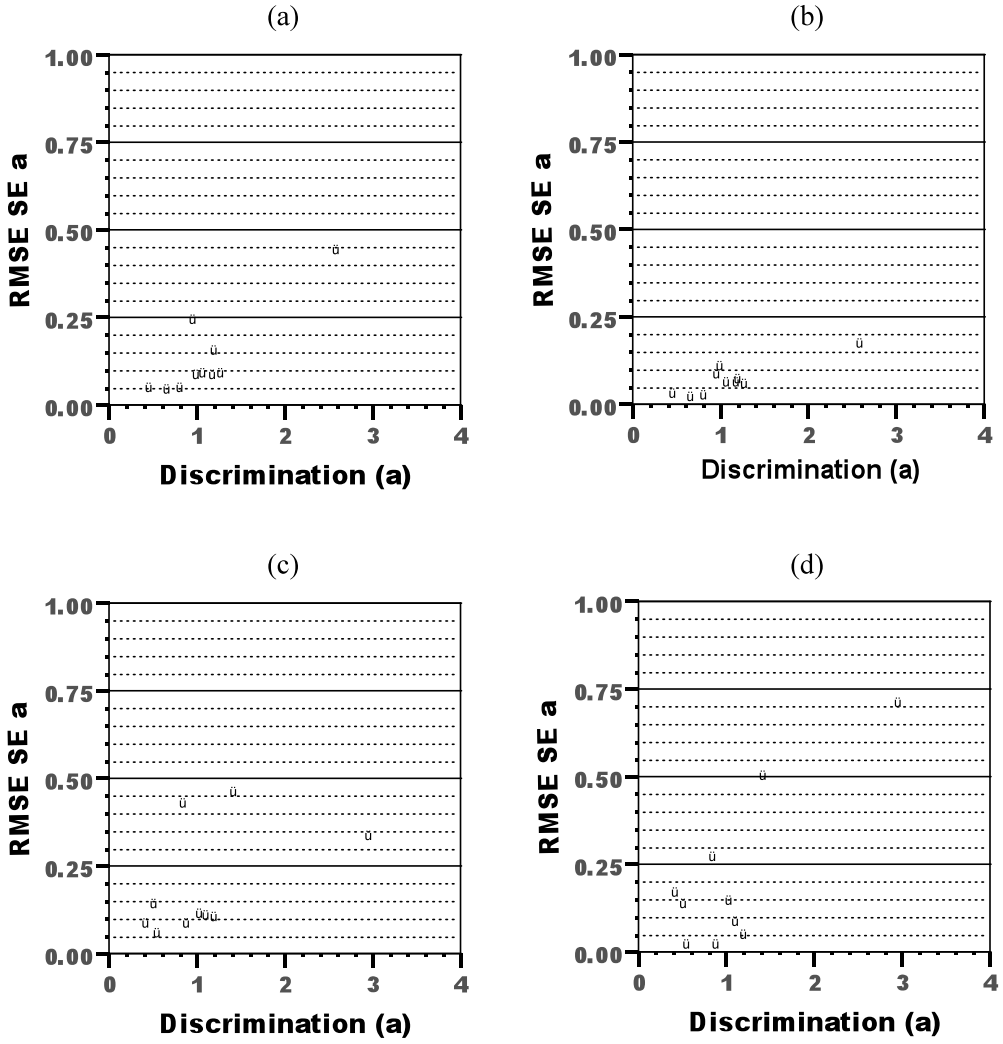

Figure 10. Relationship between the RMSE standard error of a and the item discrimination parameter under the 3PL model (J = 10). All plots come from 15 quadrature points. Plots a and b are based on $b \sim N(0,1)$, $a \sim LN(0,.25)$, $c \sim Beta4(5,17,0,1)$, and $\theta \sim N(0,1)$, while plots c and d are based on $b \sim U(-3,3)$, $a \sim (0,.36)$, $c \sim Beta4(9,33,0,1)$, and $\theta \sim \chi^2$. Also, plots a c are based on I of 500, while plots c and d are based on I of 4,000.

RMSE standard error of lower asymptote results. Figure 11 captures the typical plots of the relationship between RMSE SE_c and item c parameters, for all 15 quadrature points 3PL model conditions. Throughout the range of c parameters, RMSE SE_c was relatively uniform for all conditions. Results for the 60 quadrature points 3PL model

conditions can be inferred from these plots because they did not differ from those observed for the 15 quadrature points conditions.

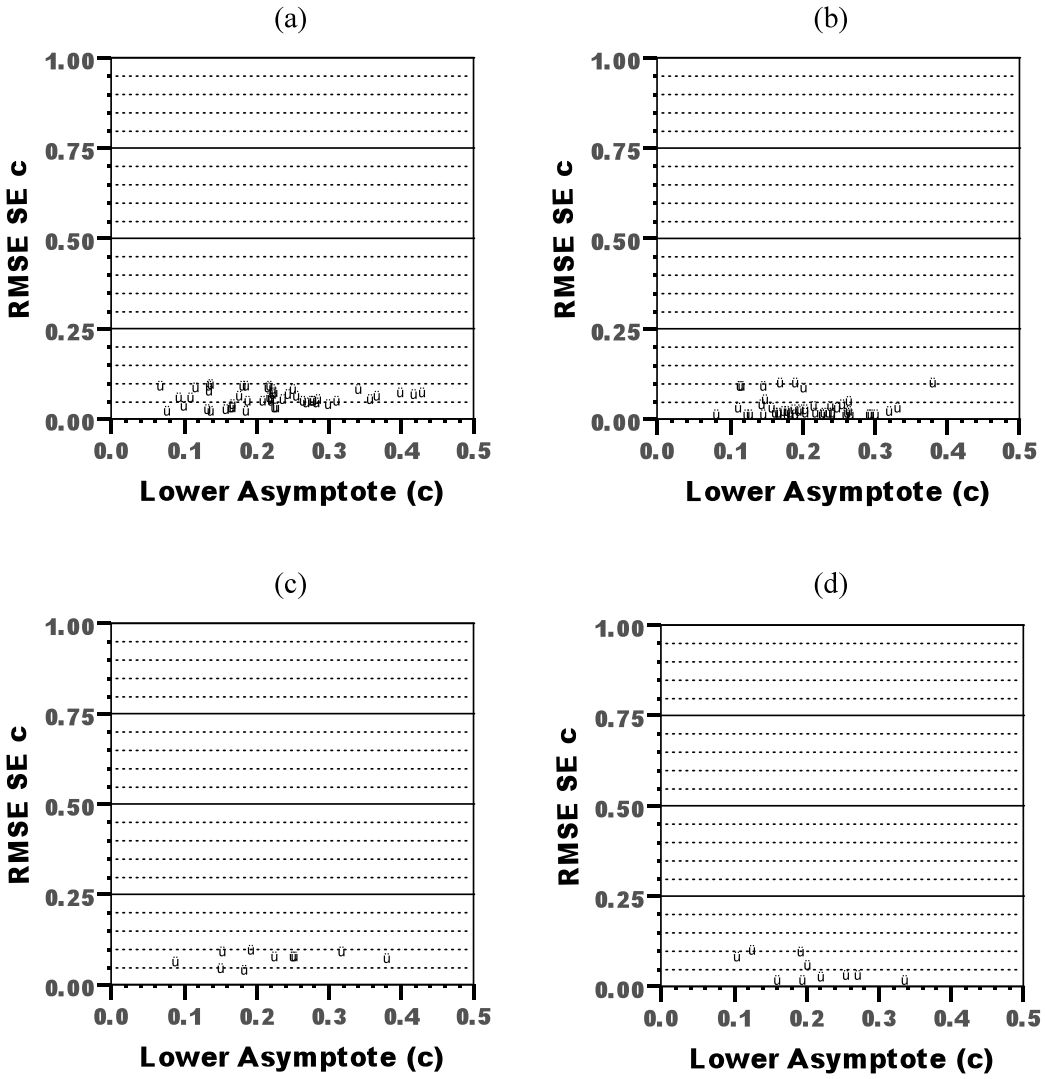

Figure 11. Relationship between the RMSE standard error of c and the item lower asymptote parameter under the 3PL model. All plots come from 15 quadrature points. Plots a and c are based on $b \sim N(0,1)$, $a \sim LN(0,.25)$, $c \sim Beta4(5,17,0,1)$, $\theta \sim N(0,1)$, and I of 500, while plots c and d are based on $b \sim U(-3,3)$, $a \sim (0,.36)$, $c \sim Beta4(9,33,0,1)$, and $\theta \sim \chi^2$ and I of 4,000. Also, plots a and b are based on J of 50, while plots c and d are based on J of 10.

Bias standard error of difficulty results. Figure 12 shows the relationship between Bias SE_b and item b parameters for 15 quadrature points under the 50-item and 10-item 1PL model conditions. Results for the 60 quadrature points 1PL model conditions can be

inferred from these plots as they did not differ from those observed for 15 quadrature points conditions. Figure 12a is typical of conditions based on *I* of 500 and underlying *b* and θ distributed *N*(0,1), while Figures 12b represents the Bias in estimation of SE_b patterns seen in conditions based on *I* of 500, underlying *b* distributed *N*(0,1), and underlying θ distributed χ^2. Figure 12c is typical of the pattern seen in conditions based on underlying *b* distributed *U*(-3,3) and *I* of 500, while Figure 12d shows the typical pattern seen in all *I* of 4,000 conditions. In general, the Bias in estimation of SE_b was not a function of *b* for 1PL model conditions. The degree of Bias in estimation of SE_b was minimal throughout the range of *b* parameters for all 1PL model conditions. However, a small negative Bias in estimation of SE_b was seen for more extreme *b*s when *I* was 500 but this excluded the conditions based on *I* of 500 and similar underlying *b* and θ distributions (Figures 12b and 12c).

Figure 13 shows the relationship between Bias SE_b and item *b* parameters for 15 quadrature points under the 50-item and 10-item 2PL model conditions. Results for the 60 quadrature points 2PL model conditions can be inferred from these plots because they did not differ from those observed for 15 quadrature points conditions. Figure 13a represents the typical pattern seen in all conditions based on *I* of 500, *J* of 50, and underlying *b* distributed *N*(0,1). Figures 13b and 13c represent the Bias in estimation of SE_b patterns seen in conditions based on *I* of 500, underlying *b* distributed *U*(-3,3), and *J* of 50 and 10, respectively. Figure 13d is typical of the pattern exhibited in all *I* of 4,000 conditions, and *I* of 500, *J* of 10, and underlying *b* distributed *N*(0,1) conditions. In general, Bias in estimation of SE_b was not a function of *b* for any of the 2PL model conditions and the degree of Bias was minimal for all 2PL model conditions.

The least amount of Bias in estimation of SE_b throughout the entire range of *b* parameters was observed for conditions based on *I* of 4,000, and *I* of 500 and underlying *b* distributed *N*(0,1) (see Figure 13a and 13d).

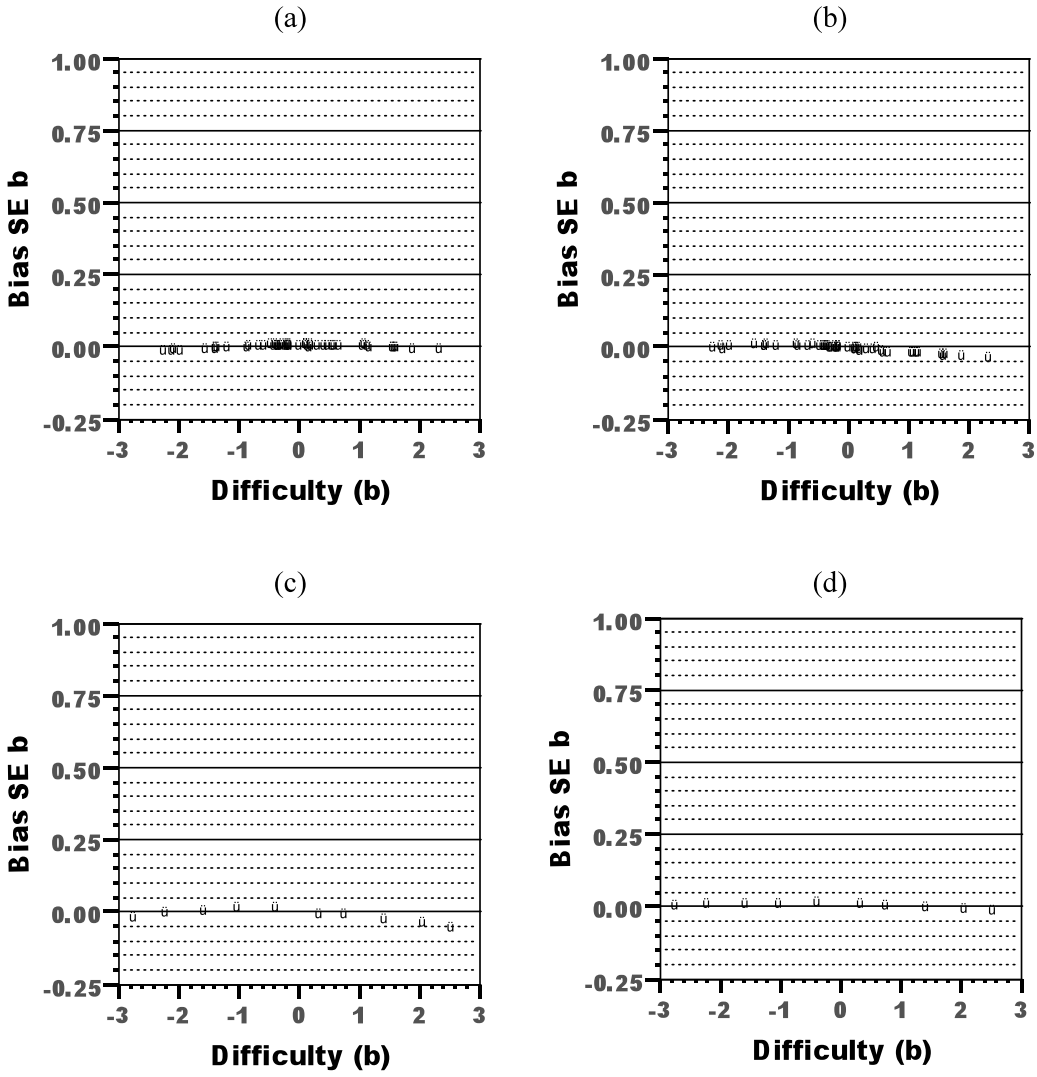

Figure 12. Relationship between the Bias standard error of b and the item difficulty parameter under the 1PL model. All plots come from 15 quadrature points. Plots a and b are based on J of 50 and $b \sim N(0,1)$, while plots c and d are based on J of 10 and $b \sim U(-3,3)$. Also, plots a, b, c are based on I of 500, while plot d is basd on I of 4,000. Moreover, plot a is based on $\theta \sim N(0,1)$, while plots b, c, and d are based on $\theta \sim \chi^2$.

