The effect of person misfit on item parameter estimation: A simulation study

by

Seyed Amin Mousavi

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in
Measurement, Evaluation, and Cognition

Department of Educational Psychology
University of Alberta

Abstract

The validity and reliability of a test may be compromised because of the presence of misfitting response patterns in test data. Consequently, the purpose of the present study was to investigate the effects of the inclusion and exclusion of misfitting response patterns on item parameter estimates under using simulated data. The four factors considered included test length (20, 40, and 60 items), item parameter estimation method (MLE and Baysian Modal), percentage of students who responded misfittingly in the sample (10%, 20%, and 30%), and percentage of items susceptible to misfitting responses (25% and 50%). Two person fit indices (l_z and H^T) were used to remove misfitting response patterns from the data set with misfitting response patterns. The 2-PL IRT model used to analyze the data. The dependent variables were the bias in the estimated *b* and *a* parameters, the standard error of the estimated parameters, and the classification accuracy of placing students into one of two performance categories. The results showed that 1) there was no difference between the two item parameter estimators, 2) the item difficulty parameter (*b*) was less affected by the presence of misfitting response patterns than the item discrimination parameter (*a*), 3) item parameters with large true *b* and *a* parameter values were affected, 4) an increase in the percentage of misfitting response patterns led to larger bias for both the *b* and *a* parameters, 5) the standard errors of estimates of the *b* and *a* parameters were small across all conditions, 6) the classification accuracy was higher for the low performing students for the fitting data set and lower and essentially constant across the data sets with all the misfitting response patterns, and the data sets with misfitting response patterns removed by l_z and by H^T, 7) the classification accuracy was lower for the high performing students than the low performing students for the fitting data set and lower and essentially constant for the remaining three data sets. Implications for practice and recommendations for future research are provided.

Acknowledgement

This study could not have been completed without the help and support I received from many others. First of all, I would like to express my gratitude to my supervisor Dr. Ying Cui for her continuous support and enthusiasm. With her expertise and breadth of knowledge in the area of Educational measurement, I could improve my research skills and extend my knowledge. I would also like to express my appreciation to my supervisory committee Dr. Todd Rogers and Dr. Michael Carbonaro. I truly believe that without the immense help and support that I received from Dr. Rogers I wouldn't be able to attain such achievement. Your advice on both research as well as on my career have been priceless.

I am heartily thankful to my beloved wife Neda Moslemi, who has helped me to go through many difficult times, and for all the emotional support, entertainment, and caring provided. And finally, my son and bundle of joy of my life, Benyamin, who brought a world of happiness and bliss, with his small feet and hands, into our life.

Table of Contents

List of Tables

List of Figures

Chapter One: Introduction

The use of large-scale tests in educational, psychological and decision-making contexts has become part of the ongoing activities of many school districts, provinces/states, and countries. Often important decisions regarding accountability and placement of students in performance categories (e.g., below basic, basic, proficient, excellent) are made on the basis of test scores generated from tests. Therefore, it is important to evaluate the validity of the inferences derived from test results, which depends on the measurement model used in the design, construction of items, scoring of the students' responses, and analyses of the scored responses. When the measurement model fails to reflect accurately the real aspects of student responses, the validity of test scores may be compromised.

One example of this failure can be found when unusual or unexpected response pattern are produced by some, but not all, students. For example, if some students produce correct answers to the more difficult items but fail to answer the easier items successfully, the students' responses are considered as "unexpected", "aberrant", "unpredictable", or "misfitting" (Meijer & Sijtsma, 2001). "Misfitting" refers to the mismatch between the observed response patterns and the expected response patterns of students derived from a given measurement model. Meijer (1997) and Schmitt, Cortina, and Whitney (1993) suggested that validity and reliability of a test might be compromised because of the existence of misfitting responses in test data. This is mainly due to the effect of the presence of misfitting response patterns on the estimation of student's ability. The existence of misfitting response patterns can distort the shape of likelihood function and result in incorrect ability estimates.

An Overview on Person Fit Statistics

To assess the fit of a student's response pattern to the measurement model a "person fit statistic" (PFS) is used. Generally, PFSs classify students into two groups: students with fitting response vectors and students with misfitting response vectors. One advantage of PFSs is that they analyze response patterns of each individual student tested. Analyzing specific patterns of item responses can disclose more information than simply analyzing test scores at the group level because students may differ in response strategies that they use for answering items on a test. There are other advantages for assessing person fit in testing situations. For example, Emons, Meijer, and Sijtsma (2001) suggested that misfitting responses may serve as an indication that the student's response behavior may have been influenced by factors that are not intended to be measured by the test.

Many person fit statistics have been proposed to help identify observed response patterns that are incongruent with the measurement model used in the test design and analysis. For example, Meijer and Sijtsma (2001) reviewed over 40 person fit statistics and Karabatsos (2003) conducted a simulation study that compared the performances of 36 person fit statistics under different testing conditions.

The different PFSs can be classified into are two main approaches: group-based and IRT-based person fit statistics. In the first approach, PFSs are computed irrespective of a particular measurement model and use observed response patterns without considering, for example, item parameters. A group-based person fit statistic classifies an observed response pattern as misfitting if easy items are answered incorrectly and hard items are answered correctly (Meijer & Sijtsma, 2001). That is, if a student's number-correct score is *r*, the student is expected to have answered the first *r* easiest items correctly. A response vector is considered as misfitting when items with relatively low proportion of correct scores (i.e., the percentage of students who answer the item correctly is low) are answered correctly but items with relatively high proportion of correct scores (i.e., the percentage of students who answer the item correctly is high) are answered incorrectly. Examples of group dependent person-fit statistics are Harnisch and Linn's (1981) modified caution index *C*, van der Flier's (1982) *U3* index, Tatsuoka and Tatsuoka's (1983) norm conformity index *NCI*, and Sijtsma's (1986) H^T coefficient .

In the second approach, PFSs assess the fit of a response pattern relative to a given IRT model such as the three parameter logistic (3-PL) model. Model based PFSs use estimated item and ability parameters for calculating person fit indices and then classifying responses as misfitting or fitting. IRT-based person-fit statistics are specifically designed to evaluate the misfit of an observed response pattern with an IRT model by calculating the response probabilities associated with a student's ability parameter and item parameters. If according to the IRT model the probability of a correct response from a student is high, the hypothesis is posited that the student should answer that item correctly, and vice versa. A misfit is found when the hypothesis is not supported by the observed data. Examples of IRT-based person-fit statistics include Wright and Stone's (1979) *U* statistic, Wright and Masters's *W* statistic (1982), Smith's *UB* and *UW* statistics (1985), and Drasgow, Levine, and Williams' l_z statistic (1985).

Gerald and Lawrence (1992) suggested what while person fit statistics are indices with high potential, there are also many questions to be answered, such as which person fit index should be used and under what circumstances or what should be done for the students with misfitting response patterns. As pointed out more recently by Tendeiro and Meijer (2014) and Rupp (2013), the most important limitation of person fit statistics is that it is hard, if not impossible, to determine the type of misfitting response behavior that underlies the misfitting responses. There's no mechanism to identity whether an observed misfitting pattern, for example, is due to carelessness or random responding (Rupp, 2013). Many individual, classroom and school characteristics can contribute to misfitting patterns. Additionally, there is no general rule or set of guidelines on how to deal with a misfitting response pattern in practical situations (Tendeiro & Meijer, 2014). There are some suggestions, but they have not been used in practice (Smith, 1985; Rupp, 2013).

Research on Person Fit Statistics

To date, the majority of the person fit literature has been heavily focused on creating and evaluating new indices (Rupp, 2013). Such efforts led to more than 40 person fit indices. Some person fit indices may be used under specific conditions (e.g., only for the Rasch model); some are sensitive to specific types of misfitting response patterns (e.g., local vs. global); and some are designed only for dichotomous items (Meijer & Sijtsma, 2001). Furthermore, the effect of person misfit has been studied on ability estimation, equating, and classification accuracy (Nering, 1998; Hendrawan, Glas, & Meijer, 2005; Sotaridona, Choi, & Meijer, 2005). In addition, there are some studies that examined factors which may contribute to person misfit. The latter studies try to provide explanations for why and how person misfit occurs (e.g., Petridou & Williams, 2007).

A few studies have been conducted to investigate the effect of misfitting response patterns in a data set on the estimates of the item parameters and the findings have been mixed. The studies conducted by Levine and Drasgow (1982), Philips (1986), Hendrawan, Glas, and Meijerl (2005), and Sotaridona, Choi, and Meijer (2005) are the only studies in the literature that (directly or indirectly) considered the effect of misfitting response patterns on item parameters. These studies provided limited information on how misfitting response patterns affect item parameters. For instance, Sotaridona et al (2005), who conducted the study most related to the focus of the present study, only used a sample of 10,000 randomly selected examinees from a previously administered assessment program in the United States. They manipulated selected response vectors in order to create copying and guessing respondents. While Hendrawan et al.'s (2005) study was more comprehensive, the focus of their study was on the effect of person misfit on classification decisions. They considered ability and item parameter estimation method, test length, different item parameters, sample size, and two types of misfitting response patterns with 100 replications. A small portion of the study was dedicated to the effect of person misfit on estimated parameters, but without a thorough explanation about how the chosen factors affected estimated item parameters. If misfitting responses patterns do affect the estimates of the item parameters, then the scores of all the students could be comprised. Consequently, the scores may not be validly interpreted.

Purpose of Study

Therefore, the purpose of the present study was to investigate the effect of inclusion and exclusion of misfitting response patterns on the *a* and *b* item parameters in the 2-PL IRT model. Four factors – test length, item parameter estimation method, percentage of misfit in the sample, and percentage of item susceptible of misfitting responding – were considered.

The purpose was addressed using simulated data. Four data sets were created: a data set with no misfitting response patterns, a data set with all the misfitting response patterns, a data set with response patterns removed by l_z, and a data set with misfitting response patterns removed by H^T. The choice of l_z was based on its popularity among researchers and practitioners due to the fact that it is one of the most studied person fit statistics and, because l_z was based on the likelihood function, its calculation is easy and straightforward. The H^T was selected for this study because it was one of the most referred used group-based statistics in previous research (e.g., Armstrong & Shi, 2009; Dimitrov & Smith, 2006; Emons et al., 2002, 2003, 2005; Karabatsos, 2003; St-Onge et al., 2011; Zhang & Wlaker, 2008). Karabatsos (2003) also reported that H^T was the best functioning person fit indices across 36 group-based and IRT-based indices.

Delimitations of the Study

The study is delimited in two ways. First, only dichotomously scored items were considered. Second, no attempt was made to investigate why misfitting response behavior occurred.

Organization of the Dissertation

The dissertation is organized in seven chapters. Chapter One contained a synopsis of the literature, identification of the problem to be investigated, rationale for the study, and the delimitations of the study. The literature review, presented in Chapter Two, starts with an illustration on how misfitting response patterns affect the likelihood function. Then, the selected person fit indices are described in detail. Last, the most recent research on person fit statistics and research that particularly examined the influence of misfitting response patterns on the estimation of item parameters are reviewed. The simulation study design, data simulation procedures and evaluation criteria are provided in Chapter Three. The results of analysis are reported and discussed in Chapters Four, Five, and Six. Finally, a summary of the purpose, method, and results is provided, discussion of the results, limitations of the study, the conclusion drawn in light of the limitations of the study is presented followed by implications for practice and recommendations for future research are presented in Chapter Seven.

Chapter Two: Literature Review

Chapter Two provides a framework for this study by reviewing relevant concepts and previous research. The chapter is organized in three main sections. The first section is devoted to an overview of IRT and the effect of misfitting response patterns on the likelihood function. This section provides a basis for a better understanding of person misfit and the need for person fit assessment. The second section consists of a review of the common person fit statistics in the literature and the two selected person fit indices that were used in this study are presented in detail. The third section contains two sub-sections. The first sub-section includes a review of the studies on different aspects related to person fit research and the second sub-section contains a review of studies on the influence of misfitting response pattern on item parameters. Finally, a summary of the chapter is provided at the end of chapter.

Effect of Misfitting Response Patterns on the Likelihood Function

Since many person fit indices were developed under the framework of Item Response Theory (IRT), it is useful to first provide a brief review of IRT. IRT refers to a family of mathematical functions that connect the probability of a correct answer of an item to the properties of the item such as item difficulty, item discrimination, and/or guessing and to the student's ability (usually denoted as θ). Item responses can be dichotomously scored (e.g., true/false or correct/incorrect), polytomously scored (e.g., Likert-scale), or continuously scored (Embretson & Reise, 2000).

The most complete model for dichotomously scored items is the 3-parameter logistic (3-PL) logistic model in which item difficulty (denoted as *b*), item discrimination (denoted as *a*), and item pseudo-guessing level (denoted as *c*) are used to calculate probability of a correct answer to an item given ability level θ. The 3PL model can be formulated as:

$$P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta_i - b_j)}} \ , \qquad (1)$$

where $P_j(\theta_i)$ is the probability of a correct answer to the jth item by the ith student. There are also a 2-PL IRT model, which assumes that there is no guessing and c is set to zero in equation (1), and a 1-PL IRT model, which assumes all items have the same discrimination level and the c parameter equals zero (de Ayala, 2008). Each of these models is based on the assumption that the test is unidimensional.

The main advantage of IRT, as opposed to classical test theory, is the focus on item level analysis instead of test level analysis. That is, models in classical test theory have connected the total test scores to true scores instead of item scores to true scores (Hambleton & Swaminathan, 1985). In this approach, the pattern of item responses is considered as a source of information when gauging a student's performance on a test. This pattern is used for estimating a student's ability parameter by finding the maximum of likelihood function that is computed based on the student's item responses and the IRT model.

Misfitting response patterns may lead to an overestimate or underestimate of student's ability regardless of the kind of educational or psychological test (Meijer & Sijtsma, 2001). The effect of misfit on ability can be illustrated by its effect on the likelihood function. In item response theory (IRT), estimation of ability measured by a test can be achieved by maximizing the likelihood function for given model and observed response pattern. The likelihood function for the ith student can be computed using the formula,

$$L_i = \prod_{j=1}^{J} P_j(\theta_i)^{X_{ij}} \left[1 - P_j(\theta_i)\right]^{1-X_{ij}} \ , \qquad (2)$$

where X_{ij} is the binary (0, 1) response to item j $(j = 1, 2, \dots, J)$ by student i, θ_i is the latent trait or ability for student i, and $P_j(\theta_i)$ is the probability of a correct answer to item j by persons i

computed based on an IRT model. The maximum likelihood estimate (MLE) of θ, $\hat{\theta}$, occurs at the maximum of likelihood function where the first derivative of likelihood function equals zero.

The following example shows how a response pattern contributes to ability estimation. In this example, two different students take a ten-item test in which items are sorted in ascending difficulty order. Item parameters and response patterns are shown in Table 1.

Table 1
Item parameters and response patterns for imaginary example

	Items									
	1	2	3	4	5	6	7	8	9	10
Discrimination	1.67	1.00	1.14	1.34	1.27	1.5	1.87	1.15	1.00	1.8
Difficulty	-2.00	-1.59	-0.85	-0.10	0.00	0.5	1.2	1.9	2.2	2.5
Guessing	0.01	0.20	0.15	0.15	0.10	0.25	0.20	0.11	0.05	0.01
Response patterns										
Student 1	1	1	1	1	1	0	0	0	0	0
Student 2	0	0	0	0	0	1	1	1	1	1

As it can be seen in Table 1, Student 1 answered the first five items correctly and the last five items incorrectly, which means that s/he failed to answer the five most difficult items. In contrast, Student 2 answered the last five items correctly and the first five items incorrectly, which means s/he answered the five most difficult items but failed to answer the five easiest items. The first student is an example of fitting response pattern and the second student is an example of misfitting response pattern. The likelihood function for Student 1 is presented in Figure1. As showed in Figure 1, a fitting response pattern results in a likelihood function with a clear maximum at the estimated trait level and sharp drop-offs at other values on the ability scale.

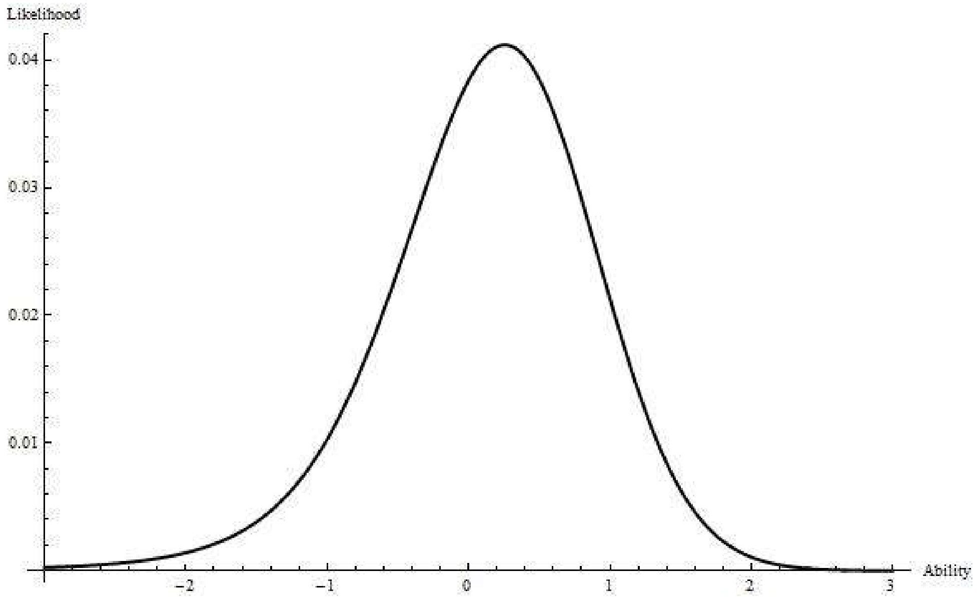

Figure 1: Likelihood function of fitting response pattern

The likelihood function for Student 2 is presented in Figure 2. As demonstrated in Figure 2, a misfitting response pattern results in a likelihood function without a clear maximum and that is mainly flat for lower values of ability, and then tails down similar to the right tail shown in Figure 1. Although both students achieved the same number-correct score of five, they have different response patterns (i.e., fitting versus misfitting) and their ability estimates are different.

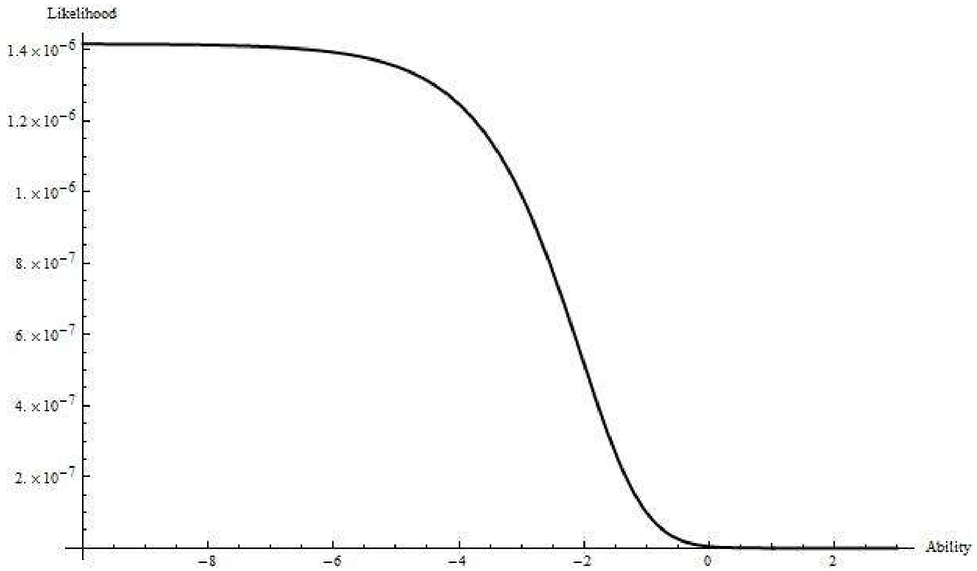

Figure 2:Likelihood function of misfitting response pattern

Whereas, the estimated ability for Student 1 is approximately 0.5, there is no obvious estimate of Student 2's ability. So, any decision making for Student 2 based on his/her ability estimate is questionable as his/her ability estimate may not be accurate. This example demonstrates the need for assessing fit of individual response patterns for the IRT model used.

Person Fit Assessment

One advantage of mathematical modeling in IRT is the possibility of assessing goodness of fit between the observed data and the given model. A goodness of fit index typically measures the discrepancy between the observed data and the values that are expected based on the model used. As a result, model-fit assessment is a major concern and has received much attention from researchers. Whereas a fitted model enables accurate predictions about patterns in observed data, a misfitted model does not (Reise & Widaman, 1999).