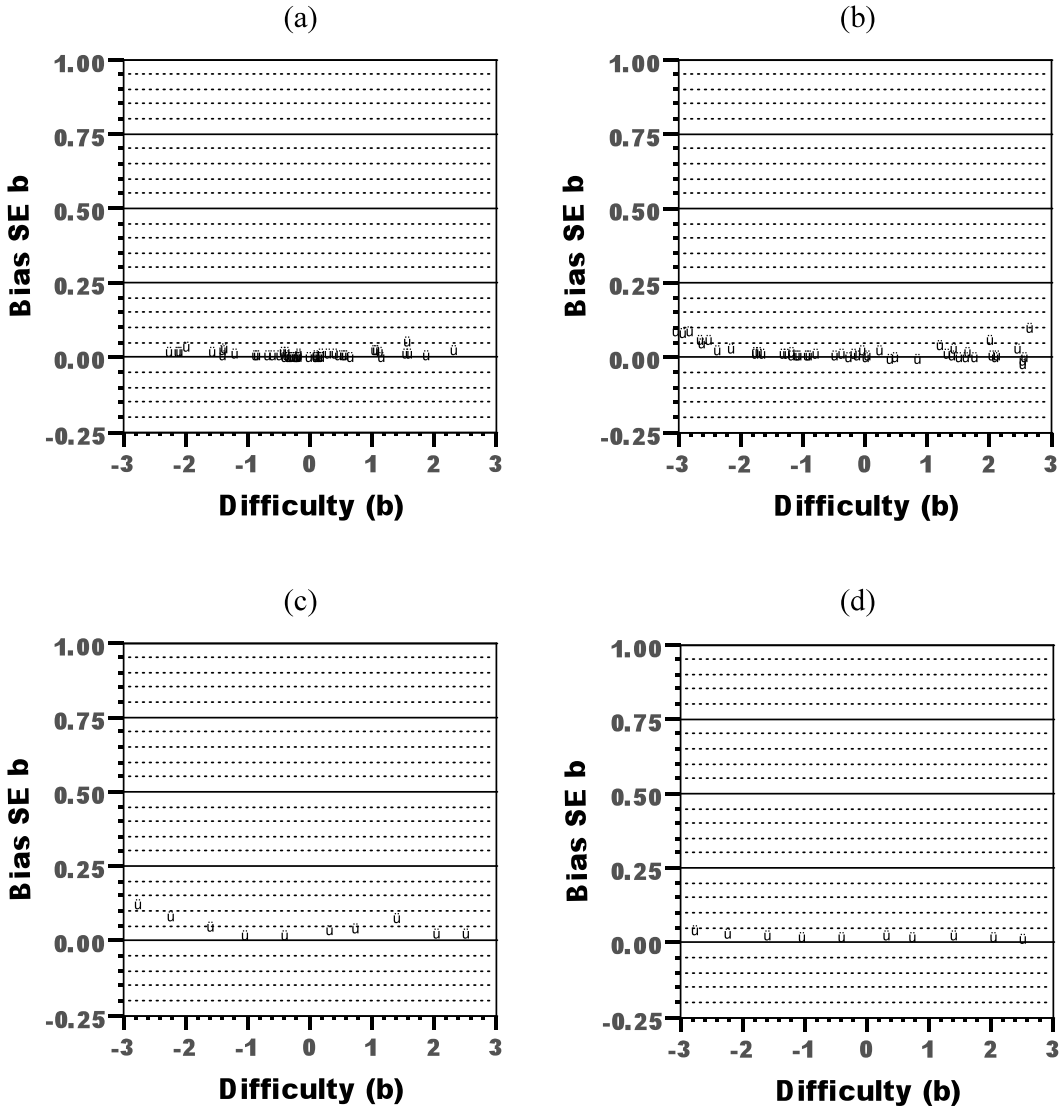

Figure 13. Relationship between the Bias standard error of b and the item difficulty parameter under the 2PL model. All plots come from 15 quadrature points. Plots a and b are based on J of 50, while plots c and d are bsed on J of 10. Also, plots a, b, c are based on I of 500, while plot d is basd on I of 4,000. Moreover, plot a is based on $b \sim N(0,1)$, $a \sim LN(0,.25)$, and $\theta \sim N(0,1)$, while plots b, c, and d are based on $b \sim U(-3,3)$, $a \sim LN(0,.36)$, and $\theta \sim \chi^2$.

Figure 14 shows the relationship between Bias SE_b and item b parameters for 15 quadrature points under the 50-item and 10-item 3PL model conditions. Results for the 60 quadrature points 3PL model conditions can be inferred from these plots as they did not differ from those presented from those observed for the 15 quadrature points

conditions. Figure 14a shows the typical trend seen for conditions based on *J* of 50, underlying *b* distributed *U*(-3,3), and underlying *a* distributed *LN*(0,.25), while Figure 14b represents conditions based on *J* of 50, underlying *b* distributed *N*(0,1), and underlying *a* distributed *LN*(0,.25). Figure 14c represents conditions based on underlying *a* distributed *LN*(0,.36) and *I* of 500, while 14d represents the same conditions except *I* of 4,000. Figures 15a and 15c, respectively, represent the trends for conditions based on underlying *b* distributed *N*(0,1) and *U*(-3,3) with *I* of 500, while Figures 15b and 15d represent the trends for conditions based on underlying *b* distributed *N*(0,1) and *U*(-3,3) with *I* of 4,000.

In general, all 3PL model conditions showed the Bias of estimation of SE_b was not a function of the *b* parameters studied (Figures 14 and 15). Also, a more positive Bias in estimation of SE_b was observed across the range of *b* parameters studied, but some *b* parameters studied did show a small amount of negative Bias in estimation of SE_b. Specifically, for the *J* of 50 conditions, the Bias SE_b decreased for larger *b*s (Figure 14), but this was not as evident for the *J* of 10 conditions (Figure 15). Morover, Bias SE_b decreased for conditions based on *J* of 50 when the underlying *a* was distributed *LN*(0,.25) (Figures 14a and 14b) relative to underlying *a* distributed *LN*(0,.36) (Figures 14c and 14d). Although the patterns seen in Figures 14c and 14d are somewhat similar, the severity of Bias in estimation of SE_b was exacerbated in the *J* of 50, *N* of 4,000, and underlying *a* distributed *LN*(0,.36) conditions (Figure 14d), which was contrary to expectations. A closer inspection of the larger Bias in estimation of SE_b showed that the corresponding *b*s tended to consist of smaller item *a* parameters (i.e., those items circled in Figure 14d) relative to the other *b* parameters. It is also important to point out that a

higher nonconvergence rate occurred in the smaller sample size conditions (Figure 14c). Moreover, if the five items circled in Figure 14d were eliminated, the results showed that a larger sample size (i.e., *I* of 4,000) gave rise to less Bias in estimation of SE_b for *J* of 50 3PL model conditions, as would be expected. However, these five items suspended this general conclusion. For the 3PL model conditions based on *J* of 10, a little positive Bias was found in the estimation of SE_b (Figure 15).

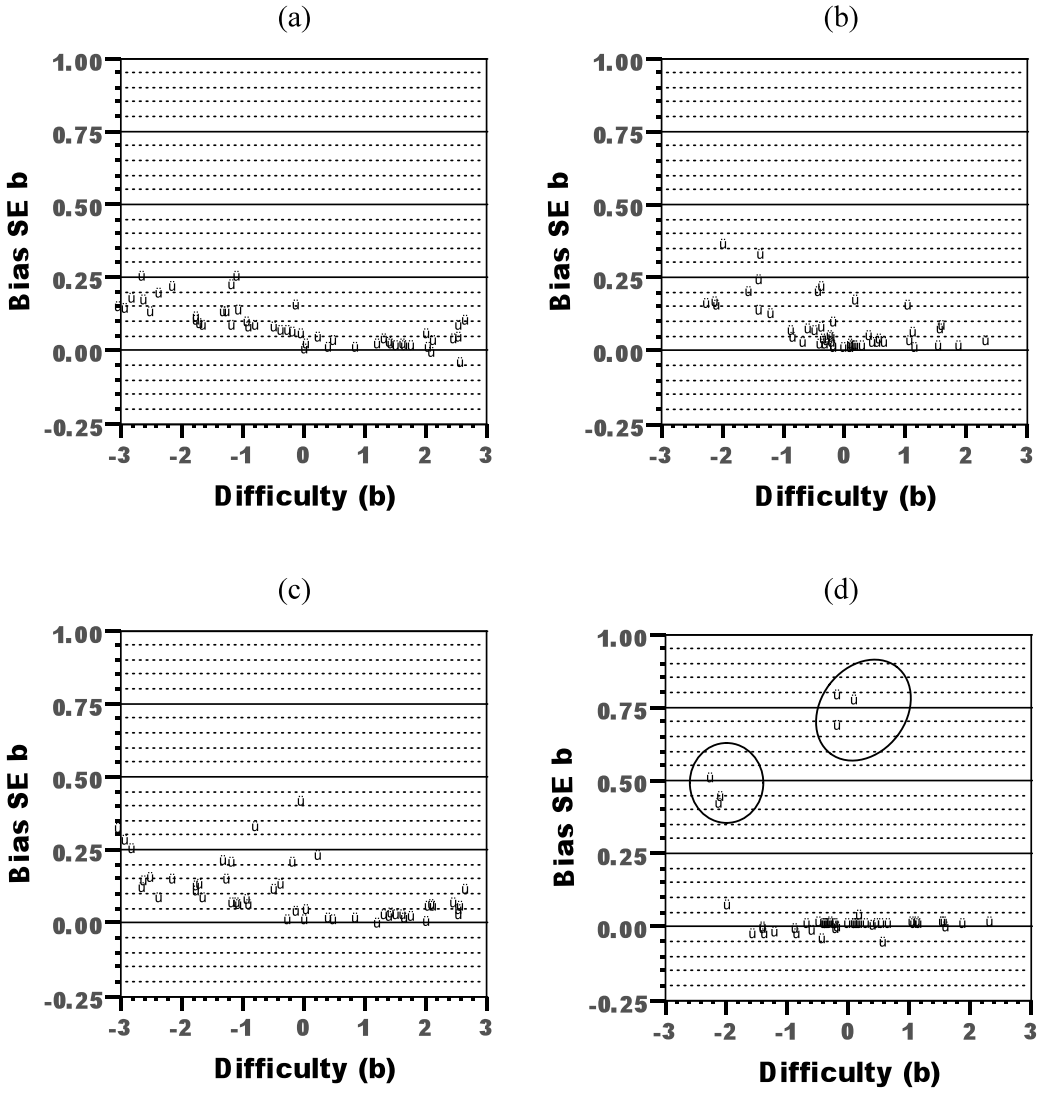

Figure 14. Relationship between the Bias standard error of b and the item difficulty parameter under the 3PL model (J = 50). All plots come from 15 quadrature points. Plots a and c are based on $b \sim U(-3,3)$ and I of 500, while plots b and d are based on $b \sim N(0,1)$ and I of 4,000. Plots a and b are based on $a \sim LN(0,.25)$, while c and d are based on $a \sim LN(0,.36)$. Also, plots a, b, and d are based on $c \sim Beta4(5,17,0,1)$, while plot c is based on $c \sim Beta4(9,33,0,1)$. Moreover, plot b is based on $\theta \sim N(0,1)$, while plots a, c, and d are based on $\theta \sim \chi^2$.

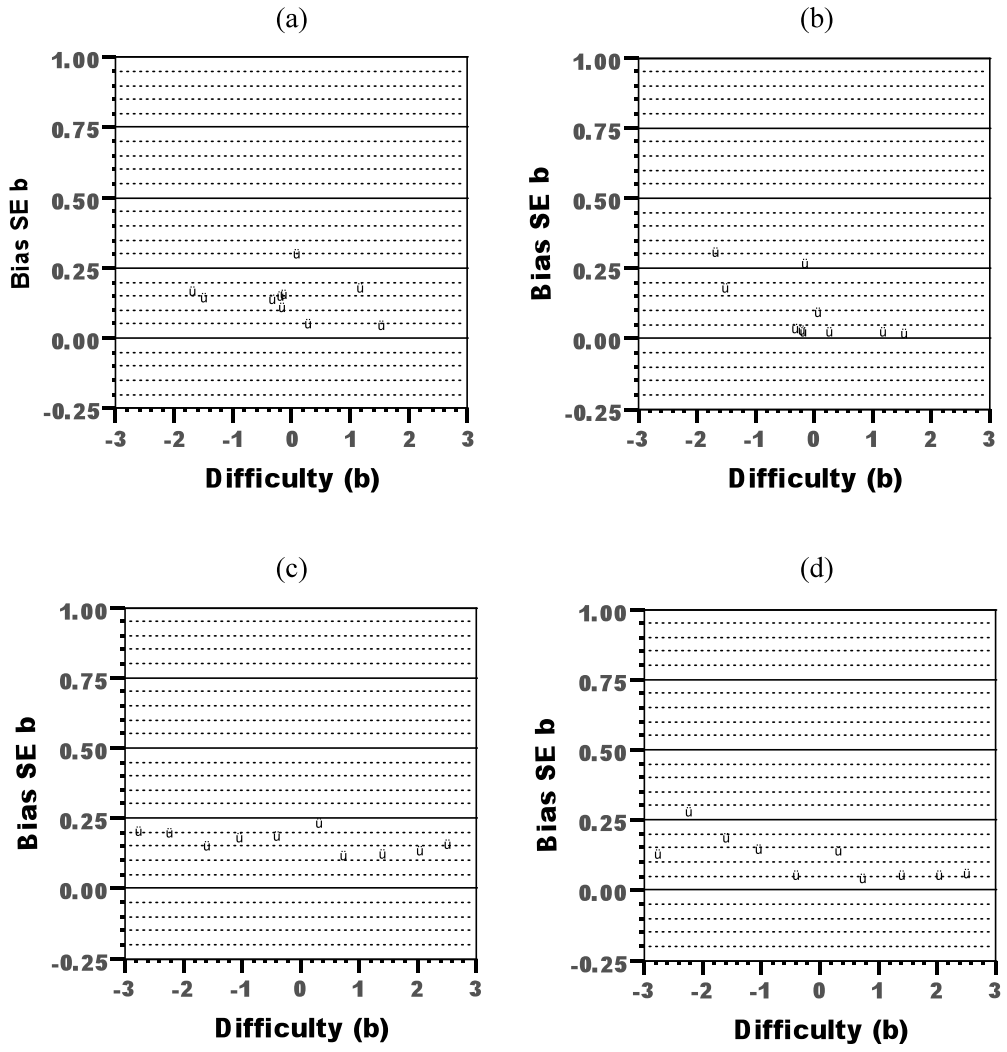

Figure 15. Relationship between the Bias standard error of b and the item difficulty parameter under the 3PL model (J = 10). All plots come from 15 quadrature points. Plots a and b are based on $b \sim N(0,1)$ and I of 500, while plots c and d are based on $b \sim U(-3,3)$ and I of 4,000. Moreover, plot a is based on $a \sim LN(0,.25)$, $c \sim Beta4(5,17,0,1)$, and $\theta \sim N(0,1)$, while plots b, c, and d are based on $a \sim LN(0,.36)$, $c \sim Beta4(9,33,0,1)$, and $\theta \sim \chi^2$.