There are numerous studies focusing on model-fit in IRT (Hambleton & Swaminathan, 1985). Model-fit assessment can take place at different levels of analysis such as model-data fit, item fit, and person fit. Studies investigating the fit of a response pattern generated by a student to an IRT model have been referred to as "Person-fit "studies (Meijer & Sijtsma, 2001). Person-fit studies use statistical methods to assess the goodness of fit of a student's response pattern to an IRT model or other response patterns in a sample of data. Person fit indices have also been called "scalability indices" (Reise & Flannery, 1996), "appropriateness measures" (Levine & Rubin, 1979), and "caution indices" (Sato, 1975). A general advantage of person-fit statistics or indices is their focus on analyzing individual response patterns. Consequently, more information about students' performance on the test rather than simply relying on test scores becomes available.

Person fit statistics can be divided into two main categories: group-based indices and IRT-based indices. The group-based indices are based on comparing an individual's observed response pattern with expectations based on aggregate-level item characteristics calculated from the overall sample (Karabtsos, 2003). IRT-based indices classify an individual's response pattern based on the extent to which the observed pattern deviates from expected pattern generated by the IRT model used.

Many person fit statistics have been proposed. However, person fit indices l_z and H^T were used in the present study. The H^T statistic, as a group-based person fit index, was selected for this study because it is the most referred to and used group-based statistic (Armstrong & Shi, 2009; Dimitrov & Smith, 2006; Emons et al., 2002, 2003, 2005; Karabatsos, 2003; St-Onge et al., 2011; Zhang & Wlaker, 2008). Karabatsos (2003) reported that H^T was the best functioning person fit index across 36 group-based and IRT-based indices. The choice of the l_z index is based on the fact that it's one of the most studied person fit statistics, and because the l_z index is based on the likelihood function.

Group-based person fit statistics

Most of the group-based person fit statistics compare an observed response pattern with the expected response pattern under the deterministic Guttman model (Guttman, 1944). A deterministic Guttman model states that:

$$\theta_i < \delta_j \quad \leftrightarrow \quad P_j(\theta_i) = 0$$

and

$$\theta_i \geq \delta_j \quad \leftrightarrow \quad P_j(\theta_i) = 1,$$

where $P_j(\theta_i)$ is the probability of student i answering item j conditional on student i's ability (θ) and δ_j is the item difficulty. Based on this model, a student with ability level equal to θ should be able to answer the r easiest items with item difficulties less than or equal to θ (second equation) and fail to answer the rest of J - r items that have item difficulty greater than θ,where J is the total number of items. Such a response pattern is known as "Guttman pattern" or "conformable pattern". If a student answered the J - r items that have item difficulty greater than θ correctly and failed to answer r easiest items with item difficulties less than or equal to θ, then his/her response pattern were considered as "error" or "inversion" or "reverse Guttman pattern" (Meijer & Sijtsma, 2001).

The H^T index

The H^T statistic is not normed against Guttman pattern (Meijer & Sijtsma, 2001). The H^T statistic (Sijtsma, 1986; Sijtsma & Meijer, 1992) is a transformed version of Loevinger's H (1948) scalability coefficient. H^T can be obtained by the following formula:

$$H^T = \frac{\sum_{i \neq j} \sigma_{ij}}{\sum_{i \neq j} \sigma_{\mathrm{ij}}^{\mathrm{max}}} \qquad (6)$$

where σ_{ij} is the covariance between students i and j response vectors and σ_{ij}^{max} is the maximum possible covariance between students i and j response vectors. Given this definition, the H^T statistic measures the correlation between the observed response vector of student i and the response vectors of the rest of N - 1 students. The values of H^T range from -1.0 to 1.0. In other words, H^T compares student's response vector with the other students in the sample. If the observed response vector is consistent with the other students' response patterns, then the numerator of H^T and resulting value of H^T is positive. If the observed response pattern by student i is not consistent with the other students, then the numerator of H^T is negative and the value of H^T is negative. A random response vector which does not correlate with other response patterns in the sample leads to a value of 0 for H^T (Meijer & Sijtsma, 2001).

Sijtsma and Meijer (1992) proposed cut-off value of 0.30 for H^T. Patterns with H^T value lower than 0.30 are classified as misfitting. Karabatsos (2003) examined the performance of 36 person fit statistics and following a sensitivity analysis using a Receiver Operating Curve (ROC) analysis determined a value of 0.22 as the critical value for H^T.

IRT-based person fit statistics

In IRT, like other model-based methods, the goodness of fit between observed data and expected outcome given the model used can be assessed. For example, for an item with difficulty parameter of 0 and using the one parameter logistic model or the Rasch model, the probability of a correct answer to the item can be calculated as:

$$P_j(\theta_i) = \frac{1}{1 + e^{-(\theta_i - b_j)}}, \qquad (7)$$

where θ_i is the i^{th} respondent's ability and $b_j$is the j^{th} item difficulty parameter. $P_j(\theta_i)$ is the probability of a correct answer to the j^{th} item for the i^{th} student. The relationship between

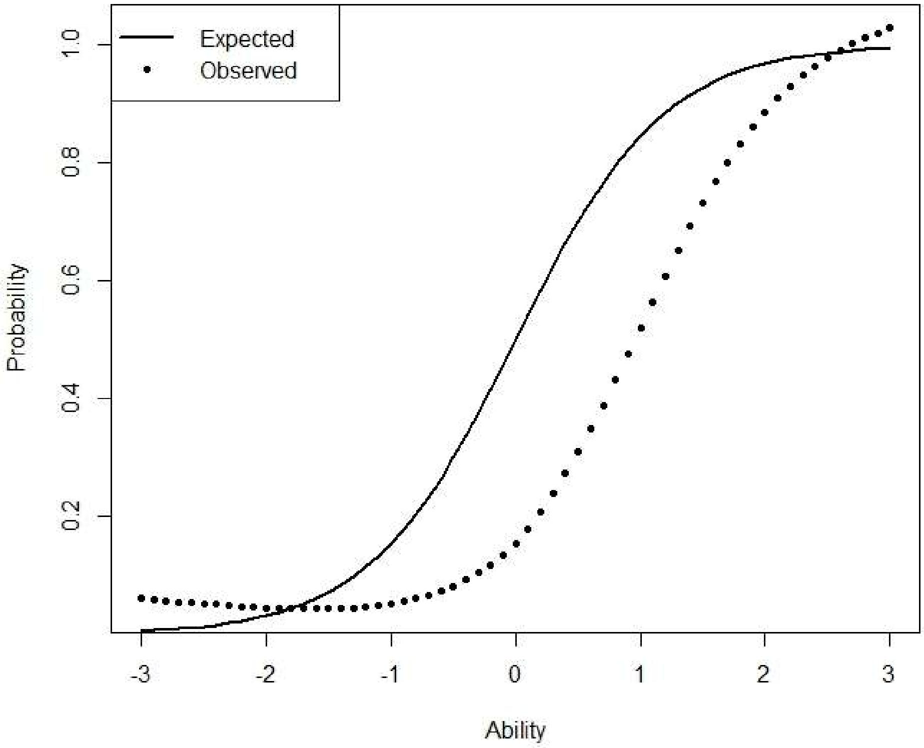

Figure 3: Expected probability (i.e., solid) and observed data (i.e., dotted)

different θ values and $P(\theta)$ can be plotted. This plot, which is known as Item Characteristic Curve (ICC), represents the expected probability of a correct answer for given θ. By plotting the observed data against ICC as shown in Figure 3, it is possible to determine if the data fits the model or not.

IRT has been used as a popular measurement framework in person fit studies (e.g., Karabatsos, 2003; St-Onge et al., 2011) and many person fit statistics have been developed for use within the IRT framework (Meijer & Sijtsma, 2001). The likelihood-based person fit statistic l_0 and its extended standardized version l_z (Levine & Rubin, 1979; Drasgow et al, 1985) are the most referred to and used IRT-based person fit indices in the literature (Molenaar & Hoijtink,1990; Nering, 1995, 1997; Reise, 1995; Krimpen-Stoop & Meijer,1999; de la Torre & Deng, 2008; Magis et al., 2012).

The l_z index

The l_z index is the standardized form of l_0 index (Levine & Rubin, 1979). l_0 is simply the log-likelihood of an observed item response pattern calculated based on an IRT model, which is given by

$$l_{0_i} = \ln\left(\prod_{j=1}^{\mathrm{J}} \mathrm{P}_j(\theta_i)^{\mathrm{X}_{ij}}\left[1 - \mathrm{P}_j(\theta_i)\right]^{1-\mathrm{X}_{ij}}\right), \qquad (8)$$

where X_{ij} is the binary (0, 1) response to item j ($j = 1, 2, \dots$,J) by student i, θ_i is the latent trait for student i, and $P_j(\theta_i)$ is the probability of correctly answering item j. A low value of l_{0_i} suggests that the probability of obtaining the response pattern produced by student i *i*s small given the hypothesized IRT model and, therefore, the response pattern can be considered as a misfit of the IRT model. As pointed out by Drasgow et al. (1985), an undesirable property of l_0 lies in the fact that l_0 is conditional on θ, which suggests that the classification of an observed response vector as fitting or misfitting is influenced by θ. Therefore, the standardized form of l_0, denoted by l_z, was derived with an asymptotic normal distribution:

$$l_z = \frac{l_0 - E(l_0)}{[Var(l_0)]^{1/2}}\,, \qquad (9)$$

where

$$E(l_0) = \sum_{j=1}^{J}\big(P_j(\theta)ln\big[P_j(\theta)\big] + \big[1 - P_j(\theta)\big]\,ln\big[1 - P_j(\theta)\big]\big) \qquad (10)$$

and

$$Var(l_0) = \sum_{j=1}^{J} P_j(\theta)\big[1 - P_j(\theta)\big]\left[ln\frac{P_j(\theta)}{1 - P_j(\theta)}\right]^2 \qquad (11)$$

Large negative values of l_z suggests potential misfit. One drawback of l_z is that its asymptotic distribution is not standard normal when the sample estimates of θ replace the true values of θ.

In order to overcome this drawback of l_z, Snijders (2001) proposed a modified version of l_z index which is referred to as l_z^*. Snijders argued that l_z^* can be used for different ability estimators and IRT models. The modifications made by Snijders are based on the mean and variance of l_z. l_z^* is not widely used mainly due to complexity that arose from modification made (see Magis et al. (2012) for more detail).

Research on Person Fit Statistics

This section is divided into two sub-sections. First, the performance of different person fit statistics is reviewed. The second section looks at the factors that may cause misfitting response patterns.

The performance of person fit indices

Research on the behavior of person fit indices includes a variety of topics and methods, but the variety can be summarized in three categories: 1) performance of person fit indices, 2) application of person fit statistics, and 3) influence of person misfit on item and test characteristics (Meijer & Sijtsma, 2001). The seminal review paper by Meijer and Sijtsma (2001) summarized thoroughly all major person fit studies in the literature before 2001. Therefore, only the research conducted after 2001 is presented here.

The literature on person fit research is dominated by evaluating the performance of person fit statistics under a variety of conditions. These conditions include, but are not limited to, test length, sample size, different item parameters, different measurement models, different ability estimators, different misfitting response types, percentages of misfit in the data set, and different person fit statistics. Most of the studies examining the performance of person fit statistics are based on simulated data.

Karabatsos (2003) studied and compared the performance of 25 IRT-based and 11 group-based person fit index under 60 conditions derived from a fully-crossed three factor design. The factors with the number of levels were: five types of misfitting responding students (cheaters, creative respondents, guessing, careless, and random respondents), four percentages of misfitting-responding students (5%, 10%, 25%, or 50%), and three test lengths (17, 33, and 65 items). The IRT model used was the Rasch model. Karabatsos' study is one the most comprehensive studies on the performance and comparison of person fit indices and based on the Google scholar (up to the time of writing this document), this paper has been cited 93 times. The results revealed that overall group-based indices outperformed IRT-based indices. The *C* (caution index, Sato, 1975), *MCI* (modified caution index, Harnisch & Linn, 1981), *U3*, and H^T, which are group based indices, and the *D(θ)* index (Trabin & Weiss, 1983), which is an an IRT-based person fit index, had the best performance. The *D(θ)* (Trabin & Weiss, 1983splits items into *S* subsets and sums within each subset the average response residuals over the *S* subsets. The problem with $D(\theta)$ is that splitting items into subsets is arbitrary and that by dividing items into S subsets in different ways, researchers may obtain different results. Karabatsos (2003) concluded that the H^T statistic outperformed the other four indices because it takes into account each observed response pattern in identifying misfitting patterns. As for the weak performance of IRT-based person fit indices, Karabatsos argued that the poorer performance of the $D(\theta)$ and the other IRT-based statistics is because of using the data twice, first for estimating item and ability parameters and then for calculating person fit statistics. However, this argument is true for group-based indices as well because the data are also used twice, once for calculating item difficulty (i.e., p values) and then using them for calculating person fit.

Contrary to the findings of Karabatsos (2003), l_z tended to be one of the most powerful indices on detecting misfitting response patterns in previous studies (e.g., Drasgow, Levine & McLaughlin, 1991; Reise, 1995; Nering 1995, 1997; Li & Olejnik, 1997). In a more recent study, Armstrong et al. (2007) reexamined the detection effectiveness of l_z under different test characteristic conditions. They used the 3-PL IRT model as the measurement model for generating and analyzing data. Two types of misfitting responding were simulated, namely spuriously low (SL) and spuriously high (SH) responding. The spuriously low responding behavior refers to the case that student's estimated ability is lower than his/her actual ability level and the spuriously high responding behavior refers to the case that student's estimated ability is higher than his/her actual ability level. As for the test characteristics, three different ranges of item difficulty, two ranges of item discrimination, and two ranges of item pseudo-guessing parameter were considered separately while other parameters were fixed. Additionally, Armstrong et al. assumed that the ability parameter was distributed evenly over 61 points from -3 to 3 and used this range for calculating empirical critical values of l_z at each θ by simulating

10,000 students and finding the lower 5th percentile. Their results illustrated that l_z should be used with caution under certain conditions. Generally, l_z showed poor performance when spuriously high (SH) cases were manipulated to have incorrect response with probability of 0.20 over the negative values of θ. Armstrong et al. concluded that l_z performed better when the range of difficulty parameters of items was aligned with the range of θ values. Moreover, they suggested that using Bayesian estimation for the ability parameter instead of the maximum likelihood estimator (MLE) may improve the detection rates.

As mentioned by Meijer and Sijtsma (2001) and Snijders (2001), when replacing the true ability parameter, θ, by its estimated value, $\hat{\theta}$, the l_z distribution under null hypothesis is no longer a standard normal distribution. de la Torre and Deng (2008) suggested a procedure for correcting the ability estimate and its reference distribution by taking into account the unreliability of the test in order to improve the accuracy of person fit. Their method involves multiplying the estimated ability, $\hat{\theta}$, based on expected *a posteriori* (EAP) estimate, by its variance (i.e., the reciprocal of test information function at given $\hat{\theta}$). The resulting ability estimate was called $\hat{\theta}^*$. Then l_z^* should be computed using new estimated ability parameter. The next step is to generate a new set of ability values from $N(\hat{\theta}^*, \frac{1}{1+I(\hat{\theta}^*)})$ to find the corrected $\hat{\theta}^{*new}$ and then computing l_z^{*new}. The final step is to compute *p*-values for each l_z^* as the proportion of l_z^{*new} values that are less or equal to l_z^*. In their study, they compared results of calculating l_z with true, estimated ($\hat{\theta}$), and correct ability estimate ($\hat{\theta}^*$) and l_z^* for different test lengths and aberrant behavior as contributing factors (i.e., 10, 30, and 50 items and six aberrance conditions) utilizing the 3-PL IRT model. The results showed thatl_z^* outperformed l_z and had Type I error rate very close to the nominal error rate. The results also indicated that l_z as well as l_z^* underestimated Type I error rates for small α and for shorter tests it was close to nominal Type I error rate. The proposed method seems to be promising but it's computationally intensive and may not be appropriate in practical situations. Furthermore, the proposed method for adjusting estimated ability parameter assumes that the initial estimate is an underestimate of the true ability but the authors did not provide any rationale for this assumption.

St-Onge et al. (2011) studied the accuracy of four person fit statistics (i.e., l_z, H^T, *U3*, and $ECI2_Z$) in their simulation study. The main goal of their study was to determine the detection rate of the four indices for data with a high percentage of misfitting response patterns in sample (i.e., more than 40%). The design of the study was a completely crossed design comprising of five factors (i.e., 2 item difficulty ranges × 2 item discrimination ranges × 21 aberrance rates × 2 aberrance types × 4 test lengths). The 2-PL IRT model was used for data generation and 10% to 60% of misfits with an increment by 2.5% were manipulated to simulate spuriously low and spuriously high aberrance types. The item parameters were estimated using the Bayesian Modal estimator (Baker & Kim, 2004) and the ability parameters were estimated using maximum likelihood. Their results indicated that for aberrance rates up to 30 to35% the detection rate increased linearly as the aberrancy rate increased and that for higher aberrance rates the relationship between aberrance rate and detection rate was no longer linear. The point that linear relationship became non-linear was different across person fit indices and data sets. St-Onge et al. (2011) explained that there is an increase in detection rate accompanied with an increase in aberrance rate until reaching a peak in detection rate. Afterwards, an increase in aberrance rate resulted in a decrease in detection power. Based on their results, all four person fit indices were generally relatively robust against an increase in aberrance rate and $ECI2_Z$ was the most robust of the four indices. They suggested that for optimal performance of person fit statistics, especially

IRT-based indices, it is better to use them when there are 25% to 35% of misfitting response patterns.

In line with Karabatsos (2003), Huang (2012) compared the performance of five group-based and five IRT-based person fit indices. The five group-based indices were the caution index (*C*; Sato, 1975), modified caution index (*MCI*; Harnisch & Linn, 1981), norm conformity index (*NCI*; Tatsuoka & Tatsuoka, 1982), and *Wc* and *Bs* (D'Costa, 1993). The five IRT-based indices were the *OUTFITz* and *INFITz* (Smith, 1991, Linacre & Wright, 1994), $ECI2_Z$ and $ECI4_Z$ (Tatsuoka & Linn, 1983), and l_z. Three factors were manipulated: aberrance type with three levels (i.e., spuriously high alone, spuriously low alone, and combination of both), aberrance severity (i.e., 10%, 20%, and 30% of items containing misfitting responses), and aberrance rate in data set (i.e., 10%, 20%, and 30% of sample). Item and ability parameters were estimated using BILOG-MG and the analytical process consisted of ten steps from simulating data based on empirical data to generating and analyzing 81,000 simulated aberrance responses. The results of Huang's study showed that group-based indices performed better than IRT-based indices generally across all conditions. The $ECI2_Z$ and *INFITz* had the lowest detection rates and l_z outperformed the other IRT-based statistics. Huang (2012) concluded that the good performance of group-based indices may be due to the sensitivity of these indices to response patterns instead of response probabilities.

Many studies have shown that when the true ability is estimated, l_z does not follow standard normal distribution (e.g., Nering, 1995; Reise, 1995; Van Krimpen-Stoop & Meijer, 1999; Snijders, 2001). Therefore using -1.65 as cutoff value for identifying misfitting response vectors can be inaccurate. Recently, Seo and Weiss (2013) showed that using empirical cutoff values derived from their simulation study using estimated item parameters is a better approach in real testing situations. They examined data from 20 exams for an introductory psychology course at a Midwestern U.S. university. Data were analyzed using the 3-PL IRT model. l_z values were calculated for three estimates of ability parameter based on maximum likelihood (MLE), maximum a posteriori (MAP), and expected a posteriori (EAP) estimation methods. The cutoff value was determined based on bottom 5% of the empirical distribution of l_z values computed via 10,000 model-fitting simulated responses using estimated item parameters. Their results illustrated that MLE estimate of ability produced closer l_z means to 0 compared to MAP and EAP, but for all three estimation methods the standard deviation of l_z was lower than 1.

Explanation of misfitting response patterns

Although person-fit statistics are sensitive to misfitting response patterns, finding these patterns does not explain why an individual has responded in a particular way (Meijer & Sijtsma, 1995). Person fit research that has focused on using person fit indices along with other statistical methods in order to provide reasonable explanations for misfitting responding behavior in different contexts, such as aptitude, personality and technology based testing, are presented next..

Lampriauou and Boyle (2004) evaluated measurement accuracy using the Mathematics National Curriculum test data (England) for ethnic minority students and students speaking English as a second language. They chose the Rasch model as the IRT analytical model and *Infit* and *Outfit* (Wright & Stone, 1979) as person fit indices. They found that students with English as a second language and students belonging to minor ethnic minorities were significantly more likely to produce misfitting responses. They concluded that results showed that students from ethnic minorities and who speak English as a second language are more likely to be mis-measured and warned about the consequential validity of the test.

Petridou and Williams (2007) used hierarchical linear modeling (HLM) to account for misfitting response patterns on a mathematics assessment. In their study, the Rasch model was used and *Infit* and *Outfit* (Wright & Stone, 1979) were the person fit indices. They found that all examined person-level variables except gender (i.e., ability, language, anxiety, and motivation) had a statistically significant effect on person aberrance as indicated by the *Infit* index. These significant relationships suggest more able students, students with an additional language spoken at home, less anxious, and less motivated students were significantly more likely to provide misfitting response patterns. According to the single-level *Infit* model, one unit increase in ability resulted in an increase in the odds of producing an misfitting pattern by approximately 70%, while a unit increase in anxiety and motivation resulted in a decrease in the odds of producing an misfitting pattern by 15% and 8%, respectively. For the *Outfit* statistic, the results suggested that ability, anxiety, and motivation had significant contributions. These significant relationships suggest that students who were more able, more motivated, and less anxious were more likely to provide misfitting response patterns as identified by the *Outfit* statistic. They also suggested that class-level factors such as general administration problems, non-standard administration practices such as interpreting questions, class "cheating" by leaving support materials on classroom walls, and instructional effect in terms of topics/items in the test not being taught at the time of test administration may be associated with aberrancy in test data.

In the study by Dodeen and Darabi (2009), four personality tests for Mathematics (i.e., attitude toward mathematics, level of mathematics anxiety, level of test anxiety, and level of motivation to learn mathematics) were considered and their relationships with l_z as person-fit index were investigated. The results showed that motivation toward learning mathematics had the strongest relationship with person-fit followed by test anxiety. While a student's ability measured by the total score had no effect on person-fit, mathematics anxiety and attitude toward mathematics seemed to have some effect on it. Studies like those conducted by Petridou and Williams (2007), Lampriauou and Boyle (2004), and Dodeen and Darabi (2009) can provide a valuable basis for explaining testing behavior undertaken by students.

Person fit indices have also been used for other purposes. For instance, Finkelman and Kim (2007) applied person fit statistics in the procedure for the “Body of Work (BOW)”standard setting procedure. In this procedure, panelists analyze students’ actual response patterns and categorize each student into one of multiple performance levels. Finally, after several rounds, panelists set cutoff values for students’ response sets. Finkelman and Kim (2007) argued that the inclusion of misfitting response patterns in the categorization procedure due to random or haphazard responding may threaten the validity of the standard setting procedure. Therefore they used person fit indices for selecting student work with fitting response patterns as the pool student work for the BOW. They used l_z (Trabin & Weiss, 1983) and a proposed group-based form of l_z which is analogous to standardized form of *U3* (*ZU3*; van der Flier, 1982). Their results showed that person fit indices were successful in reducing the selection of students whose response patterns were inadequate for the BOW procedure.

Woods et al. (2008) applied two-level logistic regression as the person fit detection method to 13 personality scales used for recruiting in the military. Their results showed that five scales were more discriminating for some recruits. Further investigation of the misfitting response patterns revealed that the aberrancy may be due to personality pathology, which was not detected during initial screening stages. An examination of differential test functioning revealed evidence related to gender, ethnicity, or both. Authors concluded that person fit assessment can be a good tool for improving psychological measurement in order to identify persons who may have provided invalid data or need special attention.

Liu and Yu (2011) used person fit indices for detecting misfitting learning in a personalized e-learning environment. They applied IRT ability estimation along with *Infit* and *Outfit* indices to detect misfitting learning. The personalized system develops the learning path for each learner based on the course difficulty and learner's estimated ability. After each unit, the learner's ability estimate was updated via information collected during the unit and values for *Infit* and *Outfit* were computed. If the results showed significant misfit then, first, a virtual tutor encourages the learner to try again and concentrate on learning modules but, second, if this strategy did not work a human tutor may get involved to help the learner with a misfitting pattern to address the misfitting. Using an experimental design, results showed that this system enhanced learning efficiency and improved the learning experience for the group that used the personalized e-learning system.

Person fit statistics have also been applied to a personality test for assessing consistency of responding to the test. Ferrando (2012) applied l_z to two personality measures (i.e., Neuroticism and Extraversion) and found 4% misfitting response patterns using -2.0 as the cutoff value for both measures. The results showed that idiosyncratic interpretation of some items and low person reliability were the main sources of aberrancy for the both measures. Person reliability refers to "the clarity and sensitivity with which individuals perceive their own trait level, and is supposed to depend on the relevance and the degree of organization with which the trait is internally organized in these individuals" (Ferando, 2012, p.719). Ferrando also argued that potential application of person fit analysis on detecting inconsistent responding behavior is very relevant and important in both clinical and validity studies.

In a recent study, Tendeiro et al (2013) used cumulative sum statistics (*CUSUM*) to detect inconsistency in unprotected Internet testing (UIT). *CUSUM* is a person fit methodology derived from Statistical Process Control to be applied in situations in which there is a sequence of responses like computerized adaptive testing (Armtsong & Shi, 2009). As mentioned by Tendeiro et al. (2013), UIT is becoming one of the most popular methods in personnel selection because of its flexibility. However, it also has several problems such as lack of enough security or possibility of cheating. The most common method for overcoming these shortcomings of UIT is to administer a follow up confirmation or verification test in a secured and supervised environment. The authors applied *CUSUM* in order to identify suspicious decreases in performance from the UIT to confirmation test. They used l_z and z statistic (Guo & Drasgow, 2010), a specific measure developed to identify cheating in UIT, for detecting misfits. The results indicated that *CUSUM* methodology is a promising approach in detecting misfit in unprotected Internet testing.

Influence of Person Misfit on Item and Test Characteristics

As reported by Meijer and Sijtsma (2001), there are some studies that have examined the effect of misfitting response patterns on test characteristics and the item parameters. Levine and Drasgow (1982) used the 3-PL IRT model to study 1) the effect of using estimated item parameters versus true item parameters on the detection rate of l_0 and 2) the effect of misfitting response patterns on the detection rate of l_0 and item parameter estimation. Estimated item parameters from a previous calibration of the verbal section of the Scholastic Aptitude Test were used to generate simulated data. Spuriously low responding was simulated using about 7% of the sample in which 20% of the item scores of the students were changed to be correct with a probability of 0.20 and incorrect with a probability of 0.80. Levine and Drasgow concluded that the presence of misfitting response patterns had no effect on the detection rate of l_0 and item parameter estimation. They argued that different misfitting responses tended to have different incorrect response patterns and that, consequently, a large number of the misfitting response patterns would have opposing effects on estimated item parameters.

Philips (1986) investigated the effect of misfitting response vectors on the fit of Rasch model, estimated item parameters, and equipercentile equating. He found that deletion of misfitting response patterns can improve the model-data fit, had small effect on estimated item difficulty parameters, and essentially no influence on the equating results. In another study, Runder, Bracey, and Skaggs (1996) analyzed the 1990 NAEP data and found almost no misfiitting response vectors. Consequently, removing misfitting patterns did not result in a significant difference in the mean of the test before and after deleting misfits.

Sotaridona et al. (2005) studied the effect of person misfit on item calibration and performance classification. In their study, they utilized l_z^* and *U3* as person fit measures and a random sample of 10,000 students from a statewide assessment program for Language Arts, Mathematics, and Science. Each test consisted of 55 multiple choice items. Two types of misfitting responding (i.e., copying and guessing) were simulated by manipulating data of selected students. As a result, two sets of data – the original data set and the data set with simulated misfitting responses – were simulated. The data sets were calibrated independently using the 3-PL IRT model and estimated item parameters were equated by the Stocking and Lord method (1983). Four criteria were chosen for comparison including differences in equated parameter estimates, differences in standard errors of the item parameter estimates, differences in test characteristic curves, and differences in test information curves. The results revealed that the estimated parameters were consistently larger in the presence of person misfit and the standard error of estimation was higher for the data with misfitting response patterns. Test characteristic and information curves were not significantly different. In addition to the original data set, two additional data sets were created, a data set comprising selected fitting responses using l_z^* and a data set containing selected fitting responses using *U3*. Each data set was calibrated and equated independently and standardized scale scores were converted into three levels below proficiency, proficient and advanced. The results suggested that the differences were negligible and for the fitting data set obtained by *U3*, differences were generally smaller. They concluded that inclusion of misfitting response patterns had reduced the accuracy of parameter estimation, but at the test level the effect was minimal.

Hendrawan et al. (2005) also investigated the effect of misfitting response patterns on classification decisions. They used three ability estimation methods – MLE, EAP, and Markov Chain Monte Carlo (MCMC) – with the three parameter normal ogive model. They used marginal maximum likelihood (MML) and Bayes Modal estimators to estimate the item parameters. The five person fit indices utilized in the study were W (Wright & Stone, 1979), *UB* (Smith, 1986), ζ_1 and ζ_2(Tatsuoka, 1984), and l_z. In the simulation, two test lengths (i.e., 30 and 60 items), two misfitting response types (i.e. guessing and item disclosure), three item discrimination values (i.e., 0.5, 1 and 1.5), two sample sizes (i.e., 400 and 1000), and three cutoff values for determining mastery/nonmastery on the test (i.e. -1, 0, and 1) were exaamined. The results showed that the presence of misfitting response patterns resulted in biased estimates of item parameters and inaccurate mastery classification decisions, especially for guessing behavior which lowered the mean of distribution of estimated abilities. This resulted in artificially higher classification accuracy for students with low ability due to the increased estimated guessing parameter and decreased the estimated difficulty parameter. In the case of item disclosure, where student has pre-knowledge of the items, results were opposite to guessing behavior. The classification accuracy was higher for students with high ability because item disclosure led to higher estimated ability. All person fit indices performed well and resulted in increased classification decision. Generally, results were the same across estimation method and MCMC was the worst in terms of classification accuracy for the fitted and misfitting response patterns. As an overall conclusion, Hendrawan et al. argued that person fit statistics are useful in finding fitting subsamples and are appropriate for using in mastery testing.

Summary

In this chapter, it was noted that there are two main categories of person fit indices: "group-based" and "IRT-based" statistics for assessing a student's response vector. The former is based on the comparison of chosen response vector with other response vectors in the sample and the latter is based on the comparing observed response pattern with expected response pattern generated by a given analytical model. A review of relevant and recent studies in person fit research showed that the main focus of research in person fit is on the improving and evaluating performance of person fit indices. In addition, several applications of person fit statistics in areas like providing explanations for misfitting testing behavior and improving test validity and classification decision making indicates the potential of such indices in enhancing the practice of testing.

Although there are a few studies on the effect of misfitting response patterns on test characteristics and item parameters, there is an obvious need for more investigations on the effect of presence of misfitting responses on the item parameter estimation utilizing simulated data based on different factors that mimic real world testing situations. More specifically, previous studies are deficient in providing information on the effect of factors like sample size, test length, item parameter estimation method, percentage of misfit in data set, and percentage of item susceptible to misfitting responses on estimated item parameters. Such information, based on simulated data, can shed light on how severely item parameters may be affected due to presence of misfitting responses. It is important to investigate the effect of misfitting responses on item parameter estimations because estimated item parameters are used not only for estimating proficiency parameter but also for other procedures like equating and scaling in testing practice.

Chapter Three: Method

The analytical methods that were employed to examine the effect of misfitting response patterns on item parameter estimation are presented in Chapter 3. The first section discusses the overall simulation design and factors that were manipulated. The second section describes data generation procedures for generating item and ability parameters as well as different misfitting response patterns.

Simulation Design

Increasingly, simulation studies are being used to investigate how IRT based methods can be applied (Harwell et al, 1996). In this study, a simulation method was used to evaluate how estimated item parameters are affected by the existence of misfitting response patterns under different conditions.

There are several design factors that are often involved in simulation studies of person fit research. These primary factors include sample size, number of simulated items, statistical model chosen for data generation, and values for item and person parameters (Rupp, 2013). Additional factors that have been used in the past are type of parameter estimator and misfitting responding (de la Tore & Deng, 2008; van Krimpen-Stoop & Meijer, 2002; St-Onge, Valois, Abdous, &Germain, 2009), type of person fit statistic (Karabatsos, 2003; St-Onge, Valois, Abdous, &Germain, 2011), and percentage of misfitting response pattern (Armstrong & Shi, 2009; Clark, 2010; Emons, 2009). Design factors and corresponding levels in this study were chosen with respect to the previous research and to mimic real testing situations.

Four factors were considered in this study. These included test length, type of item parameter estimator, percentage of misfitting response patterns, and percentage of items susceptible to misfitting responding.

Test length: There are methodological reasons for simulating longer tests such as obtaining higher reliability and increasing chance of detecting misfitting responses (Rupp, 2013). The number of items chosen for previous simulation studies on person fit statistics ranged from 10 to 121 items. In this study, tests with 20, 40, and 60 items were chosen as these test lengths are more practical and used in real world situations.

Type of item parameter estimator. Given the fact that different estimators are based on assumptions and violations of these assumptions that affect the estimated item parameters, two item parameter estimation methods were used. These estimation methods were Marginal Maximum Likelihood (MML) estimation and Bayes Modal estimation.

Percentage of misfitting response patterns. The percentage of misfitting response patterns can have an impact on the performance of PFSs such as an increase in the percentage of misfitting response patterns leads to decreased detection power (Rupp, 2013). While the percentage of aberrancy in the data set varied from 1 to 100 percent in previous studies, three percentages of 10%, 20% and 30% misfitting response patterns, which were used most frequently studied in previous research studies, were chosen for the present study.

Percentage of items susceptible to misfitting responding. When determining the number of items in a test susceptible to misftting responses, it was assumed that misfitting responses could occur for any item. An increase in the percentage of items susceptible to misfitting responding leads to higher degree of severity of person misfit. While the percentage of aberrancy in the items varied from 20% to 100% in previous studies, percentages of 25% and 50% were chosen for the present study. Given these factors were fully crossed, the design of this study was a 3×2×3×2 (test length-by-item estimator-by percent of students suscpetibile to producing misfitting response patterns, and percentage of items susceptible to misfitting responses) full crossed design. The l_z and H^T person fit indices were used to remove misfitting response patterns from the data sets with misfitting response patterns. Given the use of 100 replications is common in person fit research (e.g., Hendrawan et al., 2005) and a preliminary analysis revealed trivial difference between using 100 and 1000 replications, each of the 36 conditions was replicated 100 times. A sample size of 5,000 students was used for all conditions. All computational procedures were done using a written program in the R software (R Core Team, 2013).

Data Generation

In the simulation study, the 2-PL IRT model was the analytical IRT model used for analyzing the test data. A normal distribution with mean of 0 and standard deviation of 1 was used for generating students' abilities. Item difficulty parameters were generated using a uniform distribution on the closed interval [-2.7, 2.7]. The item discrimination parameters were generated using a uniform distribution on the closed interval [0.5, 2.5]. The reason for this choice of these two distributions was to cover all feasible values of item difficulty and item discrimination found in practice. Also, these distributions of the item parameters were used by other researchers (e.g., Armstrong et al, 2007; Choi & Cohen, 2008; Karabatsos, 2003; Sijitsma & Meijer, 2001).

Four data sets were created. The *fitting response data sets*, which had no msifitting responses, were generated with respect to test length. The procedure involved the following steps: First, the true theta values for respondents were drawn from a standard normal distribution. Next, Equation 7 was used with corresponding item parameters to calculate the probability of a correct answer for each student. Then a random number y from a uniform distribution on the closed interval [0, 1] was generated to assign 1 (as correct response) and 0 (as incorrect response) to the student response. If $y < P_j(\theta_i)$, then the response to the item j was set to 1. If $y \geq P_j(\theta_i)$, the response was set to 0. The generated data sets were analyzed to calibrate items and determine ability parameters. Marginal maximum likelihood (MML) and Bayes modal estimation methods were used for the for item parameter estimation and the Expected A Posteriori (EAP) method was used for ability estimation.

The *misfitting response data sets* were generated with respect to test length, percentage of items susceptible to misfitting responding, and percentage of misfitting response patterns. A 50-50 percent mixture of two types of misfitting behavior was considered: spuriously high (SH) responses and spuriously low (SL) responses. As argued by Rupp (2013), there are many labels utilized by researchers for distinguishing different types of misfitting response behaviors. Most of the labels are attributed to possible underlying causes of misfitting responding but statistical implementation pertinent to each label results in, generally, spuriously high or spuriously low responding. The 50-50 split was used due to the fact that it is very rare to have only one type of misfitting response pattern in real-world testing situations and it is more likely to have a mixture of both. Spuriously low responses occur when a person obtains a lower score than what would be expected based on the given model. Spuriously high responses occur when a person obtains a higher score than what would be expected based on the given model. To generate responses leading to spuriously low responding, responses of high ability students (i.e., students with $\theta \geq 0$) were chosen and responses to 25% or 50% of the randomly selected items were changed to be incorrect. For generating responses resulting in spuriously high responding, students with $\theta < 0$ were selected and responses to 25% or 50% of the randomly selected items were set as correct. By taking this approach, misfitting response patterns differed from one student to another one

(Armstrong et al, 2007). For each level of percentage of misfitting responses in sample, a 50-50 percent allocation was used to generate spuriously high and spuriously low responding behavior. For example, for the 10% misfitting response patterns in the sample, 5% of the responses were manipulated to be spuriously high and 5% were manipulated to be spuriously low.

The data sets with misfitting response sets removed by l_z and by H^T data sets were formed for each condition. The first data set had the misfitting response patterns removed by l_z. The second data set had the misfitting response patterns removed by H^T.

The true item parameters in addition to the estimated item parameters using ML and BM methods from the four above-mentioned data sets were analyzed. The Figure 4 shows the overall flowchart of obtaining all sets of item parameters used in this study.

Evaluation Criteria

To investigate the effect of exclusion of misfitting response patterns on item parameter estimates, estimated item parameters before and after deletion of misfitting response patterns were compared in terms of the magnitude and direction of change. Bias for each item parameter was computed as the difference between the mean of the estimated parameters and the true parameter value across the 100 replications for each condition. The formula for the bias in the *b* estimated parameter is:

$$Bias(b_j) = \frac{\sum_{k=1}^{100} \hat{b}_j^k}{100} - b_j \, , \qquad (13)$$

where b_j *i*s the generated true item parameter for item j and $\hat{b}_j^k$ is the estimated item parameter for item j for the k^{th} replication. The same formula was used the estimated a parameter with *b* replaced by a. Unfortunately, there were no determined criteria on how large item parameter

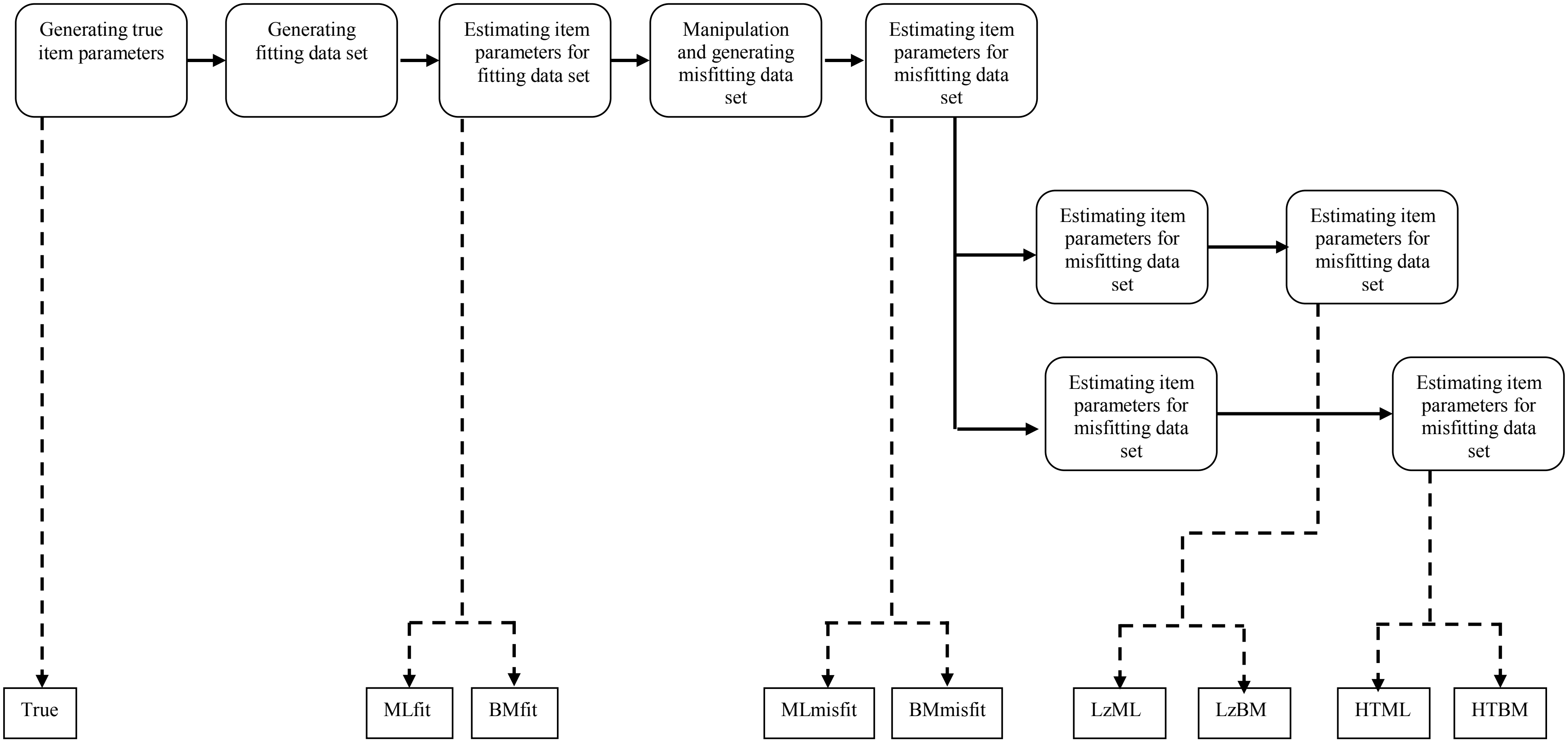
Generating true item parameters
Generating fitting data set
Estimating item parameters for fitting data set
Manipulation and generating misfitting data set
Estimating item parameters for misfitting data set
Estimating item parameters for misfitting data set
Estimating item parameters for misfitting data set
Estimating item parameters for misfitting data set
Estimating item parameters for misfitting data set
True
MLfit
BMfit
MLmisfit
BMmisfit
LzML
LzBM
HTML
HTBM

Figure 4: The procedure of obtaining different sets of item parameter estimates

estimation bias should be to be considered severe. Therefore, for the purpose of this study three ranges of bias were used: equal to or between -0.10 and 0.10, -0.20 to -0.10, and 0.10 to 0.20 and biases that are either smaller than -0.20 or larger than 0.20.