Bias standard error of discrimination results. The relationship between the Bias SE_a and a parameters for 15 quadrature points under the 50-item and 10-item 2PL model conditions are presented in Figures 16 and 17, respectively. The results for the 60 quadrature points 2PL model conditions can be inferred from these plots because they did not differ from those observed for the 15 quadrature points conditions. Figure 16a is

typical of the trend observed in all *I* of 4,000 conditions except for the conditions based on underlying θ distributed χ^2, *J* of 10, and underlying *b* distributed *U*(-3,3). Figures 16b and 16c show the trends observed for conditions based on *I* of 500, *J* of 50, and underlying *a* distributed *LN*(0,.25) and *LN*(0,.36), respectively. Figure 17 shows the range of trends observed in conditions based on *J* of 10 and *I* of 500. Additionally, the conditions based on *N* of 4,000, underlying θ distributed χ^2, *J* of 10, and underlying *b* distributed *U*(-3,3), fall somewhere between the trends shown in Figures 17c and 17d. In general, Bias of estimation of SE_a tended to be larger for larger *a*s, but improved for smaller *a*s. This trend was most evident in conditions based on *I* of 500 (Figures 16b, 16c, and 16d) and those based on *I* of 4,000, *J* of 10, underlying θ distributed χ^2, and underlying *b* distributed *U*(-3,3) (Figures 17b, 17c, and 17d).

The typical relationship observed between the Bias SE_a and *a* parameters for 15 quadrature points under the 50-item 3PL model conditions are presented in Figures 18 and 19, while Figures 20 and 21 show the same relationships for 15 quadrature points under the 10-item 3PL model conditions. Since the number of quadrature points did not influence the patterns seen under any of the 3PL model conditions, results for the 60 quadrature points conditions can be inferred from these plots. Figures 18a and 18c represent the range of trends observed in all *J* of 50 and *I* of 500 conditions. Figures 18b and 18d are characteristic of the trends observed throughout the *J* of 50 and *I* of 4,000 conditions. It is important to note that Figures 19a and 19b depict slightly different trends than those observed in Figures 18c and 18d. Specifically, Figure 19a represents the trend observed for the condition based on *J* of 50, underlying *b* distributed *N*(0,1), underlying *a*

distributed *LN*(0,.36), underlying *c* distributed *Beta4*(9,33,0,1), underlying θ distributed χ^2, and *I* of 500, while Figure 19b represents the same condition, but with an *I* of 4,000.

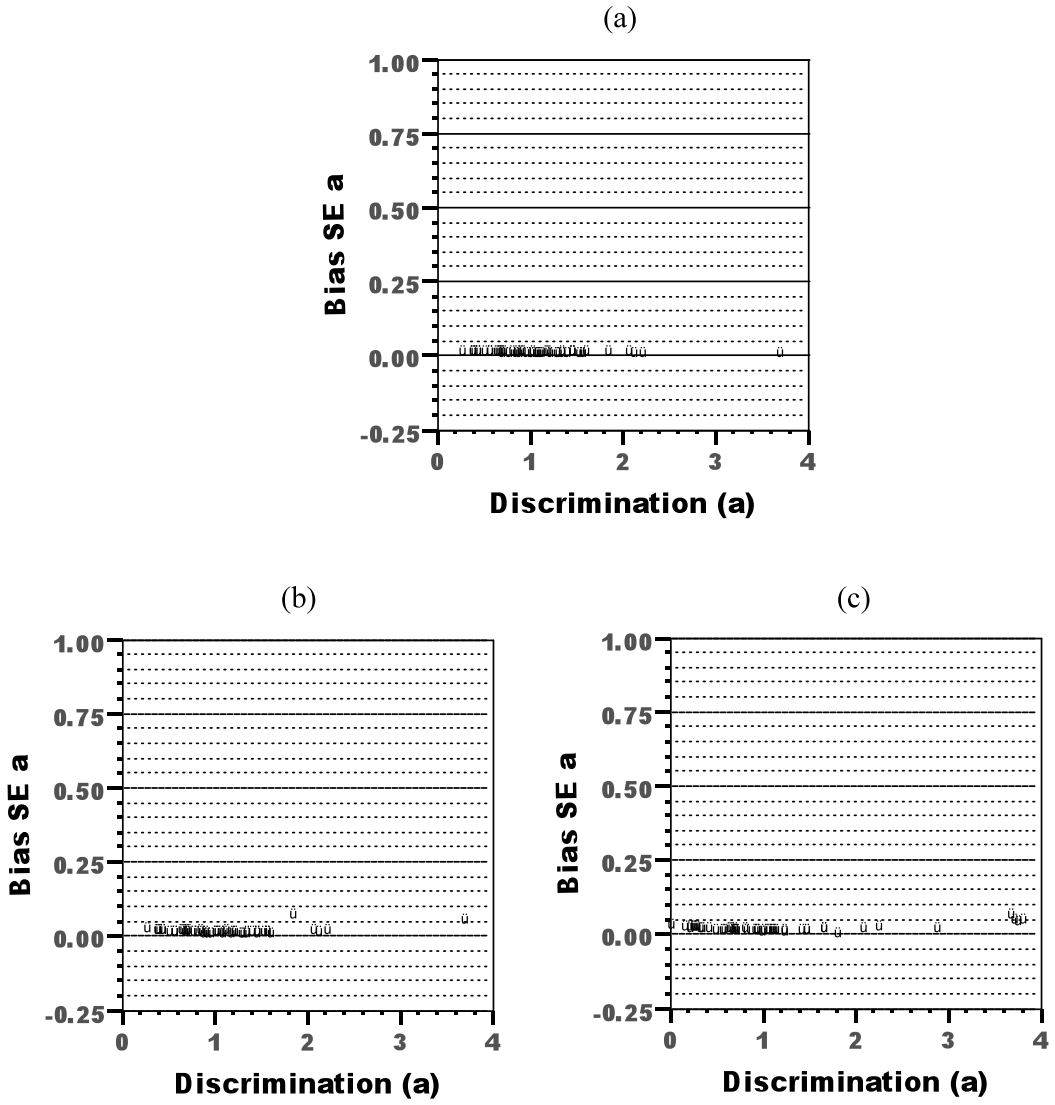

Figure 16. Relationship between the Bias standard error of a and the item discrimination parameter under the 2PL model (J = 50). All plots come from 15 quadrature points and $a \sim LN(0,.25)$. Plots a and b are based on $b \sim N(0,1)$, while plot c is based on $b \sim U(-3,3)$. Moreover, plots a and c are based on $\theta \sim N(0,1)$ and I of 500, while plot b is based on $\theta \sim \chi^2$ and I of 4,000.

Figure 20 represents the range of patterns seen in all conditions based on *J* of 10 and *I* of 500, while Figure 21 depicts the range of trends observed for conditions based on *J* of 10

and I of 4,000. Figures 18 through 21 show estimation of SE_a was a function of item a parameters. This pattern was clearly seen in conditions based on J of 50 and I of 500

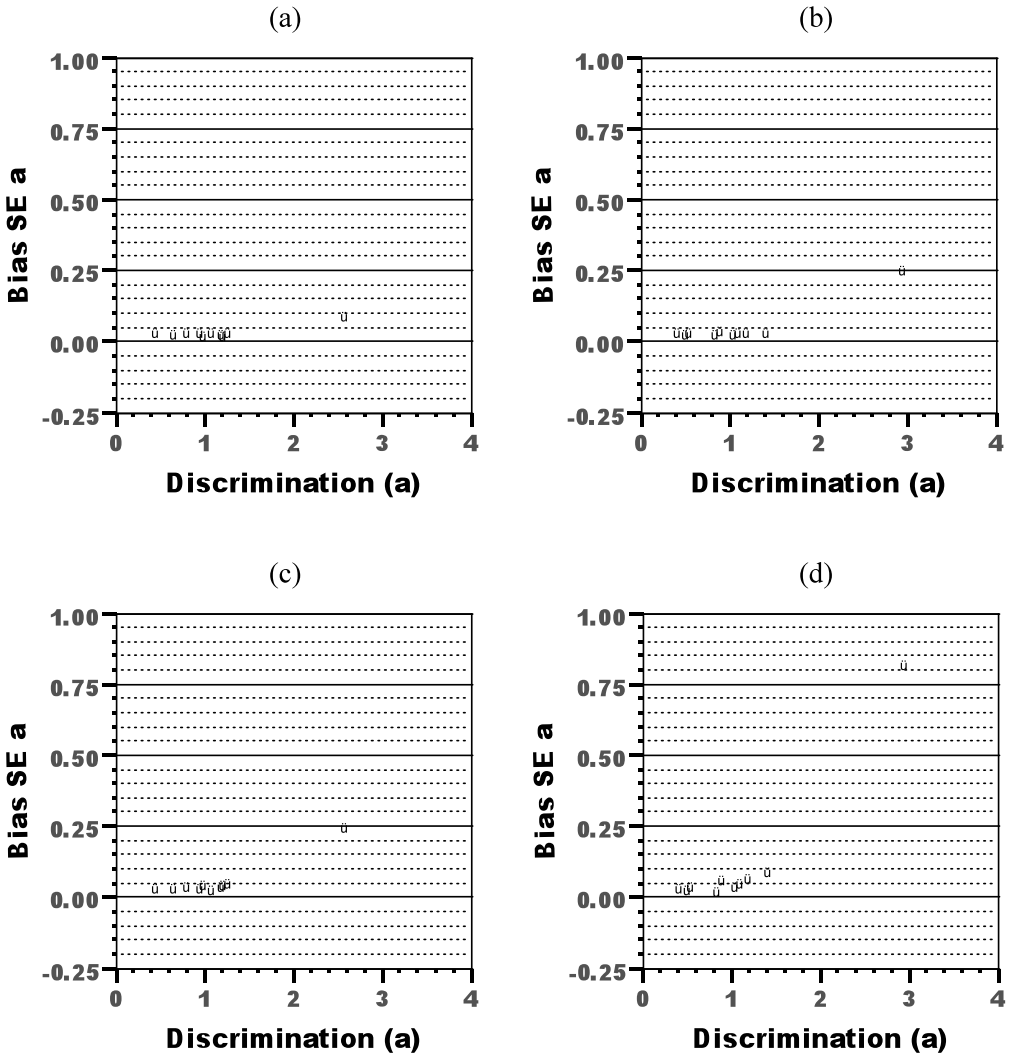

Figure 17. Relationship between the Bias standard error of a and the item discrimination parameter under the 2PL model (J = 10). All plots come from 15 quadrature points and I of 500 Plots a and b are based on $b \sim N(0,1)$, while plots c and d are a based on $b \sim U(-3,3)$. Moreover, plots a and c are based on $a \sim LN(0,.25)$ and $\theta \sim N(0,1)$, while plots b and d are based on $a \sim LN(0,.36)$ and $\theta \sim \chi^2$.

(Figures 18a and 18c). For the conditions based on J of 50 and I of 4,000, (Figures 18b and 18d) this trend became more evident when fewer of the underlying a, c, and θ distributions were similar to the prior distributions used in BILOG-MG 3 and underlying

b was distributed *U*(-3,3) (Figure 18d). This trend was also discernable in the *J* of 10 conditions (Figure 20), but this trend was only detectable because one extreme *a* parameter (i.e., $a > 2.5$) in these conditions had an overestimated SE_a. For all conditions,

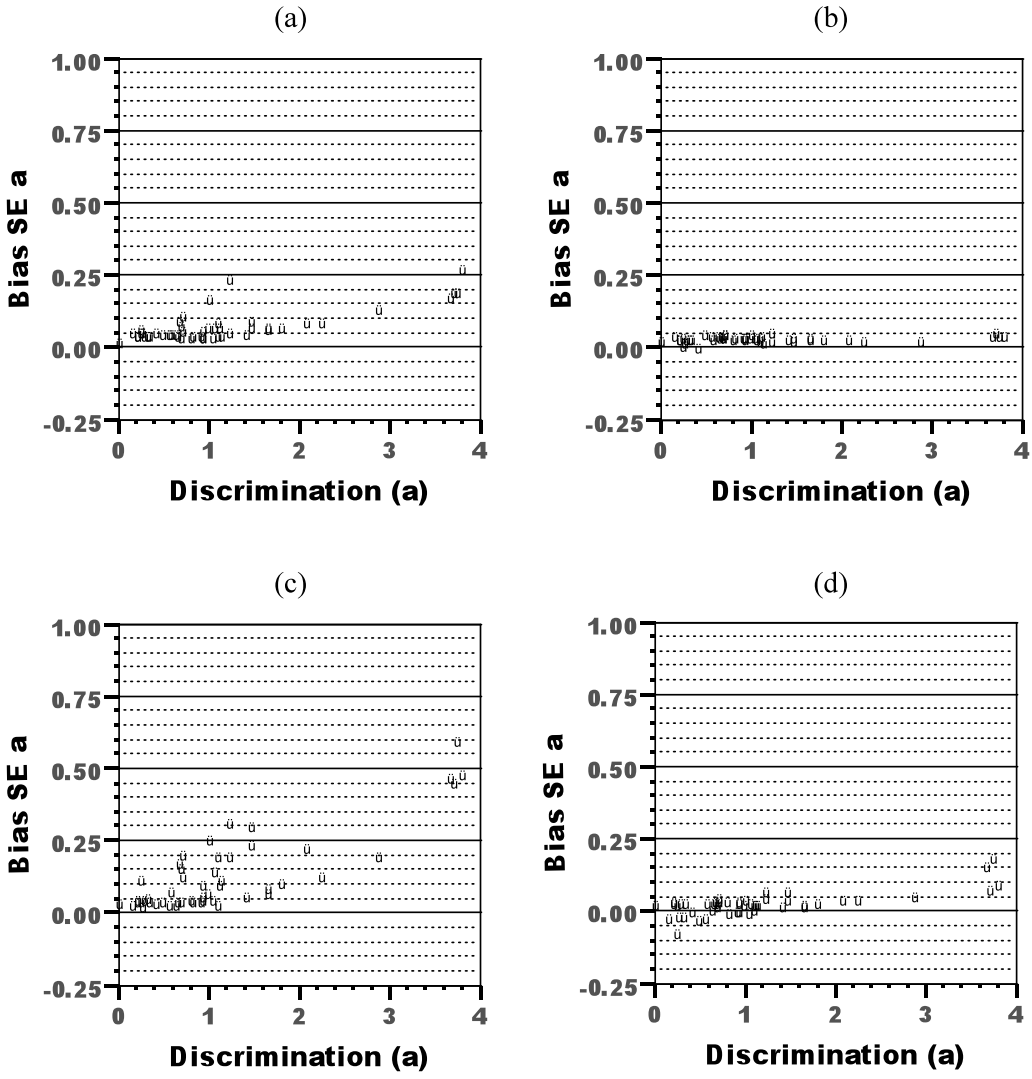

Figure 18. Relationship between the Bias standard error of a and the item discrimination parameter under the 3PL model ($J = 50$). All plots come from 15 quadrature points $a \sim LN(0,.36)$, and $c \sim Beta4(9,33,0,1)$. Plots a and b are based on $b \sim N(0,1)$ and $\theta \sim N(0,1)$, while plots c and d are based on $b \sim U(-3,3)$ and $\theta \sim \chi^2$. Moreover, plots a and c are based on I of 500, while plots b and d are based on I of 4,000.

the accuracy of estimation of SE_a improved as *I* increased (Figures 18a, 18b, 19a, 19b, 20 and 21). In addition, Bias in estimation of SE_a was generally overestimated, but some

conditions based on *I* of 4,000 (Figures 18b, 18d, 21b, and 21d) tended to underestimate the SE_a for some smaller *a*s and overestimate SE_a for some larger *a*s.