For the two data sets with misfitting response patterns removed, a cut score for differentiating a misfitting response pattern from a fitting response pattern needed to be set. The bootstrap method was used to set cutoff values. For each test length, a data set comprising of fitting response patterns was generated using the item parameters estimated from the data set with all the misfitting response patterns. The rational for this approach is that in practice, true item parameters are unknown and only estimated item parameters are available. This simulated fitting data set was used to determine the cut score for each data set as follows:

1. The value of person fit index (i.e., l_z and H^T) was calculated for each response pattern of the simulated fitting data set.
2. A random sample of 5,000 person fit values (i.e., equal to the sample size of the data set with all the misfitting response patterns), was selected with replacement, from the observed person fit values in step 1 and the 5th percentile rank of this random sample was found.
3. Step 2 was replicated 1,000 times to have a data set comprising of 1,000 estimates of the 5th percentile rank.
4. The median of the 1,000 estimates was used as cut score to differentiate fitting response patterns from misfitting response patterns.

This cut-score was then used for the two data sets in which misfitting response patterns were removed by l_z and by H^T.

In order to assess the effect of inclusion and exclusion of misfitting response patterns on the estimated parameters at the test level, the value of mean absolute deviation (MAD) was calculated for each condition and then the mean of the MADs for the 100 replications was computed taking into account the number of items or test length for each condition:

$$MAD = \frac{\sum_{k=1}^{100} \frac{\sum_{j=1}^{J} \left| \hat{b}_j^k - b_j \right|}{J}}{100} \quad , \qquad (14)$$

where J is the test length (i.e., 20, 40 or 60).

To investigate to what extent bias in the estimated *b* and *a* parameters influenced the estimate of examinee ability, the classification accuracy of placing students in one of two classes was examined. Students with true $\theta < 1.00$ were classified as "low" performing and students with true $\theta \geq 1.00$ were classified as "high" performing. Students with estimated $\hat{\theta} < 1.00$ were classified as low performing and students with estimated $\hat{\theta} \geq 1.00$ were classified as high performing. The 2 x 2 classification contingency tables were developed with the frequency of students in each cell for each of the 100 replications. The mean number of students for each cell was then calculated across 100 replications. The values were then rounded to the nearest whole number using scientific rounding.

Chapter Four: Results for the Test with 20 Items

Chapter Four presents the results of the analyses of the simulation for the test with 20 items. The results for the 40 item test are provided in Chapter Five and the results for the 60 item test are provided in Chapter Six.

The results for the 20 item test are presented in two sub-sections. In the first sub-section, the results for item parameter and classification accuracy estimates using a complete set of tables (i.e., item bias, summary table of graph, and classification accuracy) and the graph with two panels for the first simulation condition. Given the length of the first table and the overlap with the graph of the results reported in this table, the second sub-section includes only the graph and tables for the summary of the graph and classification accuracy for the remaining conditions for the 20 item test and for the 40 item test (Chapter Five and 60 item test (Chapter Six). The tables for the bias in the *b* and *a* parameters are provided in Appendix A. Since the range of standard errors of estimate of the estimated item parameters was small for all the six conditions for each test length, ranging from 0.002 to 0.20 for all simulation conditions, only the bias of item parameter estimates are presented for each condition in the tables with the bias results.

Effect of Misfitting Response Patterns on Item parameters

The complete set of results is provided for the simulation condition in which there were 20 items, 25% of the items were susceptible to mifitting responses, and 10% of the students responded misfittingly. The complete set consists of a table in which the true difficulty (*b*) and discrimination (*a*) parameters used to generate the fitting data and the degree of bias in estimating *b* and *a* yielded through Maximum Likelihood estimation (ML) and Bayes Modal estimation (BM) are presented for the fitting response data (fit), response data with all the misfitting responses (misfit), response data after students with misfitting response patterns identified using the l_z procedure were removed, and response data after the students with misfitting response patterns using the H^T procedure were removed. The ML estimate of *b* and *a* for each type of student sample (e.g., fitting, misfitting, l_z, and H^T) are then displayed in a graph for *b* and for *a.* The results in first table and the corresponding graphs are summarized in a second table in terms of the magnitude of the bias. Lastly the classification accuracy of placing students in one of two performance categories is presented for each type of sample to determine if the presence of bias in some items resulted in different estimates of ability, thereby leading to different placements. In the second sub-section, the results for the remaining 17 simulation conditions are provided. The results include the graphs for *b* and *a*, the summary table, and the classification table. The corresponding tables containing the bias for each item are provided in Appendix A for the 17 conditions.

20 items, 10% of students with misfitting response patterns in 25% of items

Table 2 contains the bias estimates for the set of 20 items with 25% of the items susceptible to misfitting responses and 10% of the students with misfitting response patterns. The bias estimates for item difficulty are provided in the top panel and the bias estimates for item discrimination are provided in the lower panel. The items are listed in the same order in each panel. The column with item numbers is shaded with light and dark grey. The light grey signifies that only one parameter (i.e., either *b* or *a*) had estimation bias beyond ± 0.20 in at least one data set (e.g., fitting, misfitting, l_z, and/or H^T). The dark grey signifies that the shaded item had estimation bias beyond ± 0.20 for both *b* and *a* in at least one data set. The first column of the table contains the true value of the *b* parameter (top panel) and the *a* parameter (bottom panel) of each item. Lastly, the mean absolute deviation (MAD) values are presented at the bottom of each

Table 2
Bias of estimated item parameters for manipulation of the 25% of items and 10% of sample

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4810	4813	4706	4708
Item difficulty	Item1	-1.05	0.00	0.00	-0.04	-0.04	-0.04	-0.03	-0.10	-0.09
	Item2	0.42	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00
	Item3	0.78	0.01	0.01	0.03	0.02	0.03	0.03	0.07	0.06
	Item4	2.00	0.01	-0.01	0.03	0.01	0.06	0.03	0.10	0.08
	Item5	-0.84	0.00	0.00	-0.06	-0.06	-0.04	-0.04	-0.09	-0.09
	Item6	1.47	0.02	0.02	0.04	0.04	0.05	0.05	0.13	0.13
	Item7	-0.59	0.00	0.00	-0.02	-0.02	-0.03	-0.03	-0.06	-0.06
	Item8	1.01	0.01	0.01	0.03	0.03	0.04	0.04	0.08	0.08
	Item9	1.00	0.01	0.01	0.02	0.02	0.03	0.03	0.08	0.08
	Item10	1.94	0.02	0.03	0.15	0.16	0.07	0.08	0.27	0.28
	Item11	-2.24	-0.04	-0.03	-0.25	-0.24	-0.07	-0.07	-0.32	-0.31
	Item12	2.38	0.01	-0.01	0.03	0.01	0.06	0.04	0.16	0.13
	Item13	0.79	0.00	-0.01	0.02	0.00	0.02	0.01	0.01	0.00
	Item14	2.22	0.01	0.02	0.13	0.14	0.05	0.07	0.31	0.32
	Item15	1.93	0.01	0.02	0.07	0.08	0.04	0.05	0.20	0.21
	Item16	-0.37	0.01	0.01	0.00	0.00	-0.01	-0.01	-0.04	-0.04
	Item17	-0.99	0.00	0.01	-0.04	-0.03	-0.03	-0.03	-0.09	-0.08
	Item18	1.00	0.00	0.01	0.01	0.02	0.02	0.02	0.07	0.08
	Item19	-2.62	-0.02	-0.01	-0.30	-0.28	-0.03	-0.03	-0.37	-0.35
	Item20	-0.35	-0.01	-0.01	-0.02	-0.01	-0.03	-0.02	-0.06	-0.05
MAD	NA	NA	0.01	0.01	0.06	0.06	0.04	0.04	0.13	0.13
Item discrimination	Item1	0.83	0.00	0.00	-0.07	-0.07	-0.02	-0.02	-0.04	-0.04
	Item2	0.61	0.01	0.01	-0.03	-0.02	-0.02	-0.01	-0.01	-0.01
	Item3	1.16	-0.01	-0.01	-0.08	-0.08	-0.06	-0.06	-0.08	-0.09
	Item4	0.59	0.00	0.01	-0.03	-0.02	-0.02	-0.01	-0.01	-0.01
	Item5	1.88	-0.01	-0.02	-0.37	-0.37	-0.12	-0.13	-0.25	-0.26
	Item6	1.28	-0.01	-0.01	-0.10	-0.11	-0.05	-0.06	-0.11	-0.11
	Item7	0.94	0.00	0.00	-0.08	-0.07	-0.03	-0.03	-0.05	-0.05
	Item8	0.91	-0.01	0.00	-0.06	-0.06	-0.04	-0.04	-0.05	-0.05
	Item9	1.56	-0.02	-0.02	-0.13	-0.14	-0.09	-0.09	-0.14	-0.14
	Item10	2.50	0.01	-0.04	-0.67	-0.69	-0.23	-0.28	-0.59	-0.61
	Item11	0.91	-0.01	-0.01	-0.15	-0.14	-0.02	-0.02	-0.09	-0.08
	Item12	0.61	0.00	0.01	-0.03	-0.02	-0.01	-0.01	-0.02	-0.01
	Item13	0.51	0.00	0.01	-0.02	-0.01	-0.01	-0.01	-0.01	0.00
	Item14	1.66	0.00	-0.02	-0.30	-0.31	-0.07	-0.09	-0.29	-0.30
	Item15	1.63	-0.01	-0.02	-0.22	-0.23	-0.06	-0.08	-0.21	-0.22
	Item16	0.73	0.00	0.01	-0.04	-0.04	-0.02	-0.02	-0.03	-0.02
	Item17	0.80	0.01	0.01	-0.06	-0.06	-0.02	-0.02	-0.03	-0.03
	Item18	2.20	0.02	0.00	-0.23	-0.25	-0.13	-0.15	-0.24	-0.25
	Item19	0.86	0.00	0.00	-0.14	-0.13	0.00	0.00	-0.08	-0.07
	Item20	0.60	0.00	0.00	-0.04	-0.03	-0.02	-0.01	-0.02	-0.01
MAD	NA	NA	0.01	0.01	0.14	0.14	0.05	0.06	0.12	0.12

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

item. Lastly, the mean absolute deviation (MAD) values are presented at the bottom of each panel. On the top of the last four columns, the sample size used for estimating item parameters from the data sets with removed misfitting response patterns using person fit indices are reported. The slight difference between sample sizes for the same person fit, for example l_z, is due to the difference between ML and BM estimation methods.

The next eight columns are in pairs to determine if there is a difference between the ML and BM estimators and then across the pairs to see how the values of the estimates are influenced by the presence and removal of students with misfitting response patterns. Looking at the difficulty of the first item, the true value of b is -1.05 and the bias for b is the same or within 0.01 within each pair but varies in magnitude across the pairs: 0.00 for fully fitting responses (MLfit and BMfit), -0.04 (sample including all students with misfitting response patterns (MLmisfit and BMmisfit), -0.04 and -0.03 when students identified by the l_z were removed (LzML and LzBM, respectively), and -0.10 and -0.09 when the students identified with misfitting response patterns by H^T were removed (HTML and HTBM, respectively). The true value of a for Item 1 is 0.83 and the bias is 0.00 for both the MLfit and BMfit estimates, -0.07 for MLmisfit and BMmisf, -0.02 for LzML and LzBM estimates, and -0.04 for HTML and HTBM.

b parameter: There were five items with estimation bias beyond ± 0.20 in the b parameter. Bias was present for Items 10, 11, 14, and 19 for the data set in which all the misfitting response patterns were included and the data set in which the misfitting response patterns were removed using H^T. For Items 10 and 14, the bias was greater when misfitting response patterns were removed using H^T than when all the misfitting response patterns were included. For Items 11 and 19, the values of the biases were more comparable. Lastly, the bias in b for Item 15 was beyond ± 0.20 only for the data set for which the misfitting response patterns were removed using H^T. All five items had large absolute true *b* values. For example Item 19 had the smallest true *b* value (-2.62) and Item 14 had the highest true *b* value (2.22). Lastly, whereas items with large negative *b* parameters tended to be underestimated (e.g., Item 19), items with large positive *b* parameters tended to be overestimated (e.g., Item 14).

a parameter: For the *a* parameter, there were five items with estimation bias beyond ± 0.20. The bias for Item 10 was the largest for the data set with all the misfitting response patterns and the two data sets in which misfitting response patterns were removed. For Items 5, 14, 15 and 18, negative bias in the *a* parameter estimates was present for the data set with all the misfitting response patterns and the data set in which the misfitting response patterns were removed using H^T. All five items had large *a* values, with Item 10 with the largest with true *a* value of 2.50 and Item 15 with the lowest with true *a* value of 1.63.

Both b and a parameters: Of the five items identified with bias for *b* and five items identified with bias for *a*, three were in common. The three items were Item10 (true *b* of 1.94, true *a* of 2.50), Item14 (true *b* of 2.22. true *a* of 1.66), and Item15 (true *b* of 1.93, true *a* of 1.63). What is common among these three items is that they were among the items that had both relatively high (in absolute value) true *b* values and true *a* values.

Figure 5 shows a graphical representation of results in Table 2. The horizontal dotted line represents bias estimates in the range of ± 0.10 and the dashed line represents bias estimates outside of ± 0.10 and less than or equal to ± 0.20. The upper graph represents results for the *b* parameter and lower graph represents results for the *a* parameter. Items on the x-axis are sorted increasingly in terms of the values of true *b* and true *a* and are presented from left to right. Items that had at least one bias value outside of ± 0.20 are labeled with their corresponding true value. As shown in Table 1, the results reveal that the difference between estimated bias derived using

Maximum Likelihood estimation and Bayes Modal estimation of item parameters were, at most, within $|0.10|$ for each of the four data sets. Further, this was the case for the remaining 17 simulation conditions. Consequently, the line graphs for the *b* and *a* parameter estimates for each estimation method for each of the four data sets were essentially the same line. Therefore, to facilitate the reading and interpretation of the graphs, only the results derived from Maximum Likelihood estimation are presented in the graphs for the simulation study.

b parameter. As it can be seen from Figure 5, the bias of estimation for the *b* parameter occurred mostly in the tails of the distribution of the *b* parameters. As mentioned in Chapter 3, a uniform distribution was used to generate item parameters. Consequently, there are an equal number of easy and hard items. Items with *b* parameters in the middle of distribution are less biased for the data set with all the misfitting response patterns included, the data set after misfitting response patterns identified using l_z were removed, and the data set after the misfitting response patterns identified using H^T were removed. Looking at the left tail, whereas the *b* parameters estimated using the data set with misfitting response patterns and the data set with misfitting response patterns removed by H^T have estimation bias beyond ± 0.20, the *b* parameters estimated from data set with misfitting response patterns removed by l_z have estimation bias in the range of ± 0.10. The same thing is true for the right tail. For both tails, the H^T resulted in estimation bias larger than the misftting data set. The large estimation bias in *b* parameter happened for items with *b* values less than -2.24 and greater than 1.92. As mentioned above, whereas items with large negative *b* parameters tended to be underestimated (e.g., Item 19), items with large positive *b* parameters tended to be overestimated (e.g., Item 14).

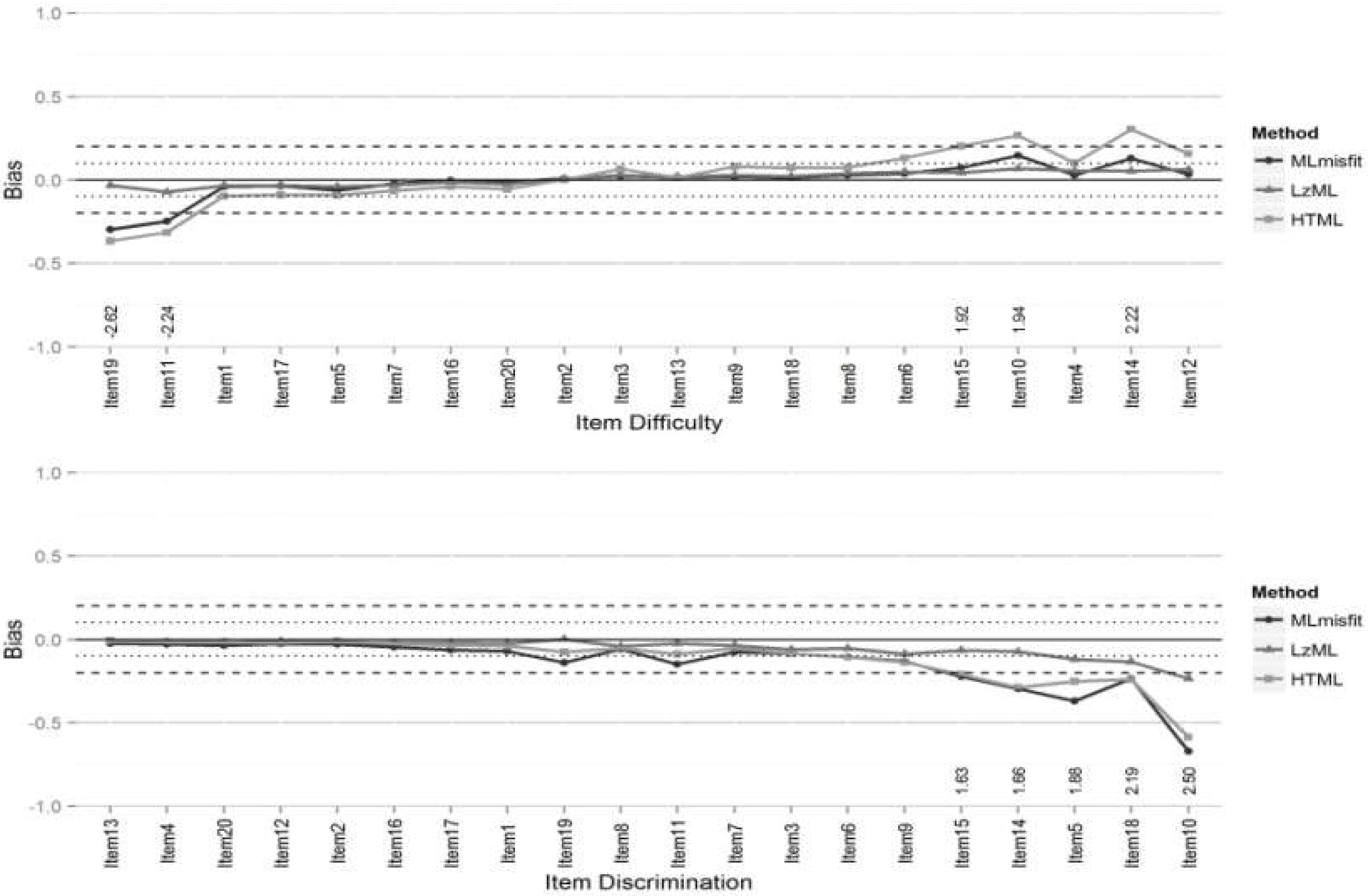

Figure 5: Bias of estimation for manipulating 25% of items and 10% of sample for test with 20 items

a parameter. In case of the *a* parameter, the majority of the items up to Item 19 had estimation bias in the range of ± 0.10 for all three data sets with two exceptions. For Items 11 and 19, the bias for the misfitting data set was less than -0.10 but beyond -0.20. The bias for the data set with all misfitting items and the data set with misfitting items removed using H^T next three items were less than -0.20. Lastly, the bias of the last item for all three data sets was less than -0.20.