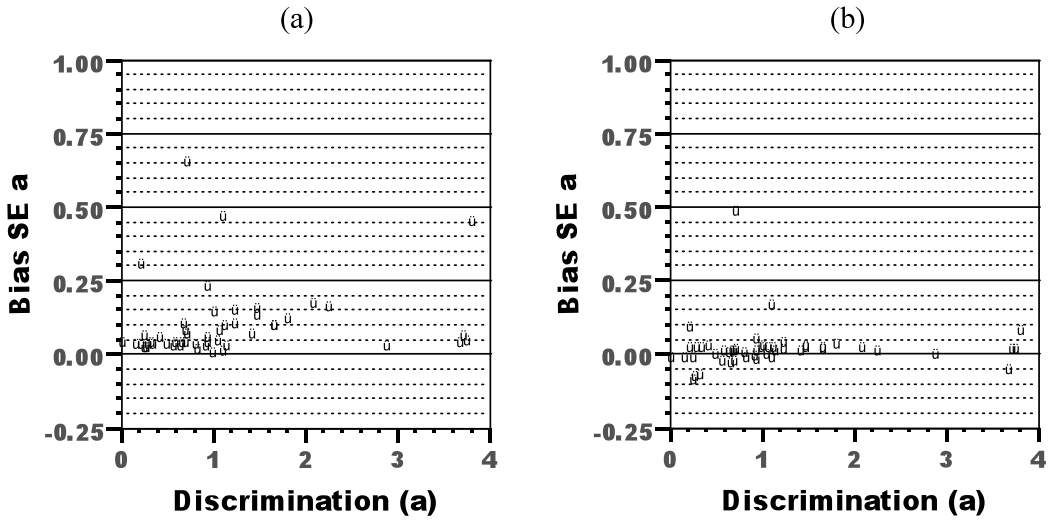

Figure 19. Relationship between the Bias standard error of a and the item discrimination parameter under the 3PL model (more J = 50). Both plots come from 15 quadrature points, $b \sim N(0,1)$, $a \sim LN(0,.36)$, $c \sim B4(9,33,0,1)$, and $\theta \sim \chi^2$. Plot a is based on I of 500, while plot b is based on I of 4,000.

Bias standard error of lower asymptote results. Figures 22a and 22b are characteristic of the relationship between Bias SE_c and item *c* parameters, for all 50- and 10-item 3PL model conditions, respectively. Results for the 60 quadrature points and 3PL model conditions can be inferred from these plots because they did not vary from those found for the 15 quadrature points conditions. Throughout the range of *c* parameters, Bias SE_c was relatively uniform and close to zero, regardless of condition.

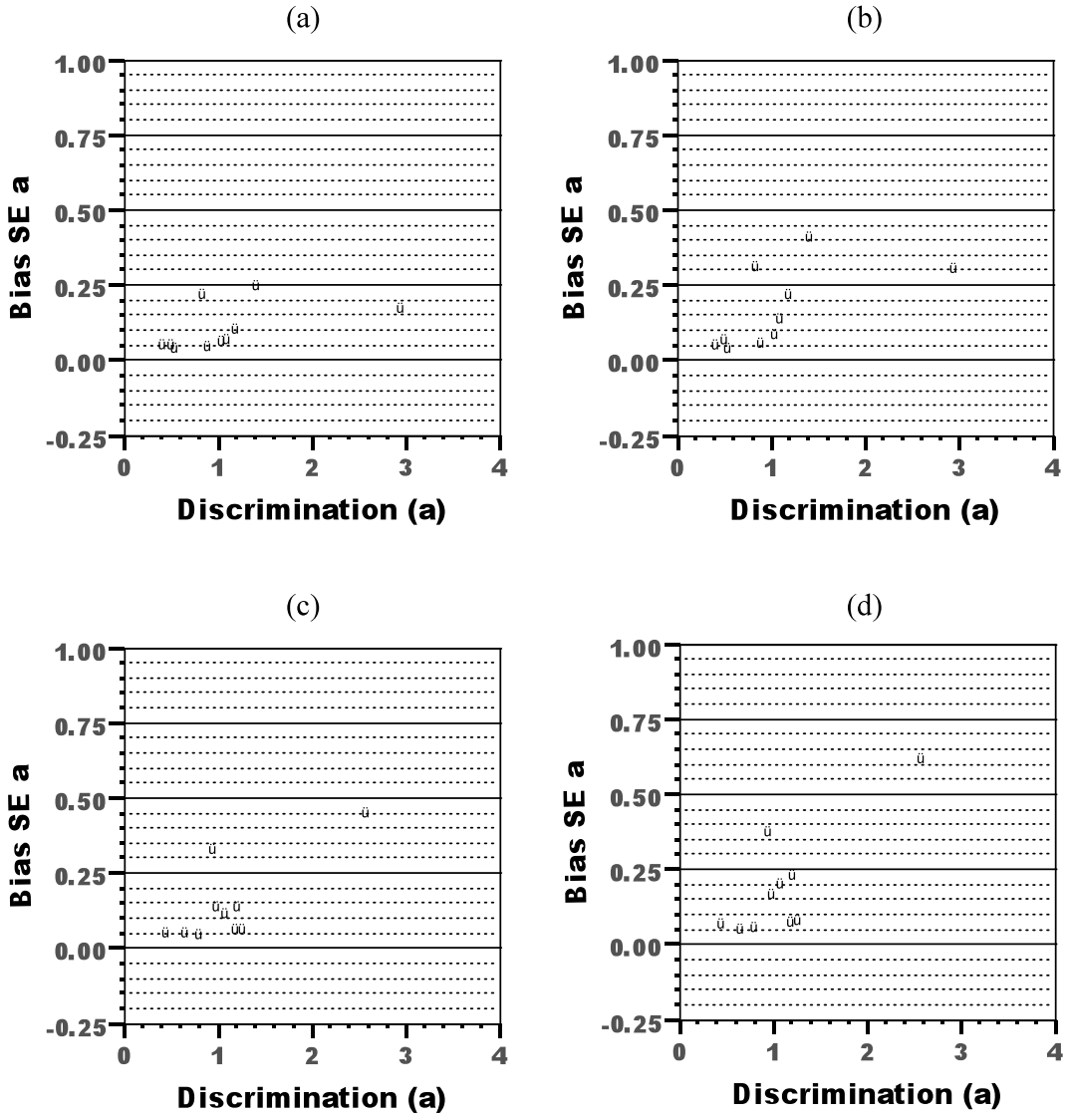

Figure 20. Relationship between the Bias standard error of a and the item discrimination parameter under the 3PL model (J = 10 and I = 500). All plots come from 15 quadrature points. Plots a and b are based on $b \sim N(0,1)$, $a \sim LN(0,.36)$, and c ~ *Beta4* (9,33,0,1), while plots c and d are based on $b \sim U(-3,3)$, $a \sim LN(0,.25)$, and $c \sim Beta4(5,17,0,1)$. Moreover, plots a and c are based on $\theta \sim N(0,1)$, while b and d are based on $\theta \sim \chi^2$.

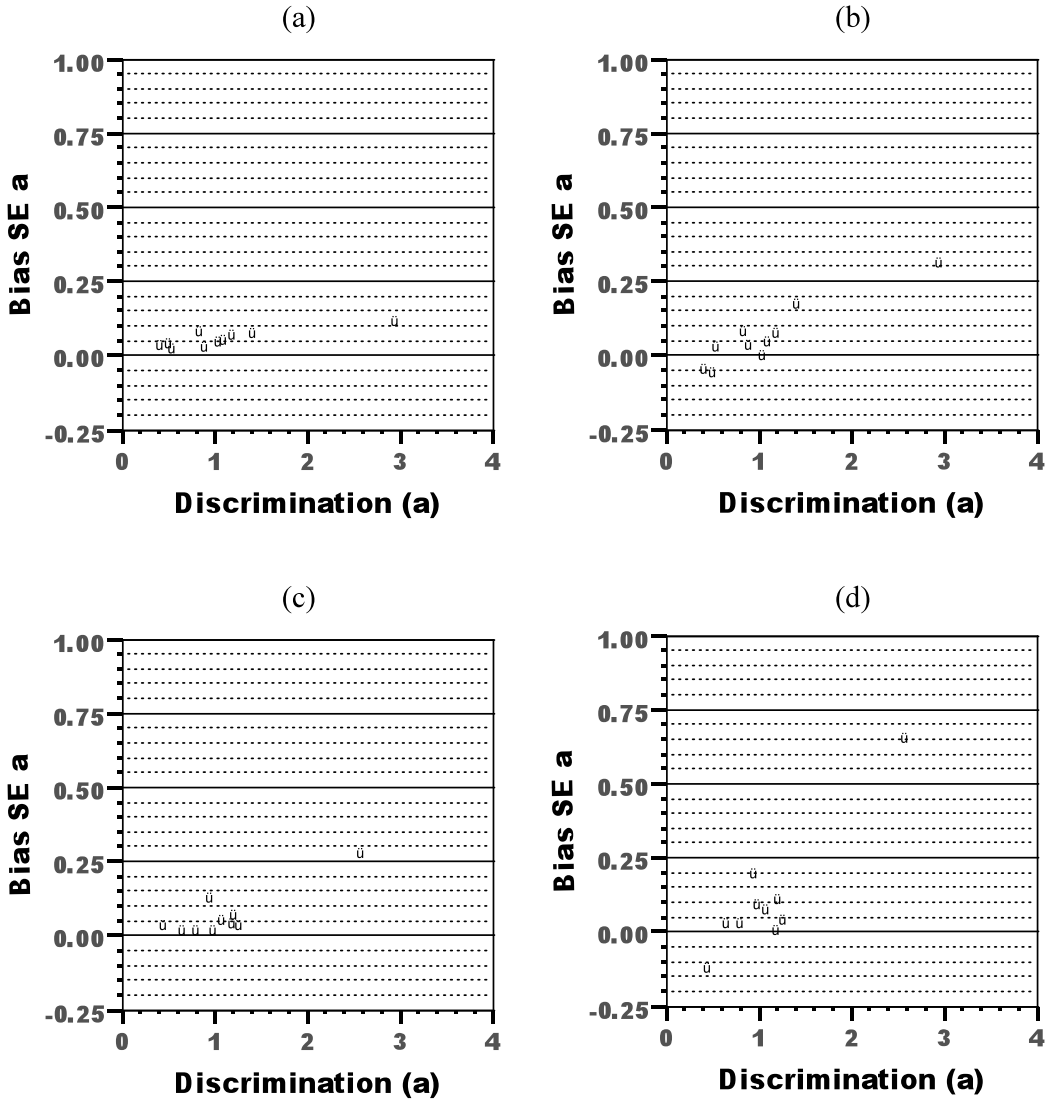

Figure 21. Relationship between the Bias standard error of a and the item discrimination parameter under the 3PL model (J = 10 and I = 4,000). All plots come from 15 quadrature points. Plots a and b are based on $b \sim N(0,1)$, $a \sim LN(0,.36)$, and c ~ *Beta4* (9,33,0,1), while plots c and d are based on $b \sim U(-3,3)$, $a \sim LN(0,.25)$, and $c \sim Beta4(5,17,0,1)$. Moreover, plots a and c are based on $\theta \sim N(0,1)$, while b and d are based on $\theta \sim \chi^2$.

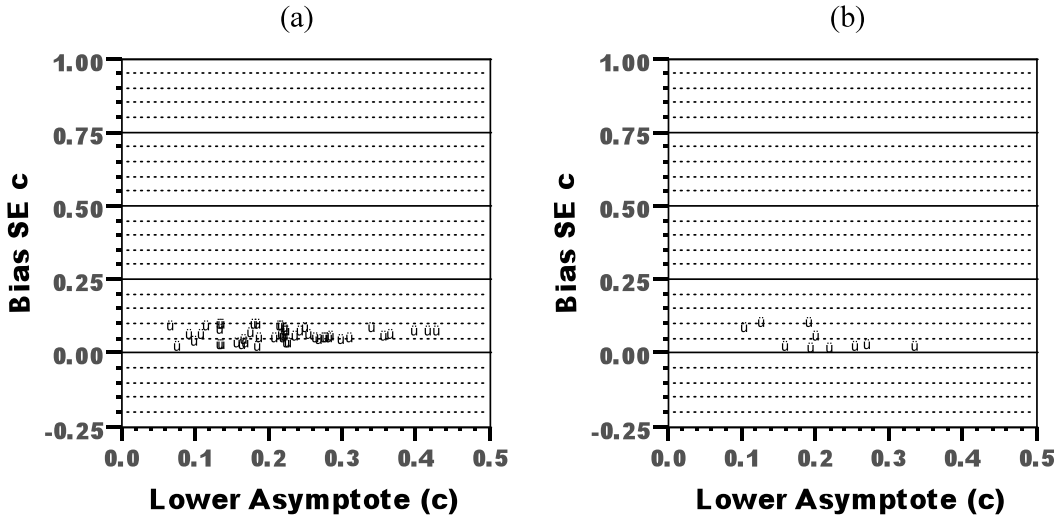

Figure 22. Relationship between the Bias standard error of c and the item lower asymptote parameter under the 3PL model. Both plots come from 15 quadrature points. Plot a is based on J of 50, $b \sim N(0,1)$, $a \sim LN(0,.25)$, $c \sim Beta4(5,17,0,1)$, I of 500, and $\theta \sim N(0,1)$, while plot b is based on J of 10, $b \sim U(-3,3)$, $a \sim LN(0,.36)$, $c \sim Beta4(9,33,0,1)$, I of 4,000, and $\theta \sim \chi^2$.

Chapter Five

Discussion

Currently, BILOG-MG 3 it is one of the most popular IRT programs used for calibrating item parameter estimates from dichotomously scored items. However, little attention has been given to the accuracy of item parameter SEEs produced by the program. The goal of this simulation study was to inform users of BILOG-MG 3 regarding the accuracy of item parameter SEEs produced in the program. Therefore, a Monte Carlo simulation was conducted to allow a direct examination of the accuracy of estimation of the 1PL, 2PL, and 3PL models item parameter SEEs, under a variety of conditions.

To recap, hypothesis one predicted that as sample size increased, item parameter SEEs would be more accurate. Hypothesis two predicted that increasing the number of quadrature points would improve the accuracy of item parameter SEEs. The third hypothesis was that test length would not have an impact on the accuracy of item parameter SEEs. The final hypothesis predicted that item parameter SEEs would be more accurate when more of the underlying item parameter and ability distributions were similar to the prior item parameter and ability distributions specified in BILOG-MG 3, than when the underlying distributions and prior distributions were not similar.

Results from the RMSE and Bias plots showed the accuracy of the estimated SE_b under the 1PL, 2PL, and 3PL models depended on the magnitude of the difficulty parameter being estimated for select conditions. Under the 1PL model, results were not consistent with the first hypothesis because the accuracy of the estimated SE_b was related to I, underlying θ distribution, and underlying b distribution. Specifically, accurate

estimation of SE_b (i.e., both RMSE < .05 and Bias < .1 for all *b*s) was found throughout the range of *b* parameters studied for *I* of 500, underlying *b* distributed *N*(0,1), and underlying θ distributed *N*(0,1) conditions or *I* of 4,000 conditions. For all other 1PL model conditions, accuracy of SE_b tended to decrease for larger *b* parameters. As indicated above, neither increasing the number of quadrature points nor changing the test length had an influence on the accuracy of the estimated SE_b under the 1PL model. Thus, results were consistent with hypothesis two, but were not consistent with hypothesis three. Also, the data was consistent with hypothesis four under the 1PL model because the accuracy of SE_b improved when the underlying θ distribution was similar to the prior θ distribution specified in BILOG-MG 3, but only when *I* was 500. Taken as a whole, the 1PL model results are consistent with those found by Wang and Chen (2005), who examined accuracy of SE_b under the Rasch model by means of WINSTEPS.

Consistent with the first hypothesis, results for the 2PL model showed the accuracy of the estimated SE_b was related to *I*. For *I* of 4,000, consistent estimation of SE_b was found throughout the range of difficulty parameters studied. When *I* was 500, accuracy of SE_b decreased for larger *b* parameters. Consequently, no other variables in this study had an impact on the accuracy of the estimated SE_b under the 2PL model. This means the results were consistent with hypothesis three, but the data was not consistent with either hypothesis two or four.

For the 3PL model, results showed the overall accuracy of the estimated SE_b tended to be impacted by *I*, which is consistent with hypothesis one. Results were not consistent with hypothesis two, in that no gain in accuracy of the estimated SE_b was

found by increasing the number of quadrature points. Consistent with hypothesis three, results did not show a difference in accuracy of the estimated SE_b between test lengths. However, RMSE and Bias SE_b results showed certain combinations of *J*, *I*, underlying *b* distribution, and underlying *a* distribution had consistently uniform accuracy of the estimated SE_b across the range of *b* parameters studied. These conditions were: (1) *J* of 50, underlying *b* distributed *N*(0,1), and underlying *a* distributed *LN*(0,.25); (2) *J* of 50, *I* of 4,000, underlying *b* distributed *U*(-3,3), and underlying *a* distributed *LN*(0,.25); (3) *J* of 10, *I* of 500, and underlying *b* distributed *N*(0,1); and (4) *J* of 10 and *I* of 4,000. Results did show the accuracy of the estimated SE_b improved when the underlying item parameters and ability distributions were similar, but only for *J* of 50 (i.e., compare Figures 4b to 4c and 4d). Thus, results are consistent with hypothesis four. For the remaining *J* of 50 conditions, an inconsistent estimation of SE_b was found throughout the range of *b* parameters studied.

When considering the accuracy of the estimated SE_a, the RMSE and Bias plots under the 2PL and 3PL models showed that the accuracy depended upon the magnitude of the *a* parameter being estimated. For the 2PL model, results showed the accuracy of the estimated SE_a was related to *J*, *I*, underlying θ distribution, underlying *b* distribution, and underlying *a* distribution when the entire range of *a* parameters was considered. It is important to note that when only small *a* parameters were considered (i.e., $a < 1.4$), a small advantage in accuracy of the estimated SE_a was found when using *I* of 4,000 versus *I* of 500. When the full range of item *a* parameters were considered it was found that results were not consistent with hypothesis one, two, or three, but they were consistent with hypothesis four. Although the effect of the above mentioned variables on the

accuracy of SE_a was small, it is still important to discuss. For instance, for J of 10 and I of 500 the accuracy of the estimated SE_a improved throughout the range of a parameters studied as the underlying a and θ distributions became more similar to the prior a and θ distributions identified in BILOG-MG 3. Results also showed consistent and accurate estimates of SE_a throughout the range of a parameters studied for I of 4,000, but this did not hold for combinations of I of 4,000, J of 10, underlying θ distributed χ^2, and underlying b distributed $U(-3,3)$. In these conditions, large RMSE and Bias SE_a values were found for the largest a parameter studied. In the remaining I of 500 conditions, accuracy of the estimated SE_a also tended to diminish for larger a parameters. Moreover, the poorest estimation of SE_a across the range of a parameters occurred for J of 10, I of 500, underlying b distributed $U(-3,3)$, and when the underlying a and θ distributions were different from the prior a and θ distributions used in BILOG-MG 3. These results were consistent with hypothesis four.

When the 3PL model was considered, results showed the accuracy of the estimated SE_a was related to J, I, underlying b, a, and θ distributions. With the exception of two conditions (see Figure 10), RMSE and Bias SE_a data showed that as the magnitude of the a parameter increased, the accuracy of the estimated SE_a consistently decreased. Similar to the 2PL model results, an increase in sample size drove the accuracy of the estimated SE_a under the 3PL model. This is consistent with hypothesis one. However, although small, results showed that when more of the underlying item parameter distributions were similar to the prior item distributions used in BILOG-MG 3, smaller RMSE and Bias SE_a values were found across the range of item a parameters studied.

This finding was not consistent with hypothesis one and two, but was consistent with hypothesis three and four.

The RMSE and Bias plots showed the accuracy of the estimated SE_c under the 3PL model was independent of the magnitude of the item *c* parameter being estimated. Furthermore, results from these plots showed the accuracy of SE_c was consistently estimated across the range of *c* parameters studied for all conditions. Consequently, results were not consistent with hypothesis one, three, and four, but results were consistent with hypothesis three because *J* did not have an effect on the accuracy of the estimated SE_c.

These findings suggested some general conclusions, but they should be interpreted with caution because they assume the underlying item and ability distributions are known to the researchers. One, BILOG-MG 3 produced accurate estimates of SE_b under the 1PL and 2PL models throughout the range of difficulty parameters studied for all conditions studied. This means users can have confidence in the accuracy of SE_b from the 1PL and 2PL models for use in other applications. The problems associated with trying to get accurate estimates of SE_b under all conditions studied for the 3PL model seemed to be challenging. For instance, users of BILOG-MG 3 can get reasonably accurate estimates of SE_b for a 50-item test under the 3PL model when sample size is 4,000 and all item *a* parameter estimates collectively have a distribution similar to that assumed by the default prior *a* distribution in BILOG-MG 3. BILOG-MG 3 produced reasonable estimates of SE_b for a 10-item length test under the 3PL model, but this did not hold when the sample size was 500 and the estimated item *b* parameters followed a uniform distribution restricted to the range (-3,3). In addition, the accuracy of SE_b across

the range of *b* parameters studied seemed to have a greater dispersion than that found under the 1PL and 2PL models. That is, the trends for the 1PL and 2PL models seemed to be smoother, under the 3PL model patterns were difficult to identify. Due to the poor estimation of the SE_b under the 3PL, it is not recommended that they be used beyond descriptive purposes.

A second conclusion that can be drawn from this study is that users of BILOG-MG 3 can get reasonably accurate estimates of SE_a under the 2PL model for smaller item *a* parameters (i.e., $a < 1.4$), but items with larger *a* parameters tended to have poor SE_a estimates under some study conditions. Unfortunately, under the 3PL model, accurate estimates of SE_a throughout the range of *a* parameters studied tended to be limited to 50-item tests calibrated with *I* of 4,000. So, users of the 3PL SE_a should use them with caution.

A third conclusion is that users can use BILOG-MG 3 to get reasonably accurate estimates of SE_c throughout the range of *c* parameters studied under all of the situations examined in this study. However, the tendency was toward a small positive Bias. Generally speaking, a positive Bias in item parameter SEEs was seen across all models.

Given the fixed factor design, generalizations beyond the conditions considered should be made with caution. For instance, this simulation study is limited to BILOG-MG 3. Clearly, one would not generalize findings from this study to those using another IRT estimation program. Also, only a limited number of testing conditions were considered. For example, it is unknown how these item parameter SEE will perform with very long tests and smaller sample sizes. Equally, it is unknown how dimensionality or missing data will impact the accuracy of SEEs in BILOG-MG 3. Future research might

also explore how accurate SEEs from BILOG-MG 3 compare with other programs for dichotomously scored items.

Another limitation to the study was that the default ridge constant of RIDGE = (2, 0.1, 0.01) was changed to RIDGE = (2, 0.01, 0.2) on the BILOG CALIB line. Although this was only done for the 3PL model to offset an initially high level of nonconvergence rates, it is possible that the modification lead to the increase in variability of SE_a and SE_b estimates. Future research should explore this possibility by generating thousands of replications per condition and then removing nonconverged files to arrive at a conclusion. Then, and only then, we could rule out the modification to the ridge constant as an explanation for the inconsistency in accuracy of SE_b and SE_a produced under the 3PL model.

References

Abdel-fattah, A. A. (1994, April). *Comparing BILOG and LOGIST estimates for normal, truncated normal, and beta ability distributions.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Arnould, C. (2006). *Hand functioning in children with cerebral palsy*. Unpublished Dissertation, Université catholique de Louvain, Brussels, Belgium.

Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. *Applied Psychological Measurement, 14*, 139-150.

Baker, F. B. (1998). An investigation of the item parameter recovery of a Gibbs sampling procedure. *Applied Psychological Measurement, 22*, 153-169.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). College Park, MD: Eric Clearing House on Assessment and Evaluation. (ERIC Document Reproduction Service No. ED 458 219).

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker, Inc.

Ban, J-C., Hanson, B. A., Wang, T., Qing, Y., & Harris, D. J. (2001). A comparative study of on-line pretest item - calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 191-212.

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

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443-459.

Carlson, R. D., & Locklin, R. H. (1995). *Item response theory: Comparing BILOG and MicroCAT calibration for a mathematics ability test.* (ERIC Document Reproduction Service No. ED 393 881)

Cohen, A. S., Kim, S., & Subkoviak, M. J. (1991). Influence of prior distributions on detection of DIF. *Journal of Educational Measurement, 28*, 49-59.

DeMars, C. E. (2003). Equating multiple forms of a competency test: An item response theory approach. (ERIC Document Reproduction Service No. ED480126).

DeMars, C. E., (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education, 17*, 265-300.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*, 77-90.

du Toit, M. (Ed.). (2003). *IRT from SSI. BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* Lincolnwood, IL: Scientific Software International.

Educational Testing Services. (1998). *Algebra end-of-course examination report*. Princeton, NJ.

El-Korashy, A-F (1995). Applying the rasch model to the selection of items for a mental ability test. *Educational and Psychological Measurement, 55*(5), 753-763.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20*, 369-377.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 253-262.

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

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

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist, 27*, 353-383.

Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement, 15*, 375-389.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics, 13*, 243-271.

Harwell, M. R. & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279-291.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

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

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*, 146-162.

Li, Y. H., & Lissitz, R. W. (2004). Applications of the analytically derived asymptotic standard errors of item response theory item parameter estimates. *Journal of Educational Measurement, 41*, 85-117.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*, 164-174.

Linacre, J. M. (2001). WINSTEPS Rasch measurement computer program (Version 3.31) [Computer software]. Chicago: Winsteps.com.

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

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*, 187-194.

Mislevey, R. J. (1986). Bayes model estimation in item response models. *Psychometrika, 51*, 177-195.

Mislevy, R. J. & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.

Muraki, E., & Bock, R. D. (1996). PARSCALE (Version 3) [Computer program]. Chicago: Scientific Software International.

Novick, M., & Jackson, P. (1974). *Statistical methods for educational and psychological research*. New York: McGraw Hill.

Obiekwa, J. C. (2001). *An item response theory analysis of Palmore's facts on aging quiz (FAQ) using the three parameter model.* Paper presented at the annual meeting of the Association for Gerontology in Higher Education, San Jose, CA.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*, 1-17.

Parshall, C. G., Kromrey, J. D., & Chason, W. M. (1996, June). *Comparison of alternative models for item parameter estimation with small samples.* Paper presented at the Annual Meeting of the Psychometric Society, Banff, Alberta, Canada.

Parshall, C. G., Kromrey, J. D., Chason, W. M., & Yi, Q. (1997, June). *Evaluation of parameter estimation under modified IRT models and small samples.* Paper presented at the Annual Meeting of the Psychometric Society, Gatlinburg, TN.

Patsula, L. N., & Gessaroli, M. E. (1995, April). *A comparison of item parameters estimates and ICCs produced with TESTGRAF and BILOG under different test lengths and sample sizes*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement, 19*, 353-368.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Richichi, R. (1996, October). An item response theory analysis of multiple-choice items chosen at random from a publisher's test bank. Paper presented at the annual conference of the Northeastern Educational Research Association Conference, Ellenville, NY.

Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for windows. *International Journal of Testing, 3*, 365-384.

Sass, D. A., Schmitt, T. A., & Walker, C. M. (2004, April). *An evaluation of BILOG-MG with skewed theta distributions using various estimation procedures: A simulation study.* Poster presented at the National Council on Measurement in Education, San Diego, California.

Seong, T. -J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299-311.

Smith, R. M. (1996). A comparison of the Rasch separate calibration and between-fit methods of detecting item bias. *Educational and Psychological Measurement, 56*, 403-418.

Smith, R. M., & Suh, K. K. (2003). Rasch fit statistics as a test of item parameter estimates. *Journal of Applied Measurement, 4*, 153-163.

Stone, G. E., & Lunz, M. E. (1994, April). *Item calibration considerations: A comparison of item calibrations on written and computerized adaptive examinations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Swamminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50*, 349-364.