Comparison a and b parameters. For both the *b* parameter estimates and the *a* parameter estimates estimated using data set with misfitting response patterns and data set with misfitting response patterns removed by H^T have estimation bias beyond ± 0.20 for a greater number of items than when the *b* and *a* estimates were obtained using the data set with misfitting response patterns removed by l_z. Further, the degree of bias tends to be greater for *a* than for *b*, particularly for Item 10.

Table 3 provides a summary of biases across data sets and item parameters in terms of number of items in each of the three ranges of bias as equal to or between -0.10 and 0.10, -0.20 to -0.10 or 0.10 to 0.20 and biases that are either smaller than -0.20 or larger than 0.20. As stated in Chapter 3, there are no determined criteria on how large item parameter estimation bias should be to be considered severe. Therefore, for the purposes of this study, the above mentioned categories of bias were used to assess the extent of change in estimation bias .

Based on the results in Table 3, there are a greater number of highly biased estimates of item discrimination than highly biased estimates of item difficulty (i.e., 5 vs 2) for the misfitting data set. Using the l_z data set resulted in fewer biased *b* and *a* item parameter estimates than using the H^T data set. For example, whereas the number of items with bias beyond ±0.20 in using

l_z is 0 for the b parameter and 1 for the a parameter, the number of items with bias beyond ±0.20 using H^T is 5 for both the b and a parameters for both ML and BM esrtimates.

Table 3
Bias of estimation across different methods (25% of items and 10% of sample) for test with 20 items

		MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
Item difficulty	-0.10 to 0.10	16	16	20	20	12	13
	±0.20 to ±0.10	2	2	0	0	3	2
	Beyond ±0.20	2	2	0	0	5	5
Item discrimination	-0.10 to 0.10	11	11	17	17	13	13
	±0.20 to ±0.10	4	4	2	2	2	2
	Beyond ±0.20	5	5	1	1	5	5

Note: ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set usingl_z; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T.

The values of the mean absolute deviation (MAD), which provide a summary across the 20 items, are provided in the last row of each panel in Table 2. The value of MAD was 0.01 for both MLfit and BMfit, 0.06 for both MLmisfit and BMmisfit, 0.04 for LzML and 0.06 for LzBM, and 0.13 for both HTML and HTBM for the b parameter. The corresponding values for the a parameter were 0.01 for both MLfit and BMfit, 0.14 for both MLmisfit and BMmisfit, 0.05 for LzML and 0.06 for LzBM, and 0.12 for both HTML and HTBM. Overall, the data sets with students with misfitting response patterns removed using l_z had smaller values of MAD for both the a and b parameters than the data sets with students with misfitting response patterns removed using H^T.

Classification accuracy. In order to investigate to what extent the bias in b and the bias in a estimated parameters influenced the estimate of examinee abilities, the classification accuracy of placing students in one of two classes was examined. Students with true $\theta < 1.00$ were classified as "low" performing and students with true $\theta \geq 1$ were classified as "high" performing. The same classification was used for the estimated θ values. The total sample size was 5,000 students. The 2 x 2 classification contingency tables were developed with the frequency of students in each cell for each of the 100 simulations. The mean number of students for each cell was then calculated across 100 replications. The values were then rounded to the nearest whole number using scientific rounding. This procedure resulted in total number of students of 4999 for some conditions, 5000 for some conditions, and 5001 for the remaining conditions.

The classifications are reported in Table 4 in terms of row percentages. The rows correspond to the true ability parameters and are therefore the true placements. For the fitting data set (Fit), the item and ability parameters were estimated from data with no misfitting response patterns. The item and ability parameters for the misfitting data set (Misfit) were estimated using the data set that contained all the students with misfitting response patterns. The item parameters for the l_z data set were estimated using the data set in which students with misfiting response patterns using the l_z procedure. Lastly, the item parameters for the the H^T data set were estimated using the data set in which students with misfitting response patterns were removed using the H^T procedure. The estimated item parameters derived from each of the four data sets were the used to estimate the abilities of the sample of 5000 or so students used to determine the classification indices.

Table 4
Classification accuracy (25% of items and 10% of sample) for test with 20 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4067 (96.6)	142 (3.4)	4059 (96.4)	151 (3.6)	4063 (96.5)	147 (3.5)	4058 (96.4)	152 (3.6)
	High	266 (33.7)	524 (66.3)	306 (38.7)	485 (61.3)	309 (39.1)	481 (60.9)	305 (38.6)	486 (61.4)
% of total agreement		91.8		90.9		90.9		90.9	
BM	Low	3415 (81.1)	794 (18.9)	4058 (96.4)	151 (3.6)	4062 (96.5)	147 (3.5)	4058 (96.4)	152 (3.6)
	High	0 (0)	790 (100)	306 (38.7)	485 (61.3)	309 (39.1)	482 (60.9)	305 (38.6)	486 (61.4)
% of total agreement		84.1		90.9		90.9		90.9	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfi,t,, l_z, and H^T.

As shown in Table 4, of the 4,209 students classified as low performing using the true ability parameter, 96.6% were classified as low performing using the ML item parameter estimates derived from the Fit data set and 3.4% were classified as false positives (FP; low ability student classified as high ability). Of the 790 students classified as high performing using the true ability parameter, 66.3% were classified as high performing using the ML item parameter estimates derived from the Fit data set and 33.7% were classified as false negatives (FN; high ability student classified as low ability). While the overall classification percentage for this data set was 91.8%, clearly the accuracy was greater for the low performing students than for the high performing students.

Examination of the eight classification tables included in Table 4 reveals that the percentage of correct decisions for the students with true low ability were essentially constant across the four data sets (either 96.4% or 96.5%). In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to essentially a constant (approximately 61% across the remaining three data sets). Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions except for BM with fitting data set for which there were not false negative placements. Lastly, with ML estimator the overall classification percentage was highest for the Fit data set and constant for the remaining data sets (90.9%); in contrast, with the BM estimator, the overall classification percentage was lowest for the Fit data set (84.1%) and constant for the remaining data sets (90.9%).

As can be seen from the results provided, the results obtained from the ML estimator and the results obtained from the BM estimator are, with one exception, very similar for this simulation and also the remaining 17 simulation conditions. As mentioned above, the absolute value of the differences between the bias obtained using the ML estimator and the bias obtained using the BM estimator were within the absolute value of 0.10. Likewise, as seen in Table 3, the numbers of biased items in each size interval are the same for each estimator. Where the results did differ is in Table 4 for the fitting data. Whereas with the ML estimator, there were both false positive and false negative placements for all simulation conditions for the fitting data set, with the BM estimator there were no false negative placements for all simulation conditions for the fitting data set. Therefore, given the high similarity between the results obtained using the ML estimator and the results obtained using the BM estimator and the same classification difference for the fitting data set across simulations conditions, only the results for the ML estimator are provided and discussed in the text. Appendix A contains tables like Table 1 for each of the remaining 17 simulation conditions in which the BM results are reported.

20 items, 20% of students with misfitting response patterns in 25% of items

The results for the 20 item test with 20% of the students with misfitting responses to 25% of the items are presented graphically in Figure 6. The full set of results are reported Table A1 in Appendix A.

b parameter: As for the previous case, the bias of estimation for the *b* parameter occurred in the tails of the distribution of *b* parameters. Of the items with bias, three items, all which had the highest positive true *b* parameters, had estimation bias beyond ± 0.20 across all three data sets. The largest bias for these three items occurred for the data set with all the misfitting response patterns and the smallest bias occurred for the data set data set with misfitting response patterns removed by l_z . Looking at the left tail, the five items with the lowest values of the *b* parameters estimated using the data set with misfitting response patterns removed by H^T had estimation bias beyond ± 0.20. In contrast, the b parameters estimated from data set with all the misfitting response patterns and data set with misfitting response patterns removed by l_z had estimation bias in the range of ± 0.10 except Items 3, 8 and 17, which had estimation bias between -0.10 and -0.20. Lastly, whereas items with large negative b parameters tended to be underestimated (e.g., Item 8), items with large positive b parameters tended to be overestimated to a greater degree (e.g., Item 10).

a parameter: The a parameter was underestimated, with one exception, across all three data sets for the 12 items with the largest true a parameters. The bias in a for these 12 items was less than -0.20 except for Item 15 for the data set with misfitting response patterns removed by l_z. For the remaining 11 items, whereas the bias in a was less for the data set with misfitting response patterns removed by l_z, the bias for the other two data sets were comparable.

Both b and a parameters: Of the eight items identified with high bias for b and the 12 items identified with high bias for a, seven items were in common. The seven items were Item 3 (true b of -1.51, true a of 2.13), Item 8 (true b of -1.72; true a of 1.97), Item 10 (true b of 1.92; true a of 2.44), Item 15 (true b of -1.34;. true a of 1.45), Item 17 (true b of -1.66; true a of 1.81), Item 19 (true b of -1.24;. true a of 2.09), and Item 20 (true b of 2.16; true a of 1.41). What is common to these items is that they were among the items that had both relatively high (in absolute value) true b and true a values.

Comparison a and b parameters. In comparison to the b parameter, a greater number of items with marked bias in the a parameter due to presence of misfitting response patterns were

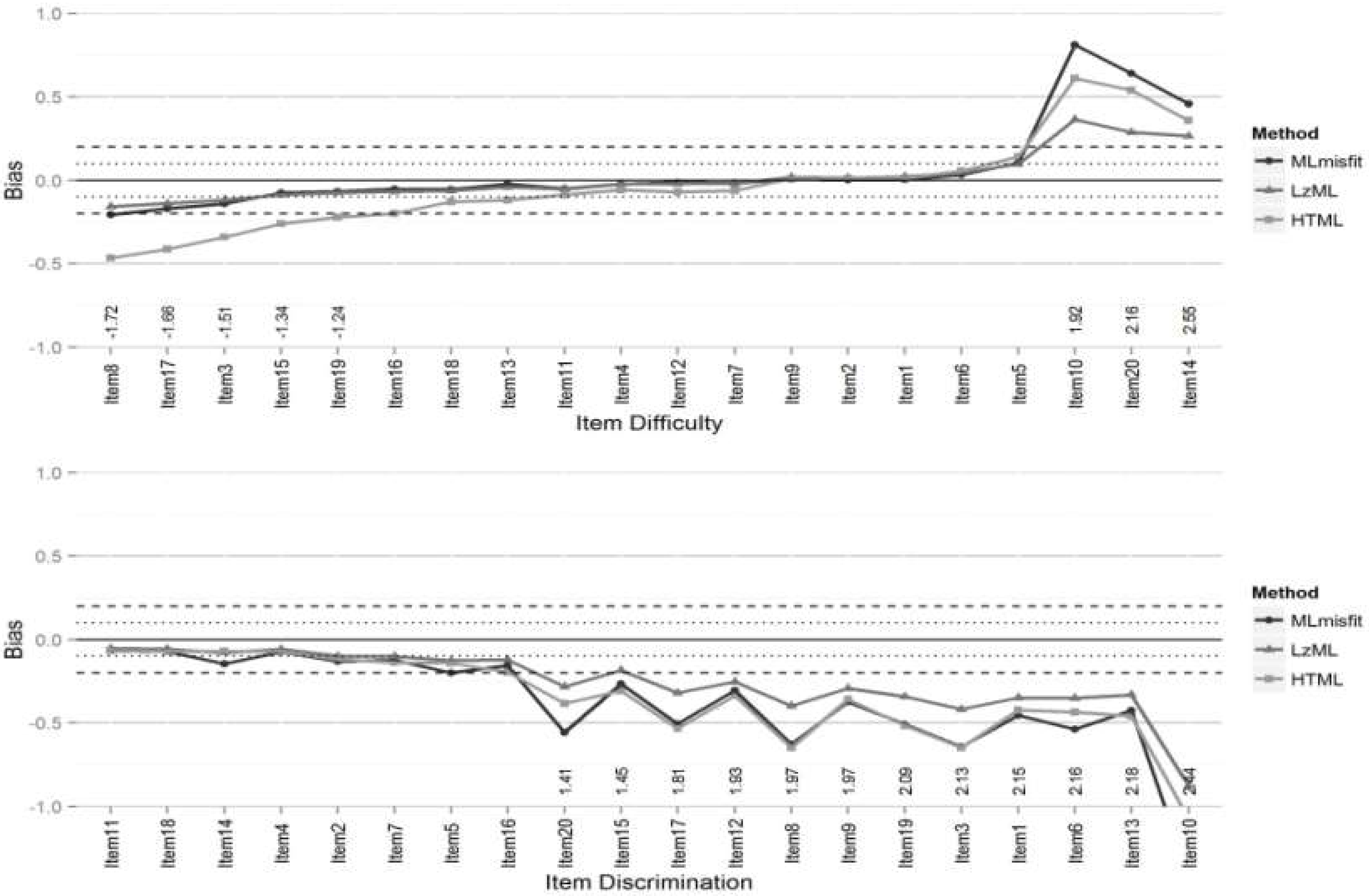

Figure 6: Bias of estimation for manipulating 25% of items and 20% of sample for test with 20 items

found. Further, the bias was larger for the a parameter than for the b parameter except for the b parameters for the three items with the largest true b. The bias for the b parameter estimated using the data set with misfitting response patterns removed by H^T tended to be the largest except for the b parameters for the three items with the highest true b values. In contrast the bias in a was approximately the same for the data set with all the mifitting response patterns and the data set with misfitting response patterns removed by H^T. For both the b parameter and the a parameter, the bias estimated using the data set with misfitting response patterns removed by l_z tended to be the smallest.

The results in Table 5 reveal that there were more highly biased estimates of item discrimination parameter in misfitting data set (12) compared to difficulty parameter (4). The use of l_z to remove misfitting response patterns resulted in fewer biased item parameters than H^T. For example, whereas the number of items with bias beyond ±0.20 in using l_z is 3 for the b parameter and 11 for the a parameter, the number of items with bias beyond ±0.20 using H^T is 8 for the b parameter and 12 for the a parameter.

Table 5
Bias of estimation across different methods (25% of items and 20% of sample)for test with 20 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	13	14	8
	±0.20 to ±0.10	3	3	4
	Beyond ±0.20	4	3	8
Item discrimination	-0.10 to 0.10	3	5	4
	±0.20 to ±0.10	5	4	4
	Beyond ±0.20	12	11	12

The values of the mean absolute deviation (MAD) for the b parameter were 0.01 for fitting data set, 0.15 for the data set with all the misfitting response patterns, 0.10 for the data set in which students with misfitting response patterns were removed using l_z, and 0.21 for the data set in which students with misfitting response patterns were removed using H^T (see Table A1, Appendix A). The corresponding values of MAD for the *a* parameter were 0.01, 0.38, 0.26, and 0.35. Overall, the data sets in which students with misfitting response patterns were removed using the l_z provided smaller values of MAD for both the *b*and *a* parameters.

Classification accuracy. The values of the classification accuracy are reported in Table 6. For the fitting data, 96.4% of the 4,202 students classified as low performing using their true ability parameters were classified as low performing using the item parameter estimates derived from the Fit data set and 3.6% were classified as false positives. Of the 798 students classified as high performing using the true ability parameters, 64.8% were classified as high performing using the item parameters derived from the Fit data set and 35.2% were classified as false negative. While the overall classification percentage for this data set was 91.4%, the accuracy was greater for the low performing students than the high performing students.

Table 6
Classification accuracy (25% of items and 20% of sample) for test with 20 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4051 (96.4)	151 (3.6)	4027 (95.9)	174 (4.1)	4025 (95.8)	177 (4.2)	4022 (95.7)	180 (4.3)
	High	281 (35.2)	517 (64.8)	356 (44.6)	442 (55.4)	356 (44.6)	442 (55.4)	353 (44.2)	445 (55.8)
% of total agreement		91.4		89.4		89.3		89.3	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

Examination of the four classification tables included in Table 6 reveals that the percentages of correct decisions for the students with true low ability were essentially constant across the four datasets (approximately 96%). In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to essentially a constant (approximately 55% across the remaining three data sets). Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (89.4% and 89.3%).

20 items, 30% of students with misfitting response patterns in 25% of items

The results for the 20 item test with 30% of the students with misfitting responses to 25% of the items are presented graphically in Figure 7. The full set of results are reported Table A2 in Appendix A.

b parameter: As for the two previous conditions, the bias of estimation for the *b* parameter occurred in the tails of the distribution of *b* parameters. Of the items with bias, the four items with the highest positive true *b* parameters had estimation bias beyond ± 0.20 across all three data sets. The largest bias for these four items occurred for the data set with all the misfitting response patterns and the smallest bias occurred for the data set with misfitting response patterns removed by l_z . Looking at the left tail, the four items with the lowest negative true *b* parameters had estimation bias beyond ± 0.20 across all three data sets. Again, the largest bias for these four items occurred for the data set with all the misfitting response patterns. However, in contrast to the four items in the right tail, the *b* parameters estimated from data set with misfitting response patterns removed by l_z and H^T were more similar. As before, whereas items with large negative *b* parameters tended to be underestimated (e.g., Item 12), items with large positive *b* parameters tended to be overestimated (e.g., Item 1).

a parameter: The bias in *a* parameter was within ±0.20 for only Items 13, 6, and 14. These three items had the lowest, second lowest, and fourth lowest values of *a*. The remaining 17 items had at least one of the three data sets for which the bias in *a* was smaller than -0.20. Further, the bias for all 17 items tended to increase as *a* increased. Whereas the bias in *a* for

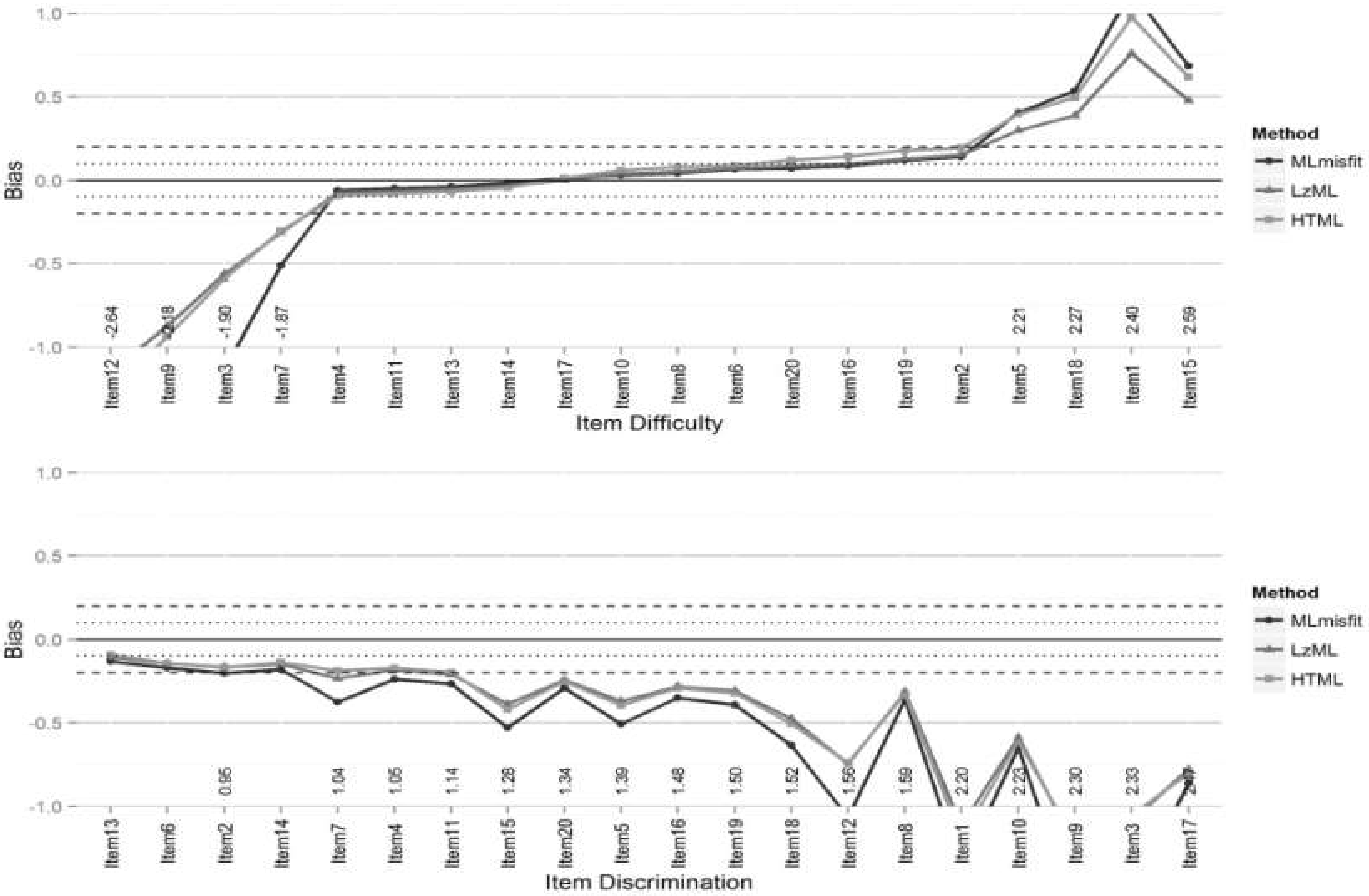

Figure 7: Bias of estimation for manipulating 25% of items and 30% of sample for test with 20 items

these 17 items was more pronounced for the data set with all the misfitting response patterns, the biases for the other two data sets were smaller and more comparable in size.