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397-412.

van der Linden, W. J., & Hambleton (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

van der Linden, W. J., & Glas, C. (Eds.). (2000). *Computer-adaptive testing: Theory and practice*. Boston: Kluwer.

Veerkamp, W. J. J., & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics, 25*, 373-389.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., et al. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, W. & Chen, C. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program. *Educational and Psychological Measurement, 65*, 376-404.

Wells, C. S., Subkoviak, & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*, 77-87.

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). *Applied Psychological Measurement, 27*, 299-300.

Wightman, L. E. & De Champlain, A. F. (1994). *A comparison of the properties of IRT parameter estimates using two different calibration designs* (ETS Research Rep. No. 64-19). Princeton, NJ: Educational Testing Service.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton NJ: Educational Testing Service.

Wright, B. D. & Stone, M. (1979). *Best test design*. Chicago: MESA.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG for Windows: Multiple-group IRT analysis and test maintenance for binary items (Version 3.0) [Computer software]. Chicago, IL: Scientific Software International.

Appendix A

Descriptive Statistics of Statistical Distributions Used in Generating Item Parameters and Sampled Parameters

Statistical Generating Distributions

b							*a*							*c*						
Distribution	*M*	*SD*	*Skew*	*Kurt*	*Min*	*Max*	Distribution	*M*	*SD*	*Skew*	*Kurt*	*Min*	*Max*	Distribution	*M*	*SD*	*Skew*	*Kurt*	*Min*	*Max*
N(0,1)	0	1	0	0	−∞	+∞	*LN*(0,.25)	1.133	.604	1.75	5.898	0	+∞	*Beta4* (5,17,0,1)	.227	.087	.52	.15	0	1
U(-3,3)	0	1.73	0	-1.2	-3	3	*LN*(0,.36)	1.197	.788	2.26	10.273	0	+∞	*Beta4* (9,33,0,1)	.214	.063	.415	.119	0	1

Sampled Parameters (J = 50)

b							*a*							*c*						
Distribution	*M*	*SD*	*Skew*	*Kurt*	*Min*	*Max*	Distribution	*M*	*SD*	*Skew*	*Kurt*	*Min*	*Max*	Distribution	*M*	*SD*	*Skew*	*Kurt*	*Min*	*Max*
N(0,1)	-.027	1.044	.005	-.006	-2.184	2.401	*LN*(0,.25)	1.159	.601	1.805	5.952	.321	3.754	*Beta4* (5,17,0,1)	.226	.086	.522	.099	.073	.435
U(-3,3)	0	1.737	-.003	-1.202	-2.965	2.73	*LN*(0,.36)	1.199	.96	1.632	2.249	.064	3.864	*Beta4* (9,33,0,1)	.213	.063	.423	.122	.087	.386

Sampled Parameters (J = 10)

b							*a*							*c*						
Distribution	*M*	*SD*	*Skew*	*Kurt*	*Min*	*Max*	Distribution	*M*	*SD*	*Skew*	*Kurt*	*Min*	*Max*	Distribution	*M*	*SD*	*Skew*	*Kurt*	*Min*	*Max*
N(0,1)	-.017	1	-.048	.009	-1.606	1.613	*LN*(0,.25)	1.157	.577	1.988	5.359	.489	2.626	*Beta4* (5,17,0,1)	.225	.085	.493	.149	.095	.385
U(-3,3)	-.02	1.811	-.018	-1.265	-2.688	2.605	*LN*(0,.36)	1.13	.729	2.093	5.326	.461	2.996	*Beta4* (9,33,0,1)	.212	.069	.398	.138	.111	.342

Note. *U* = uniform; *N* = Normal; *LN* = lognormal; *Beta4* = 4-parameter Beta distribution; *Skew* = Skewness; *Kurt* = Kurtosis, *Min* = Minimum; *Max* = Maximum.

Appendix B

Sampled Item Parameters for Conditions with 50-Item Length Test

	b		*a*		*c*	
Item No.	*N* (0,1)	*U* (-3,3)	*LN* (0,.25)	*LN* (0,.36)	*Beta4* (5,17,0,1)	*Beta4* (9,33,0,1)
1	-2.184	-2.854	0.772	0.338	0.141	0.195
2	-2.030	-2.965	0.735	0.275	0.073	0.175
3	-2.022	-2.743	0.757	0.394	0.192	0.122
4	-1.903	-2.570	0.427	0.857	0.142	0.119
5	-1.479	-2.529	0.688	0.992	0.223	0.152
6	-1.312	-2.431	0.900	0.715	0.221	0.207
7	-1.309	-2.286	0.564	0.064	0.121	0.150
8	-1.281	-2.074	0.482	0.751	0.186	0.154
9	-1.119	-1.688	0.954	1.110	0.229	0.194
10	-0.802	-1.670	1.170	0.974	0.230	0.204
11	-0.763	-1.622	1.406	0.879	0.249	0.209
12	-0.585	-1.568	1.626	1.711	0.260	0.180
13	-0.508	-1.238	1.000	0.616	0.181	0.164
14	-0.397	-1.178	0.972	0.986	0.114	0.255
15	-0.334	-1.090	0.615	0.543	0.256	0.117
16	-0.321	-1.084	1.617	2.932	0.290	0.267
17	-0.301	-1.024	0.488	2.312	0.140	0.247
18	-0.283	-0.983	0.935	1.854	0.228	0.168
19	-0.274	-0.842	1.229	1.463	0.226	0.236
20	-0.225	-0.809	1.662	1.712	0.284	0.224
21	-0.202	-0.706	1.336	0.381	0.193	0.269
22	-0.154	-0.416	1.132	1.156	0.224	0.326
23	-0.131	-0.289	1.352	0.692	0.242	0.185
24	-0.120	-0.169	1.159	3.864	0.227	0.172
25	-0.105	-0.090	1.575	0.305	0.362	0.221
26	-0.103	-0.037	0.830	1.168	0.346	0.087
27	-0.102	0.054	1.405	0.264	0.371	0.386
28	0.063	0.109	3.754	3.768	0.083	0.234
29	0.177	0.135	2.120	1.054	0.231	0.188
30	0.180	0.329	1.453	0.208	0.288	0.261
31	0.192	0.477	2.276	3.724	0.232	0.265
32	0.210	0.567	1.513	2.137	0.317	0.246
33	0.234	0.919	2.175	3.803	0.164	0.197
34	0.254	1.296	0.460	0.327	0.098	0.244
35	0.277	1.391	1.515	1.120	0.305	0.233
36	0.383	1.495	1.262	1.191	0.275	0.308
37	0.483	1.501	0.913	0.645	0.214	0.270
38	0.535	1.594	1.067	0.995	0.281	0.272
39	0.636	1.704	1.061	1.158	0.269	0.184
40	0.645	1.733	0.853	0.767	0.404	0.174
41	0.723	1.853	1.077	1.532	0.173	0.134
42	1.115	2.094	0.431	0.466	0.225	0.338
43	1.149	2.122	0.759	1.529	0.106	0.297
44	1.220	2.175	0.719	1.284	0.423	0.301
45	1.235	2.220	1.518	0.753	0.139	0.242
46	1.619	2.529	1.115	0.738	0.174	0.211
47	1.662	2.730	0.321	0.300	0.171	0.200
48	1.667	2.665	0.629	0.777	0.435	0.190
49	1.954	2.627	1.256	1.070	0.143	0.129
50	2.401	2.620	1.892	1.293	0.191	0.153

Note. *U* = uniform; *N* = Normal; *LN* = lognormal; *Beta4* = 4-parameter Beta distribution;

Appendix C

Sampled Item Parameters for Conditions with 10-Item Length Test

	b		*a*		*c*	
Item No.	*N* (0,1)	*U* (-3,3)	*LN* (0,.25)	*LN* (0,.36)	*Beta4* (5,17,0,1)	*Beta4* (9,33,0,1)
1	-1.606	-2.151	0.701	0.592	0.199	0.199
2	-1.420	-2.688	0.841	0.929	0.160	0.132
3	-0.245	-1.506	1.310	1.228	0.256	0.111
4	-0.112	-0.967	1.225	1.136	0.258	0.208
5	-0.084	-0.322	1.242	1.075	0.095	0.276
6	-0.073	0.389	1.113	0.461	0.231	0.260
7	0.146	0.830	0.489	0.543	0.323	0.225
8	0.360	1.482	2.626	2.996	0.156	0.342
9	1.256	2.129	1.026	1.462	0.385	0.201
10	1.613	2.605	0.997	0.881	0.190	0.166

Note. *U* = uniform; *N* = Normal; *LN* = lognormal; *Beta4* = 4-parameter Beta distribution;

Appendix D

Modified Version of the Whittaker et al. (2003) SAS Macro Program

```
PROC PRINTTO NEW LOG = 'C:\MIKE1to4.LOG';
RUN;
%MACRO DATAGEN(CN=, NE=, SEED=, SEEDUNI=);
%DO REP = 1 %TO 1000;
        %INCLUDE 'C:\Sim\IRTGEN.sas';
        %LET SEED=(&SEED + &REP);
/*Specifies the seed number to be used when generating thetas */
        %LET SEEDUNI=(&SEEDUNI + &REP);
/*Specifies the seed number for ranuniform used to compute 0/1s */
        DATA L3;
                INFILE "C:\Sim\Par\TL50B1A1C1.TXT";
/*change text file before running*/
                        INPUT A B C;
        %IRTGEN(MODEL=L3, DATA=L3, OUT=L3OUT, NI=50, NE=&NE)
        DATA _NULL_;
        SET WORK.L3OUT;
        FILE "C:\Sim\C&CN.\In\C&CN.inR&REP..txt";
/*make sure folders are created*/
        PUT     @1  ID  4.0
                @6  R1  1.0 @7  R2  1.0 @8  R3  1.0 @9  R4  1.0 @10 R5  1.0
                @11 R6  1.0 @12 R7  1.0 @13 R8  1.0 @14 R9  1.0 @15 R10 1.0
                @16 R11 1.0 @17 R12 1.0 @18 R13 1.0 @19 R14 1.0 @20 R15 1.0
                @21 R16 1.0 @22 R17 1.0 @23 R18 1.0 @24 R19 1.0 @25 R20 1.0
                @26 R21 1.0 @27 R22 1.0 @28 R23 1.0 @29 R24 1.0 @30 R25 1.0
                @31 R26 1.0 @32 R27 1.0 @33 R28 1.0 @34 R29 1.0 @35 R30 1.0
                @36 R31 1.0 @37 R32 1.0 @38 R33 1.0 @39 R34 1.0 @40 R35 1.0
                @41 R36 1.0 @42 R37 1.0 @43 R38 1.0 @44 R39 1.0 @45 R40 1.0
                @46 R41 1.0 @47 R42 1.0 @48 R43 1.0 @49 R44 1.0 @50 R45 1.0
                @51 R46 1.0 @52 R47 1.0 @53 R48 1.0 @54 R49 1.0 @55 R50 1.0
                ;
        RUN;
%END;
%MEND DATAGEN;

%DATAGEN(CN=1, NE=500,  SEED=577562, SEEDUNI=1362723)
%DATAGEN(CN=2, NE=4000, SEED=296586, SEEDUNI=1882944)
%DATAGEN(CN=3, NE=500,  SEED=596891, SEEDUNI=1706225)
%DATAGEN(CN=4, NE=4000, SEED=167312, SEEDUNI=1551759)

%LET MAXCAT=2; /* Maximum number of categories for any item */
%LET DIST='NORMAL'; /* Specifies distribution to be used when generating thetas */
```

```
%MACRO IRTGEN(MODEL=, DATA=, OUT=, NI=, NE=);
/***************MacroIRTGEN BEGINS***********/
%MACRO L3GEN;
        EU=EXP(A*(THETA-B));
/* The scaling factor D = 1 was incorporated into the IRT model*/
                P=C+((1-C)*(EU/(1+EU)));
                IF P GE RANUNI(&SEEDUNI) THEN R(J)=1;
                ELSE R(J)=0;
%MEND L3GEN;

%LET FLAG=0;
%IF %LENGTH(&MODEL)=0 %THEN %DO;
        %PUT;
        %PUT *** ERROR ** YOU MUST SPECIFY A MODEL ***;
        %PUT;
        %LET FLAG=1;
%END;
%LET MODEL=%UPCASE(&MODEL);
%IF    &MODEL=PC %THEN %LET MDL=PCGEN;
%ELSE  %IF &MODEL=GPC %THEN %LET MDL=GPCGEN;
%ELSE  %IF &MODEL=GR  %THEN %LET MDL=GRGEN;
%ELSE  %IF &MODEL=RS  %THEN %LET MDL=RSGEN;
%ELSE  %IF &MODEL=SI  %THEN %LET MDL=SIGEN;
%ELSE  %IF &MODEL=L3  %THEN %LET MDL=L3GEN;
%ELSE %DO;
        %PUT;
        %PUT *** ERROR IN MODEL SPECIFICATION: &MODEL ***;
        %PUT;
        %LET FLAG=1;
%END;
        %IF %LENGTH(&NI)=0 OR &NI=0 %THEN %DO;
                %PUT;
                %PUT *** ERROR ** YOU MUST SPECIFY NUMBER OF ITEMS;
                %PUT;
                %LET FLAG=1;
        %END;
%IF %LENGTH(&NE)=0 OR &NE=0 %THEN %DO;
                %PUT;
                %PUT *** ERROR ** YOU MUST SPECIFY NUMBER OF
EXAMINEES ***;
                %PUT;
                %LET FLAG=1;
        %END;
%IF &FLAG=0 %THEN %DO;
        %LET NCATSTR=;
        %IF &MODEL=GR %THEN
```

```
        %LET NCATSTR=%STR(NACT=&MAXCAT-NMISS(OF CB1-
CB&MAXCAT)+1;);
    %IF ((&MODEL=PC)|(&MODEL=GPC)) %THEN
        %LET NCATSTR=%STR(NACT=&MAXCAT-NMISS(OF SD1-
SD&MAXCAT)+1;);
    %IF ((&MODEL=RS)|(&MODEL=SI)) %THEN
        %LET NCATSTR=%STR(NACT=&MAXCAT-NMISS(OF H1-
H&MAXCAT)+1;);
    DATA THETA; *PRODUCES THETAS FOR ALL EXAMINEES;
        KEEP THETA ID;
        CALL STREAMINIT(&SEED);
        DO I=1 TO &NE;
            IF &DIST='UNIFORM' THEN
            THETA=RAND(&DIST)*6-3;
            ELSE THETA=RAND(&DIST);
            ID = 0 + I;
            OUTPUT;
        END;
    RUN;
DATA &OUT;
    KEEP ID THETA R1-R&NI;
    ARRAY PP(*) P1-P&MAXCAT; ARRAY PS(*) PS1-PS&MAXCAT;
    ARRAY DD(*) D1-D&MAXCAT; ARRAY ZZ(*) Z1-Z&MAXCAT;
    ARRAY BB(*) CB1-CB&MAXCAT; ARRAY SP(*) SUMP1-
SUMP&MAXCAT;
    ARRAY SD(*) SD0 SD1-SD&MAXCAT; ARRAY R(*) R1-R&NI; SD0=0;
    ARRAY TH(*) H0 H1-H&MAXCAT; H0=0;
    SET THETA;
    CALL STREAMINIT(&SEEDUNI);
    DO J=1 TO &NI;
        SET &DATA POINT=J;
            &NCATSTR
                %&MDL;
            END;
        RUN;
    %END;

    %MEND IRTGEN;
```

Appendix E

Sample 1PL Model Calibration Command File for BILOG-MG 3

```
>GLOBAL DFNAME = 'C:\Sim\C193\In\C193inR1.txt',
       NPARM = 1,
       LOGISTIC,
       SAVE;
>SAVE PARM = 'C:\Sim\C193\Out\C193outR1.txt';
>LENGTH NITems = (50);
>INPUT NTOtal = 50,
       NALT =    5,
       NIDchar = 4;
>ITEMS ;
>TEST1 TNAme = '',
      INUmber = (1(1)50);
(4A1, 1X, 50A1)
>CALIB NQPt = 15,
      CRIt = 0.01,
      ACCel = 1.0,
      Cycles = 1000,
      Newton = 2,
      Sprior,
      Gprior;
>SCORE ;
```

Appendix F

Sample 2PL Model Calibration Command File for BILOG-MG 3

```
>GLOBAL DFNAME = 'C:\Sim\C1\In\C1inR1.txt',
       NPARM = 3,
       LOGISTIC,
       SAVE;
>SAVE PARM = 'C:\Sim\C1\Out\C1outR1.txt';
>LENGTH NITems = (50);
>INPUT NTOtal = 50,
       NALT =    5,
       NIDchar = 4;
>ITEMS ;
>TEST1 TNAme = '',
      INUmber = (1(1)50);
(4A1, 1X, 50A1)
>CALIB NQPt = 15,
      CRIt = 0.01,
      ACCel = 1.0,
      Cycles = 1000,
      Newton = 2,
      Sprior,
      RIDGE=(2,0.01,0.2),
      Gprior;
>SCORE ;
```

Appendix G

Sample 1PL Model Calibration Command File for BILOG-MG 3

```
>GLOBAL DFNAME = 'C:\Sim\C1\In\C1inR1.txt',
       NPARM = 3,
       LOGISTIC,
       SAVE;
>SAVE PARM = 'C:\Sim\C1\Out\C1outR1.txt';
>LENGTH NITems = (50);
>INPUT NTOtal = 50,
       NALT =    5,
       NIDchar = 4;
>ITEMS ;
>TEST1 TNAme = '',
      INUmber = (1(1)50);
(4A1, 1X, 50A1)
>CALIB NQPt = 15,
      CRIt = 0.01,
      ACCel = 1.0,
      Cycles = 1000,
      Newton = 2,
      Sprior,
      RIDGE=(2,0.01,0.2),
      Gprior;
>SCORE ;
```

Appendix H

Levels of Conditions Manipulated in the Simulation Study

				Number of Quadrature Points							
				15				60			
				I				*I*			
				500		4,000		500		4,000	
				θ Dist.		θ Dist.		θ Dist.		θ Dist.	
J	*b* Dist.	*a* Dist.	*c* Dist.	*N*(0,1)	χ^2(5) w/ *M* of -.5	*N*(0,1)	χ^2(5) w/ *M* of -.5	*N*(0,1)	χ^2(5) w/ *M* of -.5	*N*(0,1)	χ^2(5) w/ *M* of -.5
50	*N*(0,1)	*LN*(0,.25)	*Beta4*(5,17,0,1)	3	3	3	3	3	3	3	3
			Beta4(9,33,0,1)	3	3	3	3	3	3	3	3
			c = 0	2	2	2	2	2	2	2	2
		LN(0,.36)	*Beta4*(5,17,0,1)	3	3	3	3	3	3	3	3
			Beta4(9,33,0,1)	3	3	3	3	3	3	3	3
			c = 0	2	2	2	2	2	2	2	2
		a = 1	*Beta4*(5,17,0,1)								
			Beta4(9,33,0,1)								
			c = 0	1	1	1	1	1	1	1	1
	U(-3,3)	*LN*(0,.25)	*Beta4*(5,17,0,1)	3	3	3	3	3	3	3	3
			Beta4(9,33,0,1)	3	3	3	3	3	3	3	3
			c = 0	2	2	2	2	2	2	2	2
		LN(0,.36)	*Beta4*(5,17,0,1)	3	3	3	3	3	3	3	3
			Beta4(9,33,0,1)	3	3	3	3	3	3	3	3
			c = 0	2	2	2	2	2	2	2	2
		a = 1	*Beta4*(5,17,0,1)								
			Beta4(9,33,0,1)								
			c = 0	1	1	1	1	1	1	1	1
10	*N*(0,1)	*LN*(0,.25)	*Beta4*(5,17,0,1)	3	3	3	3	3	3	3	3
			Beta4(9,33,0,1)	3	3	3	3	3	3	3	3
			c = 0	2	2	2	2	2	2	2	2
		LN(0,.36)	*Beta4*(5,17,0,1)	3	3	3	3	3	3	3	3
			Beta4(9,33,0,1)	3	3	3	3	3	3	3	3
			c = 0	2	2	2	2	2	2	2	2
		a = 1	*Beta4*(5,17,0,1)								
			Beta4(9,33,0,1)								
			c = 0	1	1	1	1	1	1	1	1
	U(-3,3)	*LN*(0,.25)	*Beta4*(5,17,0,1)	3	3	3	3	3	3	3	3
			Beta4(9,33,0,1)	3	3	3	3	3	3	3	3
			c = 0	2	2	2	2	2	2	2	2
		LN(0,.36)	*Beta4*(5,17,0,1)	3	3	3	3	3	3	3	3
			Beta4(9,33,0,1)	3	3	3	3	3	3	3	3
			c = 0	2	2	2	2	2	2	2	2
		a = 1	*Beta4*(5,17,0,1)								
			Beta4(9,33,0,1)								
			c = 0	1	1	1	1	1	1	1	1

Note. *J* = Test Length; *b* Dist. = Underlying Difficulty Distribution; *c* Dist. = Underlying Lower Asymptote Distribution; *a* Dist = Underlying Discrimination Distribution; *U* = uniform; *N* = Normal; *LN* = lognormal; *Beta4* = 4-parameter Beta distribution; θ Dist = Underlying Latent Trait Distribution; χ^2(5) w/ *M* of -.5 = Chi-square Distribution with 5 df standardized to have a *M* = -.5. The number in each cell indicates the generating model for that condition. 3 = 3PL Model; 2 = 2PL Model; 3 = 3PL Model. Cells darkened indicate conditions that did not lend themselves to meaningful IRT models, and were subsequently ignored.

Appendix I

Percentage of Nonconvergence within Condition for the 3PL Model

				Number of Quadrature Points							
				15				60			
				I				*I*			
				500		4,000		500		4,000	
				θ Dist.		θ Dist.		θ Dist.		θ Dist.	
J	*b* Dist.	a Dist.	*c* Dist.	*N*(0,1)	χ^2(5) w/ *M* of -.5	*N*(0,1)	χ^2(5) w/ *M* of -.5	*N*(0,1)	χ^2(5) w/ *M* of -.5	*N*(0,1)	χ^2(5) w/ *M* of -.5
50	*N*(0,1)	*LN*(0,.25)	*Beta4*(5,17,0,1)	2.6	3.0	0.1	0.5	1.4	1.6	0.3	0.8
			Beta4(9,33,0,1)	2.0	1.0		0.5	2.0	0.8		0.3
		LN(0,.36)	*Beta4*(5,17,0,1)	4.1	4.0		0.8	2.6	3.6		1.8
			Beta4(9,33,0,1)	2.2	3.2	0.3	3.3	2.7	2.9	0.1	1.8
	U(-3,3)	*LN*(0,.25)	*Beta4*(5,17,0,1)	5.6	3.3		0.2	5.6	2.7		0.2
			Beta4(9,33,0,1)	6.3	2.8		0.4	6.5	2.1		0.1
		LN(0,.36)	*Beta4*(5,17,0,1)	6.1	3.5	0.1	2.0	5.6	2.7		1.1
			Beta4(9,33,0,1)	6.0	4.8		1.8	7.1	4.7	0.1	1.5
10	*N*(0,1)	*LN*(0,.25)	*Beta4*(5,17,0,1)	0.6	0.5		1.4	0.5	0.5		1.2
			Beta4(9,33,0,1)		0.2		2.8		0.1	0.1	1.7
		LN(0,.36)	*Beta4*(5,17,0,1)	1.3	0.6	0.1	2.1	1.6	0.9	0.4	2.5
			Beta4(9,33,0,1)		0.7	0.1	1.8	0.1	0.4		1.8
	U(-3,3)	*LN*(0,.25)	*Beta4*(5,17,0,1)	0.9	0.7	0.3	3.8	0.6	1.5	0.1	5.4
			Beta4(9,33,0,1)	1.2	0.7	0.4	1.1	1.6	1.7	0.6	1.1
		LN(0,.36)	*Beta4*(5,17,0,1)	1.5	1.4	0.1	6.5	1.5	1.0	0.5	6.7
			Beta4(9,33,0,1)	1.6	1.1	0.5	4.3	2.3	1.1	0.4	4.2

Note. *J* = Test Length; *b* Dist. = Underlying Difficulty Distribution; *c* Dist. = Underlying Lower Asymptote Distribution; *a* Dist = Underlying Discrimination Distribution; *U* = uniform; *N* = Normal; *LN* = lognormal; *Beta4* = 4-parameter Beta distribution; θ Dist = Underlying Latent Trait Distribution; *I* = Sample Size; χ^2(5) w/ *M* of -.5 = Chi-square Distribution with 5 df standardized to have a *M* = -.5. The number in each cell indicates the percentage of nonconverging datafiles within a condition, while cells left blank indicate 0% nonconvergence.

Appendix J

Percentage of Nonconvergence within Condition for the 2PL Model

			Number of Quadrature Points							
			15				60			
			I				*I*			
			500		4,000		500		4,000	
			θ Dist.		θ Dist.		θ Dist.		θ Dist.	
J	*b* Dist.	a Dist.	*N*(0,1)	χ^2(5) w/ *M* of -.5	*N*(0,1)	χ^2(5) w/ *M* of -.5	*N*(0,1)	χ^2(5) w/ *M* of -.5	*N*(0,1)	χ^2(5) w/ *M* of -.5
50	*N*(0,4)	*LN*(0,.25)	0.8	2.3	0.3		0.6	2.4	1.1	
		LN(0,.36)	0.1	0.9				0.1		
	U(-3,3)	*LN*(0,.25)	0.2	4.3				3.7		
		LN(0,.36)	0.1	3.5				3.7		
10	*N*(0,4)	*LN*(0,.25)		0.1			0.1			
		LN(0,.36)	0.6	0.2			0.3			
	U(-3,3)	*LN*(0,.25)	2.4	8.7		3.7	1.3	8.5		3.5
		LN(0,.36)	3.8	13.6	0.6	7.4	4.1	14.2	0.5	5.3

Note. *J* = Test Length; *b* Dist. = Underlying Difficulty Distribution; *a* Dist = Underlying Discrimination Distribution; I = Sample Size; *U* = uniform; *N* = Normal; *LN* = lognormal; *Beta4* = 4-parameter Beta distribution; θ Dist = Underlying Latent Trait Distribution; χ^2(5) w/ *M* of -.5 = Chi-square Distribution with 5 df standardized to have a *M* = -.5. The number in each cell indicates the percentage of nonconverging datafiles within a condition, while cells left blank indicate 0% nonconvergence.

Appendix K

Percentage of Nonconvergence within Condition for the 1PL Model

		Number of Quadrature Points							
		15				60			
		I				*I*			
		500		4,000		500		4,000	
		θ Dist.		θ Dist.		θ Dist.		θ Dist.	
J	*b* Dist.	*N*(0,1)	χ^2(5) w/ *M* of -.5	*N*(0,1)	χ^2(5) w/ *M* of -.5	*N*(0,1)	χ^2(5) w/ *M* of -.5	*N*(0,1)	χ^2(5) w/ *M* of -.5
50	*N*(0,1)								
	U(-3,3)								
10	*N*(0,1)								
	U(-3,3)		1.0				1.1		

Note. *J* = Test Length; *b* Dist. = Underlying Difficulty Distribution; I = Sample Size; *U* = uniform; *N* = Normal; *LN* = lognormal; Beta4 = 4-parameter Beta distribution; θ Dist = Underlying Latent Trait Distribution; χ^2(5) w/ *M* of -.5 = Chi-square Distribution with 5 df standardized to have a *M* = -.5. The number in each cell indicates the percentage of nonconverging datafiles within a condition, while cells left blank indicate 0% nonconvergence.

Appendix L

Summary of Items Omitted by Condition

Summary of Items Omitted by Condition From the Calibration Process

				Number of Quadrature Points			
				15		60	
				θ Dist.		θ Dist.	
J	b Dist.	a Dist.	c Dist.	$N(0,1)$	χ^2	$N(0,1)$	χ^2
50	$N(0,1)$	$LN(0,.36)$	*Beta4* (5,17,0,1)	7(2), 30(1)	7(5)	7(2)	7(6)
			Beta4 (9,33,0,1)	7(5)	29(2)	7(1)	29(4)
			c = 0		**7(1)**	**7(2)**	
	$U(-3,3)$	$LN(0,.36)$	c = 0			**7(1)**	
		$LN(0,.36)$	*Beta4* (5,17,0,1)	7(6), 27(1)	7(3)	7(1)	7(9)
			Beta4 (9,33,0,1)	7(6)	7(2)	7(1), 27(1)	7(1)
10	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)		10(1)		
			Beta4 (9,33,0,1)		10(1)	10(1)	10(1)
		$LN(0,.36)$	*Beta4* (5,17,0,1)			9(1)	10(1)
			Beta4 (9,33,0,1)	10(1)	9(1)		

Note. J = Test Length; b dist = Underlying Difficulty Distribution; c Dist = Underlying Lower Asymptote Distribution; a Dist = Underlying Discrimination Distribution; U = uniform; N = Normal; LN = lognormal; *Beta4* = 4-parameter Beta distribution; θ Dist = Underlying Latent Trait Distribution; χ^2= Chi-square Distribution with 5 df standardized to have a M = -.5. The values in the table represent item number and those in () represent the number of times an item was omitted. All omitted items came from a sample size of 500. Numbers in bold represent items omitted from the 2PL model calibrations, while all other omitted items came from the 3PL model.