Both b and a parameters: Of the eight items identified with high bias for *b* and the 17 items identified with high bias for *a*, eight items were in common. The eight items were Item 1 (true *b* of 2.40, true *a* of 2.20), Item 3 (true *b* of -1.90; true *a* of 2.33), Item 5 (true *b* of 2.21; true *a* of 1.39), Item 7 (true *b* of -1.87;. true *a* of 1.04), Item 9 (true *b* of -2.18; true *a* of 2.30), Item 12 (true *b* of -2.64;. true *a* of 1.56), Item 15 (true *b* of 2.59;. true *a* of 1.28), and Item 18 (true *b* of 2.27; true *a* of 1.52). As for the previous conditions, these items that had both relatively high (in absolute value) true *b* and true *a* values.

Comparison a and b parameters. In comparison to the *b* parameter, a greater number of items with marked bias in the *a* parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the *a* parameter than for the *b* parameter. The bias for the *b* parameter estimated using the data set with misfitting response patterns removed by l_z tended to be the smallest for the four items with high positive *b* values but the largest for the four items with the lowest *b* parameters. In contrast the bias in *a* was approximately the same for the data set with misfitting response patterns removed by l_z and H^T.

The results in Table 7 reveal that there were a greater number of highly biased estimates of item discrimination parameter than item difficulty parameters for each of the data sets (17 vs. 8; 15 vs. 8, 13 vs. 8). Clearly, the *a* parameter was more affected by the increase in the percentage of students with misfitting response patterns than the *b* parameter.

Table 7
Bias of estimation across different methods (25% of items and 30% of sample)for test with 20 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	10	9	8
	±0.20 to ±0.10	2	3	4
	Beyond ±0.20	8	8	8
Item discrimination	-0.10 to 0.10	0	0	1
	±0.20 to ±0.10	3	5	6
	Beyond ±0.20	17	15	13

The values of the mean absolute deviation (MAD) for the b parameter were 0.01 for fitting data set, 0.49 for the data set with all the misfitting response patterns, 0.28 for the data set in which students with misfitting response patterns were removed using l_z, and 0.34 for the data set in which students with misfitting response patterns were removed using H^T (see Table A2). The corresponding values of MAD for the a parameter were 0.01, 0.59, 0.46, and 0.46. Overall, data sets in which students with misfitting response patterns were removed using the l_z yielded smaller values of MAD for both the b and a parameters.

Classification accuracy. The values of the classification accuracy are reported in Table 8. For the fitting data, of the 4,212 students classified as low performing using their true ability parameters, 96.8% were classified as low performing using the item parameter estimates derived from the Fit data set and 3.2% were classified as false positives. Of the 788 students classified as high performing using the true ability parameters, 69.5% were classified as high performing using the item parameters derived from the Fit data set and 30.5% were classified as false negative. While the overall classification percentage for this data set was 92.5%, the accuracy was again greater for the low performing students than the high performing students.

Examination of the four classification tables included in Table 8, the percentage of correct decisions for the students with true low ability were essentially constant across the four data sets (approximately 96.6%). In contrast the percentage of correct decisions for the students

Table 8
Classification accuracy (25% of items and 30% of sample) for test with 20 items

		Fit		Misfit		l_z		H^T	
	True \ Estimated	Low	High	Low	High	Low	High	Low	High
ML	Low	4077 (96.8)	135 (3.2)	4073 (96.7)	139 (3.3)	4066 (96.5)	146 (3.5)	4061 (96.4)	151 (3.6)
	High	240 (30.5)	548 (69.5)	359 (45.6)	428 (54.4)	358 (45.4)	430 (54.6)	355 (45.1)	433 (54.9)
% of total agreement		92.5		90.0		89.9		89.9	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

with true high ability dropped from the percentage for the fitted data to essentially a constant (approximately 54.7% across the remaining three data sets). Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (89.9% and 90.0%).

20 items, 10% of students with misfitting response patterns in 50% of items

The results for the 20 item test with 10% of the students with misfitting responses to 50% of the items are presented graphically in Figure 8. The full set of results are reported Table A3 in Appendix A.

b parameter: In contrast to the previous three conditions, no item in each of the three data sets had bias beyond ±0.20. For all 20 items, the bias was within the interval ±0.20. Items 3 and 9, which had the largest positive *b* parameters, had positive bias between 0.10 and 0.20 for data sets with all the misfitting response patterns and the data set with misfitting response patterns removed by H^T. Items 7 and 14, which had the lowest negative true *b* parameters, had positive bias between 0.10 and 0.20 for data sets with all the misfitting response patterns and data set with misfitting response patterns removed by l_z. In contrast, the *b* parameter for Item 7

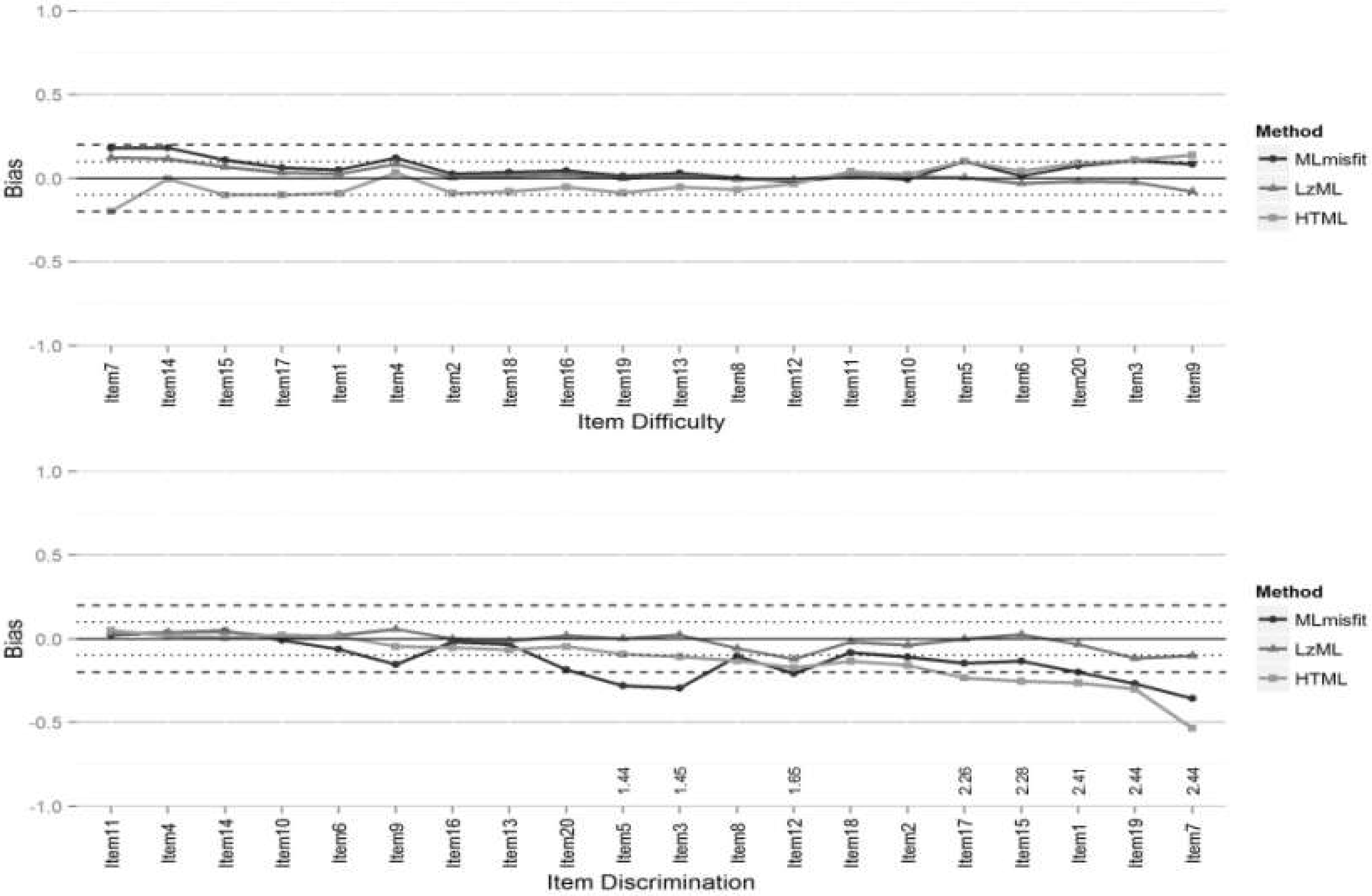

Figure 8: Bias of estimation for manipulating 50% of items and 10% of sample for test with 20 items

estimated from data set with misfitting response patterns removed by H^T had negative bias was equal to -0.20. Whereas there was a tendency for positive bias for items with the lowest b parameters for the data set with all the misfitting response patterns and the data set with misfitting response patterns removed by l_z and negative bias for the data set with misfitting response patterns removed by H^T, there was tendency for positive bias for the items with the highest *b* parameters for the data set with all the misfitting response patterns and the data set with misfitting response patterns removed by H^T and negative bias for the data set with misfitting response patterns removed by l_z.

a parameter: Only eight items had bias less than -0.20 and three of these items were in the middle of the distribution and five were in the right tail of the distribution.. For items 3, 5 and 12, which were in the middle of the distribution, had estimation bias that was -0.20 or below for data set with all the misfitting response patterns. For the items in the right tail, Items 1, 7, 15, 17 and 19 had estimation bias below -0.20 for the data set with misfitting response patterns removed by H^T. Items 17 and 19, also had bias below -0.20 for data set with all the misfitting response patterns and data set.

Both b and a parameters: There was no common items with estimation bias on both the *b* and *a* parameters. bias for a,.

Comparison a and b parameters. There were a greater number of items with marked bias in the *a* parameter than the *b* parameter due to presence of misfitting response patterns. Further, the bias was larger for the *a* parameter than for the *b* parameter. Using l_z to remove misfitting response patterns resulted in fewer biased estimated item parameters than using H^T for *b* and *a* parameters. The number of items with bias in *a* was approximately the same for data set with all the misfitting response patterns and the data set with misfitting response patterns removed by H^T.

The results in Table 9 reveal that there were more highly biased estimates of the item discrimination parameter in the misfitting data set (5) compared to the difficulty parameter (0). The use of l_z to remove misfitting response patterns resulted in fewer biased item parameters than H^T for the *b* parameter and, particularly, the *a* parameter. For example, whereas the number of items with bias beyond ±0.20 in using l_z is 0 for both the *b* and *a* parameters, the number of items with bias beyond ±0.20 using H^T is 0 for the *b* parameter and 5 for the *a* parameter.

Table 9
Bias of estimation across different methods (25% of items and 20% of sample)for test with 20 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	14	18	16
	±0.20 to ±0.10	6	2	4
	Beyond ±0.20	0	0	0
Item discrimination	-0.10 to 0.10	8	17	10
	±0.20 to ±0.10	7	3	5
	Beyond ±0.20	5	0	5

The values of the mean absolute deviation (MAD) for the *b* parameter were 0.01 for fitting data set, 0.06 for the data set with all the misfitting response patterns, 0.04 for the data set in which students with misfitting response patterns were removed using l_z, and 0.08 for the data set in which students with misfitting response patterns were removed using H^T (see Table A3, Appendix A). The corresponding values of MAD for the *a* parameter were 0.01, 0.14, 0.04, and 0.14. Overall, the data sets in which students with misfitting response patterns were removed using the l_z provided smaller values of MAD for both the *a* and *b* parameters.

Classification accuracy. The values of the classification accuracy are reported in Table 10. For the fitting data, 96.5% of the 4,208 students classified as low performing using their true ability parameters were classified as low performing using the item parameter estimates derived from the Fit data set and 3.5% were classified as false positives. Again, a smaller percentage, 63.6%, of the 791 students classified as high performing using the true ability parameters were

Table 10
Classification accuracy (25% of items and 30% of sample) for test with 20 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4059 (96.5)	149 (3.5)	3983 (94.7)	225 (5.3)	3976 (94.5)	232 (5.5)	3960 (94.1)	248 (5.9)
	High	288 (36.4)	503 (63.6)	351 (44.3)	441 (55.7)	350 (44.2)	442 (55.8)	335 (42.4)	456 (57.6)
% of total agreement		91.2		88.5		88.4		88.3	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

classified as high performing using the item parameters derived from the Fit data set and 36.4% were classified as false negative. While the overall classification percentage for this data set was 91.2%, the accuracy was greater for the low performing students than the high performing students.

The percentages of correct decisions for the students with true low ability are essentially constant across the four data sets, ranging from 94.1% to 94.7%. In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to essentially a constant (approximately 55.7% across the remaining three data sets). Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (88.3% to 88.5%).

20 items, 20% of students with misfitting response patterns in 50% of items

The results for the 20 item test with 20% of the students with misfitting responses to 50% of the items are presented graphically in Figure 9. The full set of results are reported Table A4 in Appendix A.

b parameter: In contrast to the previous case, the bias of estimation for the *b* parameter occurred in the tails of the distribution of *b* parameters. For the fight tail, on Item 13, which had the highest positive true *b* parameter, had estimation bias beyond ± 0.20 across all three data sets. The largest bias for this item occurred for the data set with all the misfitting response patterns and the smallest bias occurred for the data set with misfitting response patterns removed by H^T. Looking at the left tail, the five items with the lowest negative true *b* parameters had estimation bias beyond ± 0.20 for all three data sets. The largest bias for these items occurred for the data set with all the misfitting response patterns and the smallest bias occurred for the data set with the misfitting responses patterns removed with H^T. Lastly, and in contrast to the first three conditions, whereas items with large negative *b* parameters tended to be overestimated (e.g., Item 6), the item with large positive *b* parameters tended to be underestimated (e.g., Item 13).

a parameter: The *a* parameter was underestimatedfor Item 6 with true *a* value of 0.96 for the data set with misfitting response patterns removed using l_z. Then, the no bias for the next two items. Except for Items 4 and 14, The the items with the larger true values than 0.96 had bias in the *a* parameter for the data set with all the misfitting response patterns. The two items with the highest discrimination had bias for the data set with misfitting response patterns removed by H^T and the most discriminating item also had bias for the data set with response patterns removed using l_z.

Both b and a parameters: Of the six items identified with high bias for *b* and the 11 items identified with high bias for *a*, two items were in common. The items were Item 15 (true *b* of -2.10, true *a* of 2.20) and Item 6 (true *b* of -2.57, true *a* of 0.96). As before, the two items were among the items that had relatively high (in absolute value) true *b* and true *a* values.

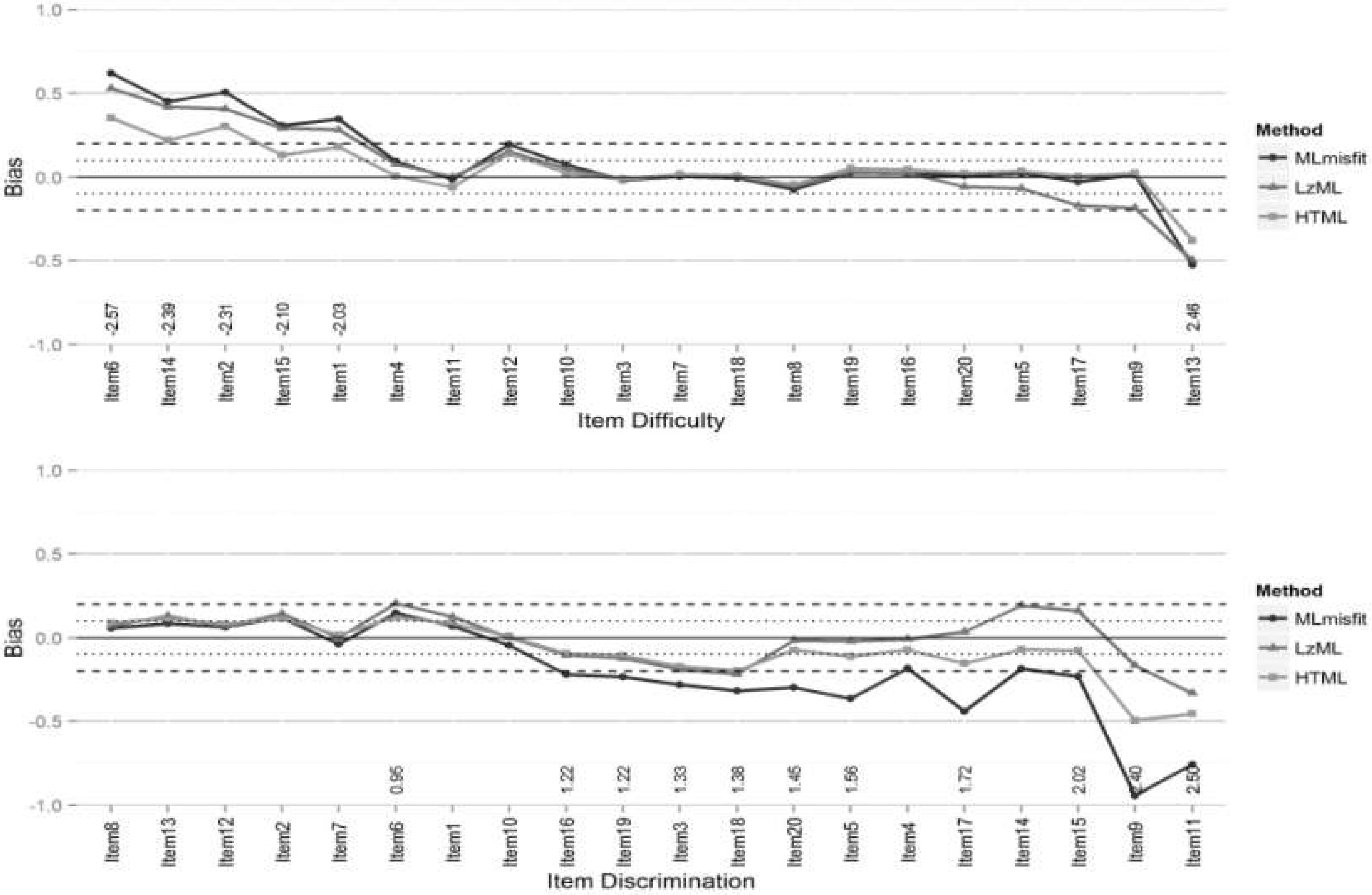

Figure 9: Bias of estimation for manipulating 50% of items and 20% of sample for test with 20 items

Comparison a and b parameters. A greater number of items with marked bias in the *a* parameter than in the *b* parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the *a* parameter than for the *b* parameter. The bias for the *b* parameter estimated using the data set with misfitting response patterns removed by H^T tended to be the smallest. In contrast the bias in *a* was approximately the same for the data set with misfitting response patterns removed by l_z and H^T, but the five items with biased *a* parameters and large true *a* values, H^T produced less biased estimates.

The results in Table 11 reveal that there were more highly biased estimates of the item discrimination parameter in misfitting data set (10) compared to the difficulty parameter (6). The use of H^T to remove misfitting response patterns resulted in fewer biased item parameters than the use of l_z for both the *b* and *a* parameters. For example, whereas the number of items with bias beyond ±0.20 in using l_z is 6 for the *b* parameter and 3 for the *a* parameter, the number of items with bias beyond ±0.20 using H^T is 4 for the *b* parameter and 2 for the *a* parameter.

Table 11
Bias of estimation across different methods (50% of items and 20% of sample)for test with 20 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	13	11	13
	±0.20 to ±0.10	1	3	3
	Beyond ±0.20	6	6	4
Item discrimination	-0.10 to 0.10	6	8	10
	±0.20 to ±0.10	4	9	8
	Beyond ±0.20	10	3	2

The values of the mean absolute deviation (MAD) for the *b* parameter were 0.02 for fitting data set, 0.17 for the data set with all the misfitting response patterns, 0.17 for the data set in which students with misfitting response patterns were removed using l_z, and 0.10 for the data set in which students with misfitting response patterns were removed using H^T (see Table A4, Appendix A). The corresponding values of MAD for the *a* parameter were 0.01, 0.25, 0.12, and

0.14. Overall, the data sets in which students with misfitting response patterns were removed using the H^T provided smaller values of MAD for the b parameter and data sets in which students with misfitting response patterns were removed using the l_z provided smaller values of MAD for the a parameter.