Appendix M

Gap Analysis Summary of Items Omitted for the 3PL Model (I = 500)

	Condition						Parameter			Estimate			
Item No.	J	b Dist	a Dist	c dist	NQ	θ Dist.	b	a	c	b	a	c	b_{SEE}
47	50	$N(0,1)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	15	$N(0,1)$	1.662	0.321	0.171	7.90	0.99	0.48	54.88
47	50	$N(0,1)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	1.662	0.321	0.171	6.47	0.99	0.47	14.34
47	50	$N(0,1)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	1.662	0.321	0.171	6.53	0.99	0.48	15.88
47	50	$N(0,1)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	60	$N(0,1)$	1.662	0.321	0.200	6.91	1.00	0.49	22.90
47	50	$N(0,1)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	60	$N(0,1)$	1.662	0.321	0.200	7.79	0.97	0.50	52.18
47	50	$N(0,1)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	60	$N(0,1)$	1.662	0.321	0.200	8.99	0.88	0.50	100.32
47	50	$N(0,1)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	60	$N(0,1)$	1.662	0.321	0.200	6.26	0.99	0.48	13.27
48	50	$N(0,1)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	60	$N(0,1)$	1.667	0.629	0.190	6.46	0.98	0.44	13.73
7	50	$N(0,1)$	$LN(0,.36)$	*Beta4* (5,17,0,1)	15	$N(0,1)$	-1.309	0.064	0.121	6.85	0.90	0.50	16.05
47	50	$N(0,1)$	$LN(0,.36)$	*Beta4* (9,33,0,1)	60	$N(0,1)$	1.662	0.300	0.200	6.73	1.00	0.49	20.48
47	50	$N(0,1)$	$LN(0,.36)$	*Beta4* (9,33,0,1)	60	$N(0,1)$	1.662	0.300	0.200	6.82	0.97	0.50	21.08
50	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	15	$N(0,1)$	2.620	1.892	0.191	7.09	1.00	0.21	15.62
50	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	15	$N(0,1)$	2.620	1.892	0.191	7.21	0.99	0.20	15.47
50	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	15	$N(0,1)$	2.620	1.892	0.191	7.09	0.97	0.20	15.16
42	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	2.094	0.431	0.225	7.74	1.01	0.47	51.10
47	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	2.730	0.321	0.171	7.15	0.99	0.43	22.70
48	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	2.665	0.629	0.435	6.81	1.00	0.49	21.02
50	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	2.620	1.892	0.191	8.29	1.01	0.22	55.73
50	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	2.620	1.892	0.191	7.62	1.00	0.20	26.26
50	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	2.620	1.892	0.191	8.65	1.01	0.20	75.39
50	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	2.620	1.892	0.191	7.03	1.01	0.25	15.11
47	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	15	$N(0,1)$	2.730	0.321	0.200	8.67	0.99	0.41	111.11
50	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	15	$N(0,1)$	2.620	1.892	0.153	7.36	1.00	0.16	16.26
47	50	$U(-3,3)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	60	$N(0,1)$	2.529	1.115	0.211	7.84	1.01	0.25	37.20

	Condition						Parameter			Estimate			
Item No.	J	b Dist	a Dist	c dist	NQ	θ Dist.	b	a	c	b	a	c	b_{SEE}
46	50	U(-3,3)	LN(0,.25)	*Beta4*(9,33,0,1)	60	N(0,1)	2.730	0.321	0.200	8.66	1.01	0.45	134.53
46	50	U(-3,3)	LN(0,.36)	*Beta4*(5,17,0,1)	15	N(0,1)	2.529	0.738	0.174	6.89	0.97	0.29	14.85
47	50	U(-3,3)	LN(0,.36)	*Beta4*(5,17,0,1)	15	N(0,1)	2.730	0.300	0.171	7.05	0.98	0.37	21.16
48	50	U(-3,3)	LN(0,.36)	*Beta4*(5,17,0,1)	15	N(0,1)	2.665	0.777	0.435	6.68	0.98	0.48	16.65
50	50	U(-3,3)	LN(0,.36)	*Beta4*(5,17,0,1)	15	N(0,1)	2.620	1.293	0.191	7.52	0.97	0.23	22.85
47	50	U(-3,3)	LN(0,.36)	*Beta4*(5,17,0,1)	60	N(0,1)	2.730	0.300	0.171	6.76	0.97	0.39	16.22
47	50	U(-3,3)	LN(0,.36)	*Beta4*(5,17,0,1)	60	N(0,1)	2.730	0.300	0.171	6.66	0.96	0.43	16.45
45	50	U(-3,3)	LN(0,.36)	*Beta4*(9,33,0,1)	15	N(0,1)	2.220	0.753	0.242	6.81	0.98	0.39	15.22
47	50	U(-3,3)	LN(0,.36)	*Beta4*(9,33,0,1)	15	N(0,1)	2.730	0.300	0.200	6.98	0.98	0.46	23.23
50	50	U(-3,3)	LN(0,.36)	*Beta4*(9,33,0,1)	15	N(0,1)	2.620	1.293	0.153	9.13	0.98	0.21	112.51
47	50	U(-3,3)	LN(0,.36)	*Beta4*(9,33,0,1)	60	N(0,1)	2.730	0.300	0.200	6.90	0.99	0.44	22.64
47	50	U(-3,3)	LN(0,.36)	*Beta4*(9,33,0,1)	60	N(0,1)	2.730	0.300	0.200	7.72	0.99	0.41	45.74
47	50	U(-3,3)	LN(0,.36)	*Beta4*(9,33,0,1)	60	N(0,1)	2.730	0.300	0.200	8.87	1.00	0.42	154.36
47	50	N(0,1)	LN(0,.25)	*Beta4*(5,17,0,1)	15	$\chi^2(5)$ w/ M = -.5	1.662	0.321	0.171	8.09	1.00	0.45	47.03
47	50	N(0,1)	LN(0,.25)	*Beta4*(9,33,0,1)	15	$\chi^2(5)$ w/ M = -.5	1.662	0.321	0.200	7.28	1.01	0.46	24.54
47	50	N(0,1)	LN(0,.25)	*Beta4*(9,33,0,1)	60	$\chi^2(5)$ w/ M = -.5	1.662	0.321	0.200	6.70	1.01	0.47	14.49
47	50	N(0,1)	LN(0,.36)	*Beta4*(5,17,0,1)	15	$\chi^2(5)$ w/ M = -.5	1.662	0.300	0.171	9.32	0.99	0.44	160.01
50	50	N(0,1)	LN(0,.36)	*Beta4*(9,33,0,1)	15	$\chi^2(5)$ w/ M = -.5	2.401	1.293	0.153	7.68	0.99	0.32	28.41
44	50	U(-3,3)	LN(0,.25)	*Beta4*(5,17,0,1)	15	$\chi^2(5)$ w/ M = -.5	2.175	0.719	0.423	8.02	0.75	0.50	19.00
48	50	U(-3,3)	LN(0,.25)	*Beta4*(5,17,0,1)	15	$\chi^2(5)$ w/ M = -.5	2.665	0.629	0.435	10.63	0.64	0.50	67.17
34	50	U(-3,3)	LN(0,.25)	*Beta4*(9,33,0,1)	60	$\chi^2(5)$ w/ M = -.5	1.296	0.460	0.244	10.33	0.68	0.50	62.80
42	50	U(-3,3)	LN(0,.25)	*Beta4*(9,33,0,1)	60	$\chi^2(5)$ w/ M = -.5	2.094	0.431	0.338	8.18	1.02	0.48	58.85
42	50	U(-3,3)	LN(0,.25)	*Beta4*(9,33,0,1)	60	$\chi^2(5)$ w/ M = -.5	2.094	0.431	0.338	8.92	1.00	0.50	139.82
42	50	U(-3,3)	LN(0,.25)	*Beta4*(9,33,0,1)	60	$\chi^2(5)$ w/ M = -.5	2.094	0.431	0.338	7.09	1.00	0.43	17.40
47	50	U(-3,3)	LN(0,.25)	*Beta4*(9,33,0,1)	60	$\chi^2(5)$ w/ M = -.5	2.730	0.321	0.200	7.08	1.02	0.40	16.60

	Condition						Parameter			Estimate			
Item No.	J	b Dist	a Dist	c dist	NQ	θ Dist.	b	a	c	b	a	c	b_{SEE}
34	50	$U(-3,3)$	$LN(0,.36)$	*Beta4* (5,17,0,1)	15	$\chi^2(5)$ w/ M = -.5	1.296	0.327	0.098	7.15	1.00	0.39	18.34
47	50	$U(-3,3)$	$LN(0,.36)$	*Beta4* (5,17,0,1)	15	$\chi^2(5)$ w/ M = -.5	2.730	0.300	0.171	7.46	0.99	0.39	21.76
50	50	$U(-3,3)$	$LN(0,.36)$	*Beta4* (5,17,0,1)	15	$\chi^2(5)$ w/ M = -.5	2.620	1.293	0.191	7.52	0.97	0.20	16.15
46	50	$U(-3,3)$	$LN(0,.36)$	*Beta4* (5,17,0,1)	60	$\chi^2(5)$ w/ M = -.5	2.529	0.738	0.174	8.96	1.00	0.26	97.05
47	50	$U(-3,3)$	$LN(0,.36)$	*Beta4* (5,17,0,1)	60	$\chi^2(5)$ w/ M = -.5	2.730	0.300	0.171	7.24	1.01	0.39	21.25
47	50	$U(-3,3)$	$LN(0,.36)$	*Beta4* (5,17,0,1)	60	$\chi^2(5)$ w/ M = -.5	2.730	0.300	0.171	7.04	1.00	0.36	14.99
47	50	$U(-3,3)$	$LN(0,.36)$	*Beta4* (5,17,0,1)	60	$\chi^2(5)$ w/ M = -.5	2.730	0.300	0.171	8.86	1.00	0.38	103.69
9	10	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	15	$N(0,1)$	2.129	1.026	0.385	6.19	0.99	0.50	16.81
10	10	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	15	$N(0,1)$	2.605	0.997	0.190	7.37	0.98	0.26	30.99
10	10	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	15	$N(0,1)$	2.605	0.997	0.190	6.63	0.99	0.26	15.41
10	10	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	2.605	0.997	0.190	9.76	1.00	0.24	389.42
10	10	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	2.605	0.997	0.190	7.15	1.00	0.24	25.26
10	10	$U(-3,3)$	$LN(0,.25)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	2.605	0.997	0.190	9.18	1.00	0.25	182.77
10	10	$U(-3,3)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	15	$N(0,1)$	2.605	0.997	0.166	6.81	0.99	0.27	19.14
10	10	$U(-3,3)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	15	$N(0,1)$	2.605	0.997	0.166	7.52	0.98	0.22	35.68
10	10	$U(-3,3)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	15	$N(0,1)$	2.605	0.997	0.166	6.86	0.98	0.27	21.86
10	10	$U(-3,3)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	60	$N(0,1)$	2.605	0.997	0.166	7.31	1.00	0.26	32.64
10	10	$U(-3,3)$	$LN(0,.25)$	*Beta4* (9,33,0,1)	60	$N(0,1)$	2.605	0.997	0.166	6.51	0.99	0.25	14.93
9	10	$U(-3,3)$	$LN(0,.36)$	*Beta4* (5,17,0,1)	15	$N(0,1)$	2.129	1.462	0.385	9.39	0.98	0.42	316.52
10	10	$U(-3,3)$	$LN(0,.36)$	*Beta4* (5,17,0,1)	15	$N(0,1)$	2.605	0.881	0.190	6.66	0.97	0.28	19.28
9	10	$U(-3,3)$	$LN(0,.36)$	*Beta4* (5,17,0,1)	60	$N(0,1)$	2.129	1.462	0.385	7.43	0.99	0.41	52.15
9	10	$U(-3,3)$	$LN(0,.36)$	*Beta4* (9,33,0,1)	15	$N(0,1)$	2.129	1.462	0.201	8.76	0.98	0.28	146.62
10	10	$U(-3,3)$	$LN(0,.36)$	*Beta4* (9,33,0,1)	15	$N(0,1)$	2.605	0.881	0.166	8.14	0.98	0.28	77.27

	Condition						Parameter			Estimate			
Item No.	*J*	*b* Dist	*a* Dist	*c* dist	NQ	θ Dist.	*b*	*a*	*c*	*b*	*a*	*c*	b_{SEE}
8	10	*U*(-3,3)	*LN*(0,.36)	*Beta4*(9,33,0,1)	60	*N*(0,1)	1.482	2.996	0.342	6.29	0.99	0.38	15.50
10	10	*U*(-3,3)	*LN*(0,.25)	*Beta4*(5,17,0,1)	15	χ^2(5) w/ *M* = -.5	2.605	0.997	0.190	8.42	0.99	0.21	69.17
9	10	*U*(-3,3)	*LN*(0,.36)	*Beta4*(5,17,0,1)	15	χ^2(5) w/ *M* = -.5	2.129	1.462	0.385	9.87	0.99	0.40	450.63
10	10	*U*(-3,3)	*LN*(0,.36)	*Beta4*(5,17,0,1)	15	χ^2(5) w/ *M* = -.5	2.605	0.881	0.190	8.12	0.98	0.23	59.52
10	10	*U*(-3,3)	*LN*(0,.36)	*Beta4*(5,17,0,1)	15	χ^2(5) w/ *M* = -.5	2.605	0.881	0.190	6.94	0.98	0.26	20.90
9	10	*U*(-3,3)	*LN*(0,.36)	*Beta4*(5,17,0,1)	60	χ^2(5) w/ *M* = -.5	2.129	1.462	0.385	6.82	0.99	0.42	29.59
9	10	*U*(-3,3)	*LN*(0,.36)	*Beta4*(5,17,0,1)	60	χ^2(5) w/ *M* = -.5	2.129	1.462	0.385	6.19	0.99	0.40	16.16
10	10	*U*(-3,3)	*LN*(0,.36)	*Beta4*(9,33,0,1)	15	χ^2(5) w/ *M* = -.5	2.605	0.881	0.166	8.27	0.98	0.23	75.35
9	10	*U*(-3,3)	*LN*(0,.36)	*Beta4*(9,33,0,1)	60	χ^2(5) w/ *M* = -.5	2.129	1.462	0.201	6.38	0.99	0.23	12.84
10	10	*U*(-3,3)	*LN*(0,.36)	*Beta4*(9,33,0,1)	60	χ^2(5) w/ *M* = -.5	2.605	0.881	0.166	6.47	0.99	0.22	13.43

Note. *J* = Test Length; *b* dist = Underlying Difficulty Distribution; a Dist = Underlying Discrimination Distribution; *c* Dist = Underlying Lower Asymptote Distribution; NQ = Number of Quadrature Points; θ Dist = Underlying Latent Trait Distribution; *U* = uniform; *N* = Normal; *LN* = lognormal; *Beta4* = 4-parameter Beta distribution; χ^2(5) w/ *M* of -.5 = Chi-square Distribution with 5 df standardized to have a *M* = -.5.