Classification accuracy. The values of the classification accuracy are reported in Table 12. For the fitting data, of the 4,211 students classified as low performing using their true ability parameters, 96.8% were classified as low performing using the item parameter estimates derived from the Fit data set and 3.2% were classified as false positives. Of the 798 students classified as high performing using the true ability parameters, 77.7% were classified as high performing using the item parameters derived from the Fit data set and 22.3% were classified as false negative. While the overall classification percentage for this data set was 93.8%, the accuracy was again greater for the low performing students than the high performing students.

Table 12
Classification accuracy (50% of items and 20% of sample) for test with 20 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4077 (96.8)	134 (3.2)	3963 (94.1)	248 (5.9)	3949 (93.8)	262 (6.2)	3915 (93.0)	296 (7.0)
	High	176 (22.3)	613 (77.7)	322 (40.8)	467 (59.2)	311 (39.4)	479 (60.6)	293 (37.1)	497 (62.9)
% of total agreement		93.8		88.6		88.6		88.2	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

For the four classification tables included in Table 12, the percentage of correct decisions for the students with true low ability varied slightly from 93.0% to 94.1% for the H^T and misfit data sets, respectively. In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to 59.2%, 60.6% and 62.9% for the misfit, l_z and H^T data sets, respectively. Thus, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (88.6% or 88.2%).

20 items, 30% of students with misfitting response patterns in 50% of items

The results for the 20 item test with 30% of the students with misfitting responses to 50% of the items are presented graphically in Figure 10. The full set of results are reported Table A5 in Appendix A.

b parameter: The bias of estimation for the *b* parameter occurred in the tails of the distribution of *b* parameters, with one item in each tail. Item 18, which had the highest positive true *b* parameter, had estimation bias beyond ± 0.20 for all three data sets. The largest bias for this item occurred for the data set with all the misfitting response patterns and the smallest bias occurred for the data set with misfitting response patterns removed by H^T. Item 10 with second smallest b parameter had estimation bias beyond ± 0.20 with the data set with all the misfitting response patterns and the data set with misfitting response patterns removed by l_z. Again, whereas Item 10 with a large negative true *b* parameter was overestimated (e.g., Item 10), Item 18 with a large positive true *b* parameter was underestimated.

a parameter: Except for the first 3 items with the lowest true *a* values, the *a* parameter was underestimated for at least one of the data sets. The bias in *a* for remaining 17 items was below -0.20 for the data set with the misfitting response patterns removed by H^T and, with one exception (Item 17), for the data set with misfitting response patterns removed by l_z. The bias was greatest for the data set with all the misfitting response patternsessentially equal for the two data setsin which misfitting response patterns were removed. Consequently, using l_z to remove misfitting response patterns resulted in slightly fewer biased estimated item parameters than using H^T specially for *a* parameter.

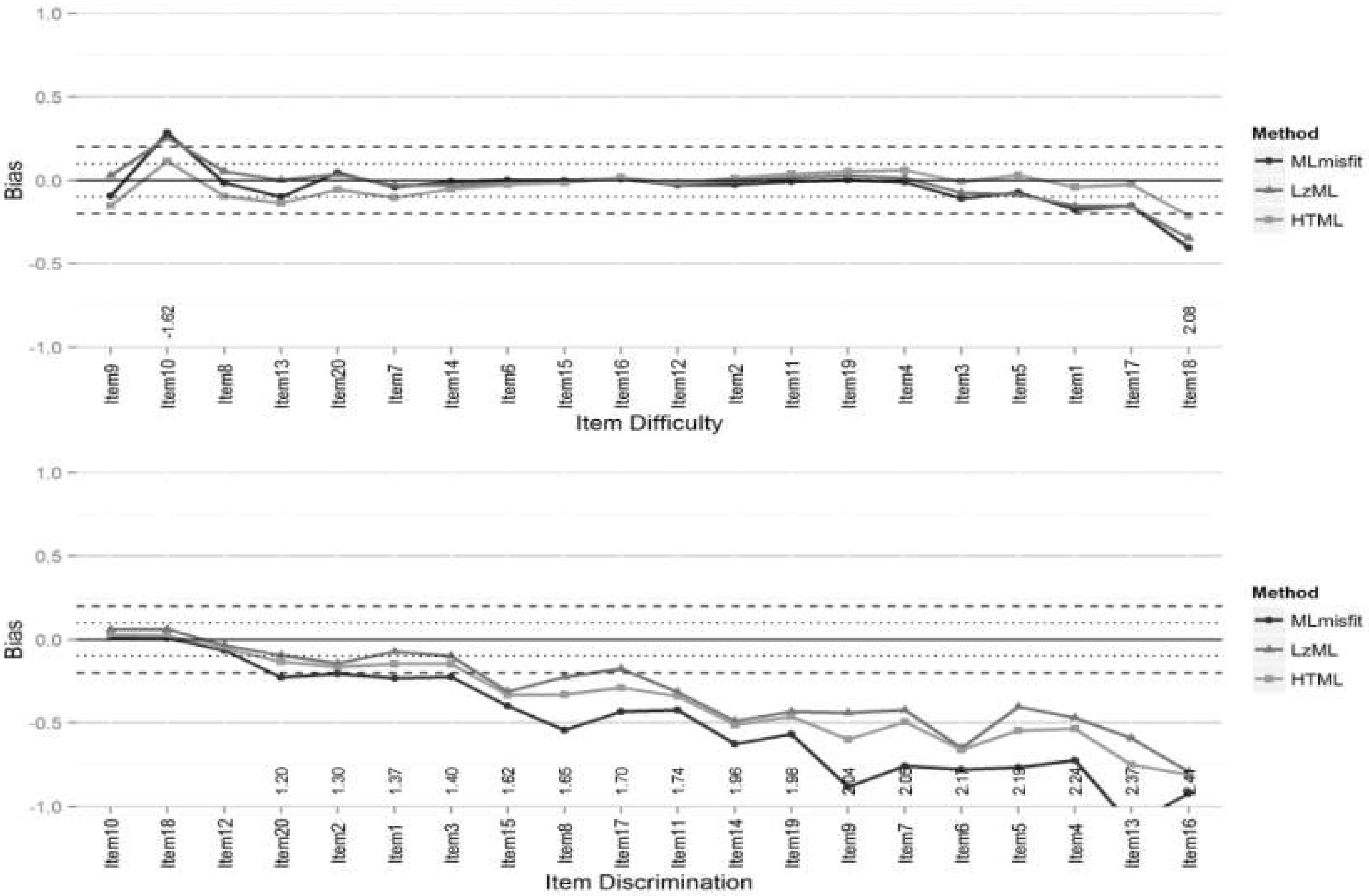

Figure 10: Bias of estimation for manipulating 50% of items and 30% of sample for test with 20 items

Both b and a parameters: There were no common items with estimation bias on both *b* and *a* parameters.

Comparison a and b parameters. A greater number of items with marked bias in the *a* parameter than marked bias in the *b* parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the *a* parameter than for the *b*. The bias for the *b* parameter estimated using the data set with misfitting response patterns removed by H^T tended to be the smallest especially for the one item with large positive *b* value. In contrast the bias in *a* was approximately the same or less for the data set with misfitting response patterns removed by l_z and H^T.

The results in Table 13 reveal that there were more highly biased estimates of item discrimination parameter than highly biased difficulty parameters) for all three data sets. The use of l_z to remove misfitting response patterns resulted in slightly fewer biased item parameters than H^T for both *b* and *a* parameter but, overall, results are comparable. For example, whereas the number of items with bias beyond ±0.20 in using l_z is 2 for the *b* parameter and 12 for the *a* parameter, the number of items with bias beyond ±0.20 using H^T is 1 for the *b* parameter and 13 for the *a* parameter.

Table 13
Bias of estimation across different methods (50% of items and 30% of sample)for test with 20 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	15	16	15
	±0.20 to ±0.10	3	2	4
	Beyond ±0.20	2	2	1
Item discrimination	-0.10 to 0.10	3	6	3
	±0.20 to ±0.10	0	2	4
	Beyond ±0.20	17	12	13

The values of the mean absolute deviation (MAD) for the *b* parameter were 0.01 for fitting data set, 0.08 for the data set with all the misfitting response patterns, 0.07 for the data set in which students with misfitting response patterns were removed using l_z, and 0.06 for the data set in which students with misfitting response patterns were removed using H^T (see Table A5, Appendix A). The corresponding values of MAD for the *a* parameter were 0.01, 0.50, 0.32, and 0.37. Overall, the data sets in which students with misfitting response patterns were removed using the H^T provided smaller values of MAD for the *b* parameter and data sets in which students with misfitting response patterns were removed using the l_z provided smaller values of MAD for the *a* parameter.

Classification accuracy. The values of the classification accuracy are reported in Table 14. For the fitting data, 96.0% of the 4,209 students classified as low performing using their true ability parameters were classified as low performing using the item parameter estimates derived from the Fit data set and 4.0% were classified as false positives. Of the 791 students classified as high performing using the true ability parameters, 74.2% were classified as high performing using the item parameters derived from the Fit data set and 25.8% were classified as false negative. While the overall classification percentage for this data set was 92.5%, the accuracy was again greater for the low performing students than the high performing students.

Table 14
Classification accuracy (50% of items and 30% of sample) for test with 20 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4039 (96.0)	170 (4.0)	3924 (93.2)	285 (6.8)	3949 (93.8)	260 (6.2)	3897 (92.6)	312 (7.4)
	High	204 (25.8)	587 (74.2)	365 (46.1)	426 (53.9)	377 (47.7)	413 (52.3)	354 (44.8)	437 (55.2)
% of total agreement		92.5		87.0		87.2		86.7	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

The percentage of correct decisions for the students with true low ability varied slightly from 92.6% to 93.8% for the H^T and l_z data sets, respectively. In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to 53.9%, 52.3% and 55.2% for the misfit, l_z and H^T data sets, respectively. Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (86.7% to 87.2%).

Summary of results for test with 20 items

A summary of findings for simulation study examining estimation bias of the *b* and *a* parameters for a test with 20 items is provided as follows:

b parameter: The bias in the *b* parameter occurred for items in the tails of the distribution and increased as the percentage of misfitting response patterns increased with the exception that of the condition with 50% of items susceptible to misfitting responses and 30% of the students with aberrant response patterns. Fewer biased *b* parameters were found with this exception that with the other "milder" conditions. The degree of estimation bias was more pronounced under the condition where 25% of items were susceptible to misfitting responses than when 50% of items were susceptible to misfitting responses. For the conditions in which 25% of items were susceptible to misfitting responses, the *b* parameters of items with large negative true *b* values were underestimated and the *b* parameters of items with large positive true *b* values were overestimated due to presence of misfitting response patterns. In contrast, for the conditions in which 50% of items were susceptible to misfittingly, the *b* parameters of item with large negative true *b* values were overestimated and the *b* parameters of items with large positive true *b* values were underestimated due to presence of misfitting response patterns.

a parameter. The number of items with bias in the *a* parameter and the size of the bias in *a* increased as the percentage of misfitting response patterns increased. Further, the bias was not found in the items with lowest true *a* values. When bias was found, the true *a* parameters were consistently underestimated across all conditions when 25% of items were susceptible misfitting responses. In contrast, the true *a* parameters were less consistently underestimated for the conditions in which 50% of items were susceptible misfitting responses.

Comparison a and b parameters: The *a* parameter was more affected by presence of misfitting response patterns than the *b* parameter across all data sets for all conditions. While there were items with estimation bias on both the *b* and *a* parameters, this did not occur for all data sets. However, what was common to these items was that they were among the items that had relatively high (in absolute value) true *b* and true *a* values. Using l_z to remove misfitting response patterns resulted in less biased estimates for both the *b* and *a* parameters, particularly when the percentages of misfitting response patterns in the samples were 10% or 20%. In contrast, for the conditions where the percentage of misfitting response patterns in the sample and the percentage of susceptible items was high, using H^T to remove misfitting response patterns resulted in less bias.

Classification accuracy: The patterns of classification of students into two categories were similar across of the four data sets (fitting, misfitting, removal of misfitting patterns using l_z, and removal of misfitting response patterns using H^T). The classification accuracy was higher and essentially equal (> 90%) across the four data sets for the low performing students. For the high performing students, the classification accuracy was lower than 90% for the fitting data set and even lower but essentially equal for the misfitting, l_z and H^T sets. Taken together, despite the presence of estimation bias in the *b* parameter and, to a greater extent, the *a* parameter, using the person fit indices l_z and H^T to remove misfitting response patterns did not improve classification accuracy over not removing misfitting response patterns.

Chapter Five: Results for the Test with 40 Items

This chapter presents the results of the analyses of the simulation for the test with 40 items. The chapter includes only the graph and tables for the summary of the graph and classification accuracy for each of the simulations. The tables for the bias of the item parameter estimates for each simulation condition are included in Appendix A.

40 items, 10% of students with misfitting response patterns in 25% of items

The results for the 40 item test with 10% of the students with misfitting responses to 25% of the items are presented graphically in Figure 11. The full set of results are reported Table A6 in Appendix A.

b parameter: The bias of estimation for the *b* parameter occurred in the tails of the distribution of *b* parameters. Bias occurred in at least for one of the three data sets for four items with the highest true *b* values. Items 2, 4, 13, and 26 had estimation bias beyond ± 0.20 for the data set with all the misfitting response patterns. Items 2 and 13 also had estimation bias beyond ± 0.20 for the data set with misfitting response patterns removed by H^T. The largest bias for these for items occurred for the data set with all the misfitting response patterns and the smallest bias occurred for the data set with misfitting response patterns removed by l_z. Four items, two of which had the lowest difficulty (Items 20 and 36) and two items which had higher but still low *b* values ((Items 7 and 28) had estimation bias beyond ± 0.20, but only for data set with all the misfitting response patterns. Like the 20 item conditions, whereas items with large negative *b* parameters tended to be underestimated (e.g., Item 20), items with large positive *b* parameters tended to be overestimated (e.g., Item 13).

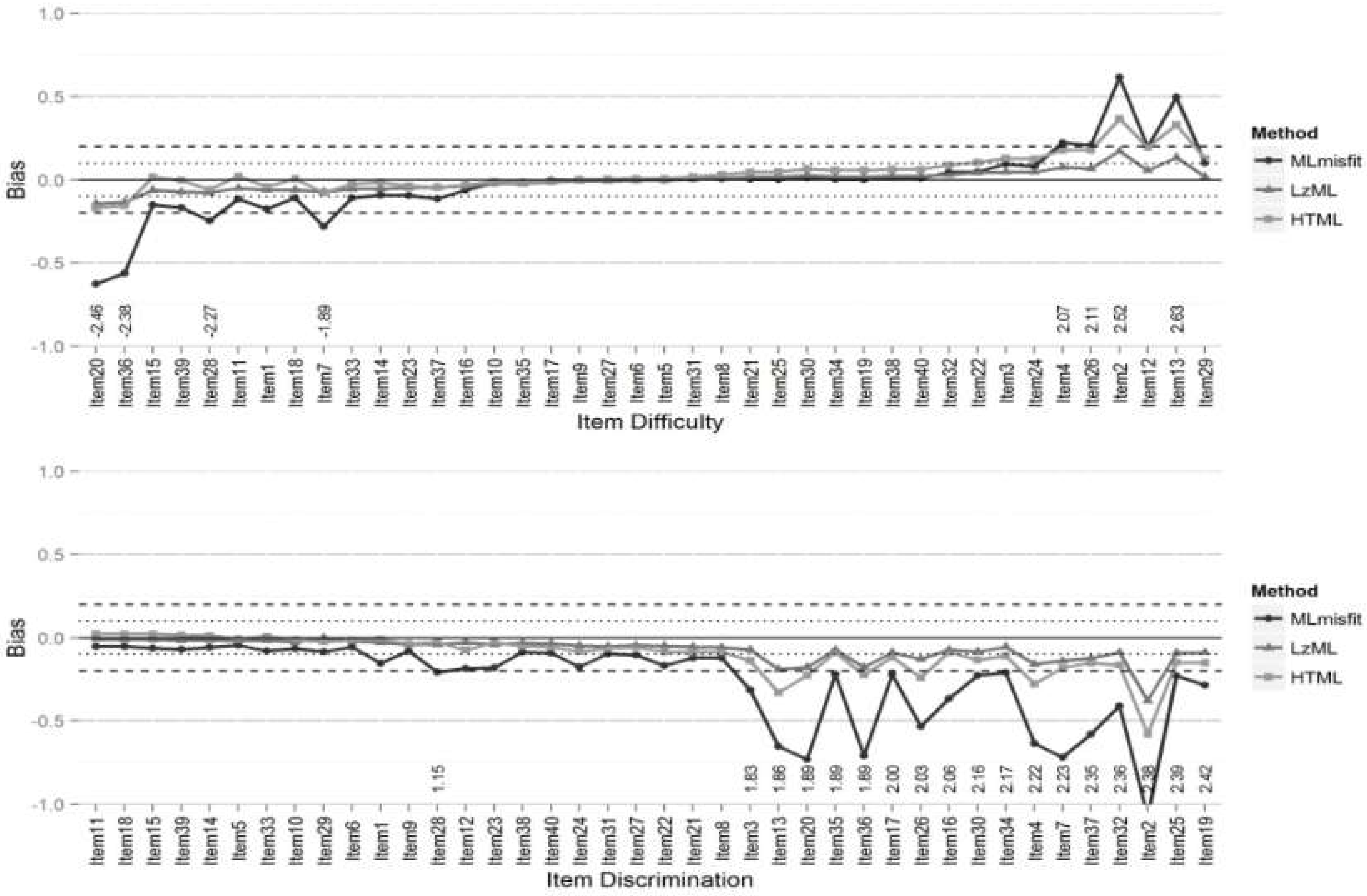

Figure 11: Bias of estimation for manipulating 25% of items and 10% of sample for test with 40 items

a parameter: The *a* parameter was underestimated for the data set with all the misfitting response sets and at least one of the data sets with misfitting response patterns removed by l_z or H^T for the 17 items with the largest true *a* parameters and for Item 28. The bias in *a* for the data set with misfitting response patterns removed by H^T was smaller than -0.20 for Items 2, 4, 13, 20, 26, and 36 and the bias in *a* for the data set with misfitting response patterns removed by l_z was smaller than -0.20 for Item2. For all 18 items, the bias in *a* was larger for the data set with all the misfitting response patterns.

Both b and a parameters: Of the eight items identified with high bias for *b* and the 18 items identified with high bias for *a*, eight items were in common. The items were Item 2 (true *b* of 2.52, true *a* of 2.38), Item 4 (true *b* of 2.07, true *a* of 2.22), Item 7 (true *b* of -1.98, true *a* of 2.23), Item 13 (true *b* of 2.63, true *a* of 1.86), Item 20 (true *b* of -2.46, true *a* of 1.89), Item 26 (true *b* of 2.11, true *a* of 2.03), Item 28 (true *b* of -2.27, true *a* of 1.15),and Item 36 (true *b* of -2.38, true *a* of 1.89). As for the 20 item test conditions, these items were among the items that had relatively high (in absolute value) true *b* and true *a* values.

Comparison a and b parameters. A greater number of items with marked bias in the *a* parameter than items with marked bias in the *b* parameter due to presence of misfitting response patterns were found. Again, the bias was larger for the *a* parameter than for the *b*. The bias for the *b* and *a* parameters estimated using the data set with misfitting response patterns removed by l_z tended to be the smallest.

The results in Table 15 reveal that there were more highly biased estimates of the item discrimination parameter in misfitting data set (18) compared to the difficulty parameter (8). The use of l_z to remove misfitting response patterns resulted in fewer biased item parameters than H^T for both *b* and *a* parameter. For example, whereas the number of items with bias beyond ±0.20 in using l_z is 0 for the b parameter and 1 for the a parameter, the number of items with bias beyond ±0.20 using H^T is 2 for the b parameter and 6 for the a parameter.

Table 15
Bias of estimation across different methods (25% of items and 10% of sample)for test with 40 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	23	36	29
	±0.20 to ±0.10	9	4	9
	Beyond ±0.20	8	0	2
Item discrimination	-0.10 to 0.10	14	32	25
	±0.20 to ±0.10	8	7	9
	Beyond ±0.20	18	1	6

The values of the mean absolute deviation (MAD) for the b parameter were 0.02 for fitting data set, 0.13 for the data set with all the misfitting response patterns, 0.05 for the data set in which students with misfitting response patterns were removed using l_z, and 0.08 for the data set in which students with misfitting response patterns were removed using H^T (see Table A6, Appendix A). The corresponding values of MAD for the a parameter were 0.01, 0.26, 0.07, and 0.11. Overall, the data sets in which students with misfitting response patterns were removed using the l_z provided smaller values of MAD for the b and a parameters.

Classification accuracy. The values of the classification accuracy are reported in Table 16. For the fitting data, of the 4,202 students classified as low performing using their true ability parameters, 96.7% were classified as low performing using the item parameter estimates derived from the Fit data set and 3.3% were classified as false positives. Of the 799 students classified as high performing using the true ability parameters, 74.3% were classified as high performing using the item parameters derived from the Fit data set and 25.7% were classified as false negative.

Table 16
Classification accuracy (25% of items and 10% of sample) for test with 40 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4062 (96.7)	140 (3.3)	4050 (96.4)	152 (3.6)	4060 (96.6)	142 (3.4)	4048 (96.3)	154 (3.7)
	High	205 (25.7)	594 (74.3)	247 (30.9)	552 (69.1)	256 (32)	543 (68.0)	245 (30.7)	553 (69.3)
% of total agreement		93.1		92.0		92.1		92.0	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

While the overall classification percentage for this data set was 93.1%, clearly the accuracy was greater for the low performing students than the high performing students.

The percentages of correct decisions for the students with true low ability varied slightly from 96.3% to 96.7% across the four data sets. In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to 69.1%, 68.0%, and 69.3% for the misfit, l_z and H^T data sets, respectively. Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (92.0% or 92.1%).

40 items, 20% of students with misfitting response patterns in 25% of items

The results for the 40 item test with 20% of the students with misfitting responses to 25% of the items are presented graphically in Figure 12. The full set of results are reported Table A7 in Appendix A.

b parameter: The bias of estimation for the *b* parameter occurred in the tails of the distribution of *b* parameters. Items 25 and 40, which had the highest positive true *b* parameter, had estimation bias beyond ± 0.20 across the data set with all the misfitting response patterns and the data set with all the misfitting response patterns removed by H^T. Item 25 also had estimation bias for the data sets with the misfitting response patterns removed by *lz*. Items 9 and 14, which have the fifth and seventh highest true *b* values, also had estimation bias beyond ± 0.20 for the data set with all the misfitting response patterns removed by H^T. The largest bias for these items occurred for the data set with all the misfitting response patterns and the smallest bias occurred for the data set with misfitting response patterns removed by l_z. With the exception of Item 11, the 12 items with the lowest negative true *b* parameters had estimation bias beyond ± 0.20 for the data set with all the misfitting response patterns. Of these 11 items, Items 2, 8, and 19 also had estimation bias beyond ± 0.20 for the data set for the two data sets with misfitting response patterns removed. Across these 11 items, the *b* parameters estimated from the data set with all the misfitting response patterns had larger estimation bias. Whereas items with large negative *b* parameters tended to be underestimated (e.g., Item 2), items with large positive *b* parameters tended to be overestimated (e.g., Item 25).

a parameter: The *a* parameter was underestimated, with seven exceptions, across all three data sets for the 25 items with the largest true *a* parameters. The seven exceptions included Items 4, 7, 9, 28, 29, 34 and 35 that showed high estimation bias only for the data set with all the misfitting response patterns.. For the remaining 18 items, the bias in *a* was largest for the data set with all the misfitting response patterns and essentially equal for the data sets with misfitting response patterns removed using l_z and using H^T.

Both b and a parameters: Of the 16 items identified with high bias for *b* and the 25 items identified with high bias for *a*, 11 items were in common. The items were Item 2 (true *b* of -2.68, true *a* of 2.40), Item 4 (true *b* of -2.65, true *a* of 0.85), Item 7 (true *b* of -2.09, true *a* of 1.29), Item 8 (true *b* of -2.61, true *a* of 1.51), Item 10 (true *b* of -2.06, true *a* of 1.98), Item 17 (true *b* of

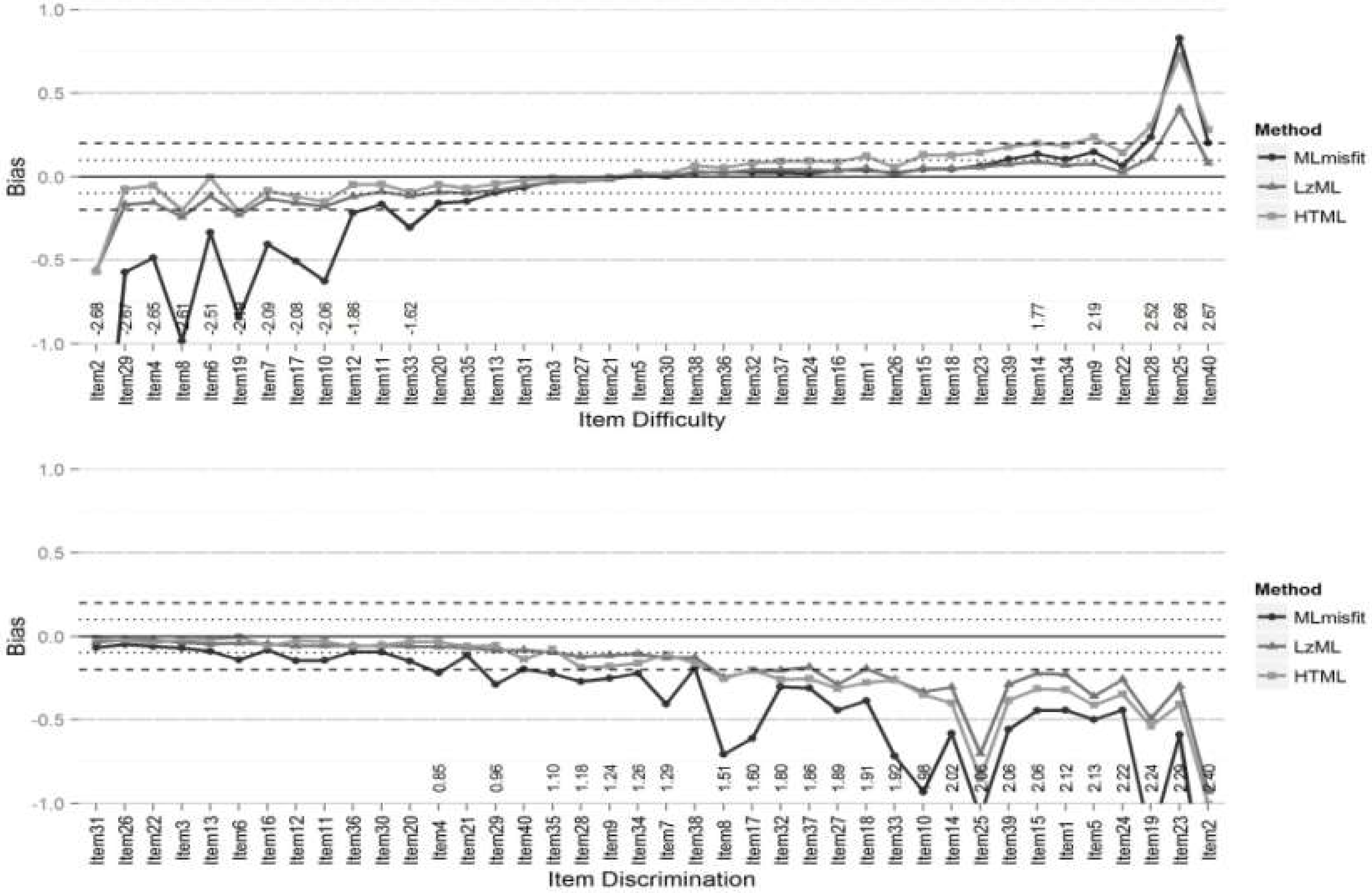

Figure 12: Bias of estimation for manipulating 25% of items and 20% of sample for test with 40 items

-2.08, true a of 1.60), Item 19 (true b of -2.19, true a of 2.24), Item 25 (true b of 2.66, true a of 2.06) , Item 28 (true b of 2.52, true a of 1.18) , Item 29 (true b of -2.67, true a of 0.96), and Item 33 (true b of -1.62, true a of 1.92). What is common to these items is that they were among the items that had relatively high (in absolute value) true b and true a values.

Comparison a and b parameters. In comparison to the b parameter, a greater number of items with marked bias in the a parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the a parameter than for the b parameter. The bias for the b and a parameters estimated using the data set with misfitting response patterns removed by l_z tended to be the smallest.

The results in Table 17 reveal that there were more highly biased estimates of the item discrimination parameter in misfitting data set (25) compared to the difficulty parameter (14). The use of l_z to remove misfitting response patterns resulted in fewer biased item parameters than H^T for both the b and a parameters. For example, whereas the number of items with bias beyond ±0.20 in using l_z is 4 for the b parameter and 16 for the a parameter, the number of items with bias beyond ±0.20 using H^T is 8 for the b parameter and 18 for the a parameter.

Table 17
Bias of estimation across different methods (25% of items and 20% of sample)for test with 40 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	19	27	23
	±0.20 to ±0.10	7	9	9
	Beyond ±0.20	14	4	8
Item discrimination	-0.10 to 0.10	8	17	16
	±0.20 to ±0.10	7	7	6
	Beyond ±0.20	25	16	18

The values of the mean absolute deviation (MAD) for the b parameter were 0.01 for fitting data set, 0.26 for the data set with all the misfitting response patterns, 0.10 for the data set in which students with misfitting response patterns were removed using l_z, and 0.13 for the data set in which students with misfitting response patterns were removed using H^T (see Table A7, Appendix A). The corresponding values of MAD for the a parameter were 0.01, 0.39, 0.19, and 0.22, respectively. Overall, the data sets in which students with misfitting response patterns were removed using the l_z provided smallest values of MAD for the b and a parameters.

Classification accuracy. The values of the classification accuracy are reported in Table 18. For the fitting data 96.9% of the 4,206 students classified as low performing using their true ability parameterswere classified as low performing using the item parameter estimates derived from the Fit data set and 3.1% were classified as false positives. Of the 794 students classified as high performing using the true ability parameters, 81.4% were classified as high performing using the item parameters derived from the Fit data set and 18.6% were classified as false negative. While the overall classification percentage for this data set was 94.4%, clearly the accuracy was greater for the low performing students than the high performing students.

Table 18
Classification accuracy (25% of items and 20% of sample) for test with 40 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4077 (96.9)	129 (3.1)	4071 (96.8)	135 (3.2)	4077 (96.9)	129 (3.1)	4083 (97.1)	123 (2.9)
	High	148 (18.6)	646 (81.4)	247 (31.1)	547 (68.9)	248 (31.3)	545 (68.7)	252 (31.7)	542 (68.3)
% of total agreement		94.4		92.4		92.5		92.5	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

The percentages of correct decisions for the students with true low ability varied slightly from 96.8% to 97.1% across the four data sets. In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to 68.9%, 68.7%, and 68.3% for the misfit, l_z and H^T data sets, respectively. Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (92.4% or 92.5%).

40 items, 30% of students with misfitting response patterns in 25% of items

The results for the 40 item test with 30% of the students with misfitting responses to 25% of the items are presented graphically in Figure 13. The full set of results are reported Table A8 in Appendix A.

b parameter: Items 3, 17, 18, and 19, which had the highest positive value of the true *b* parameter, and Item 10, which was the seventh most difficult item, had estimation bias beyond ± 0.20 across all three misfitting data sets. Items 27 and 32, the fifth and sixth most difficult items, had estimation bias beyond ± 0.20 for data set with all the misfitting response patterns. The largest bias for these seven items occurred for the data set with all the misfitting response patterns and the smallest bias occurred for the data sets with misfitting response patterns removed by H^T or l_z. Seven items, all which had the lowest negative true *b* parameters, had estimation bias beyond ± 0.20 for data set with all the misfitting response patterns. Items 15 (most difficult), 11 (second most difficult), and 24 (sixth most difficult) had estimation bias beyond ± 0.20 for data sets with misfitting response sets removed by l_z and by H^T. Again, whereas items with large negative true *b* parameters tended to be underestimated (e.g., Item 15), items with large positive true *b* parameters tended to be overestimated (e.g., Item 19).

a parameter: The *a* parameter was underestimated across all three data sets for the 22 items with the largest true *a* parameters. As well, Items 11, 17, 32, 36, and 37 had high estimation bias only for the data set with all the misfitting response patterns. For all items with bias, the bias in *a* was largest for the data set with all the misfitting response patterns.

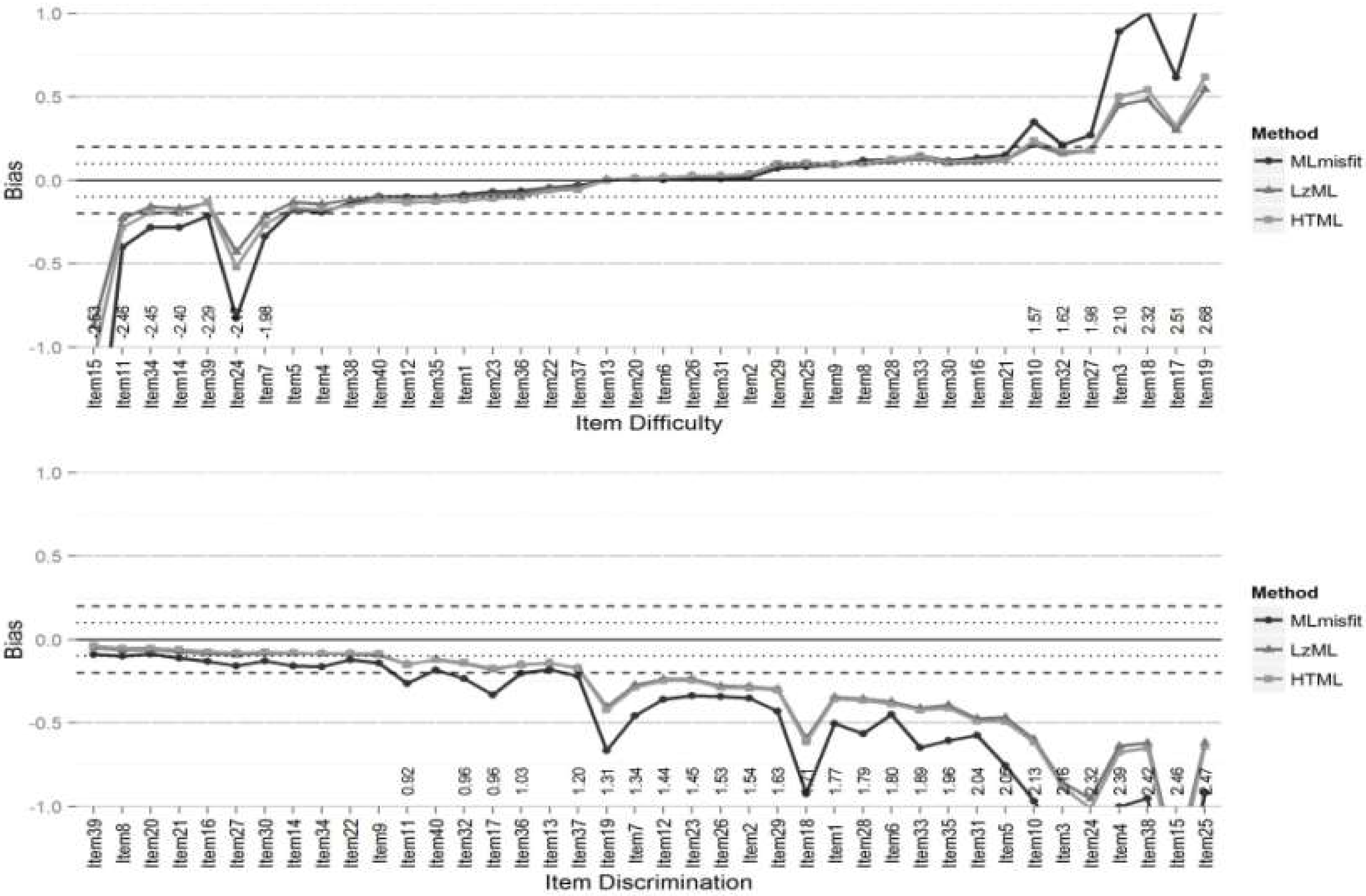

Figure 13: Bias of estimation for manipulating 25% of items and 30% of sample for test with 40 items

As seen in Figure 13, using l_z or H^T to remove misfitting response patterns resulted in approximately similar number of biased estimated item parameters for both the *b* and *a* parameters.

Both b and a parameters: Ten items had high bias for *b* and for *a*. The items inlcuded Item 3 (true *b* of 2.10, true *a* of 2.16), Item 7 (true *b* of -1.98, true *a* of 1.34), Item 10 (true *b* of 1.57, true *a* of 2.13), Item 11 (true *b* of -2.46, true *a* of 0.92), Item 15 (true *b* of -2.53, true *a* of 2.46), Item 17 (true *b* of 2.51, true *a* of 0.96), Item 18 (true *b* of 2.32, true *a* of 1.71), Item 19 (true *b* of 2.68, true *a* of 1.31) , Item 24 (true *b* of -2.17, true *a* of 2.32) and Item 32 (true *b* of 1.62, true *a* of 0.96). What is common to these items is that they were among the items that had relatively high (in absolute value) true *b* and true *a* values.

Comparison a and b parameters. In comparison to the *b* parameter, a greater number of items with marked bias in the *a* parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the *a* parameter than for the *b*. As seen in Figure 13, using l_z or H^T to remove misfitting response patterns resulted in approximately similar number of biased estimated item parameters for both the *b* and *a* parameters.

The results in Table 19 reveal that there were more highly biased estimates of item discrimination parameter in the misfitting data set (27) compared to difficulty parameter (14). Both indices performed the same for the *a* parameter. For example, whereas the number of items with bias beyond ±0.20 in using l_z is 9 for the *b* parameter and 22 for the *a* parameter, the number of items with bias beyond ±0.20 using H^T is 9 for the *b* parameter and 22 for the *a* parameter.

Table 19
Bias of estimation across different methods (25% of items and 30% of sample)for test with 40 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	17	16	9
	±0.20 to ±0.10	9	15	22
	Beyond ±0.20	14	9	9
Item discrimination	-0.10 to 0.10	3	11	11
	±0.20 to ±0.10	10	7	7
	Beyond ±0.20	27	22	22

The values of the mean absolute deviation (MAD) for the *b* parameter were 0.02 for fitting data set, 0.27 for the data set with all the misfitting response patterns, 0.17 for the data set in which students with misfitting response patterns were removed using l_z, and 0.19 for the data set in which students with misfitting response patterns were removed using H^T (see Table A8, Appendix A). The corresponding values of MAD for the *a* parameter were 0.01, 0.48, 0.32, and 0.33. Overall, the data sets in which students with misfitting response patterns were removed using the l_z provided smaller values of MAD for the *b* and *a* parameters.

Classification accuracy. The values of the classification accuracy are reported in Table 20. For the fitting data, 97.1% of the 4,204 students classified as low performing using their true ability parameterswere classified as low performing using the item parameter estimates derived from the Fit data set and 2.9% were classified as false positives. Of the 796 students classified as high performing using the true ability parameters, 80.5% were classified as high performing using the item parameters derived from the Fit data set and 19.5% were classified as false negative. While the overall classification percentage for this data set was 94.5%, clearly the accuracy was greater for the low performing students than the high performing students.

Table 20
Classification accuracy (25% of items and 30% of sample) for test with 40 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4082 (97.1)	122 (2.9)	4077 (97.0)	128 (3.0)	4081 (97.1)	124 (2.9)	4070 (96.8)	134 (3.2)
	High	155 (19.5)	641 (80.5)	298 (37.5)	497 (62.5)	303 (38.1)	493 (61.9)	296 (37.2)	499 (62.8)
% of total agreement		94.5		91.5		91.5		91.4	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

The percentage of correct decisions for the students with true low ability varies slightly from 96.8% to 97.1% across the four data sets. In contrast the percentage of correct decisions for the students with true high ability drops from the percentage for the fitted data to 62.5%, 61.9%, and 62.8% for the misfit, l_z and H^T data sets, respectively. Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (91.4% or 91.5%).

40 items, 10% of students with misfitting response patterns in 50% of items

The results for the 40 item test with 10% of the students with misfitting responses to 50% of the items are presented graphically in Figure 14. The full set of results are reported Table A9 in Appendix A.

b parameter: The bias of estimation for the *b* parameter occurred in the tails of the distribution of *b* parameters. Only one item, Item 25, which had the highest positive true *b* parameter, had estimation bias beyond ± 0.20 and only for data set with all the misfitting response patterns. The nine items with the lowest negative true *b* parameters had estimation bias beyond ± 0.20 for data set with all the misfitting response patterns. Of these seven items, Item 26, which was the most difficult item, had estimation bias larger than 0.20 for the data sets with misfitting response patterns removed by l_z. Whereas *b* parameter for the items with large

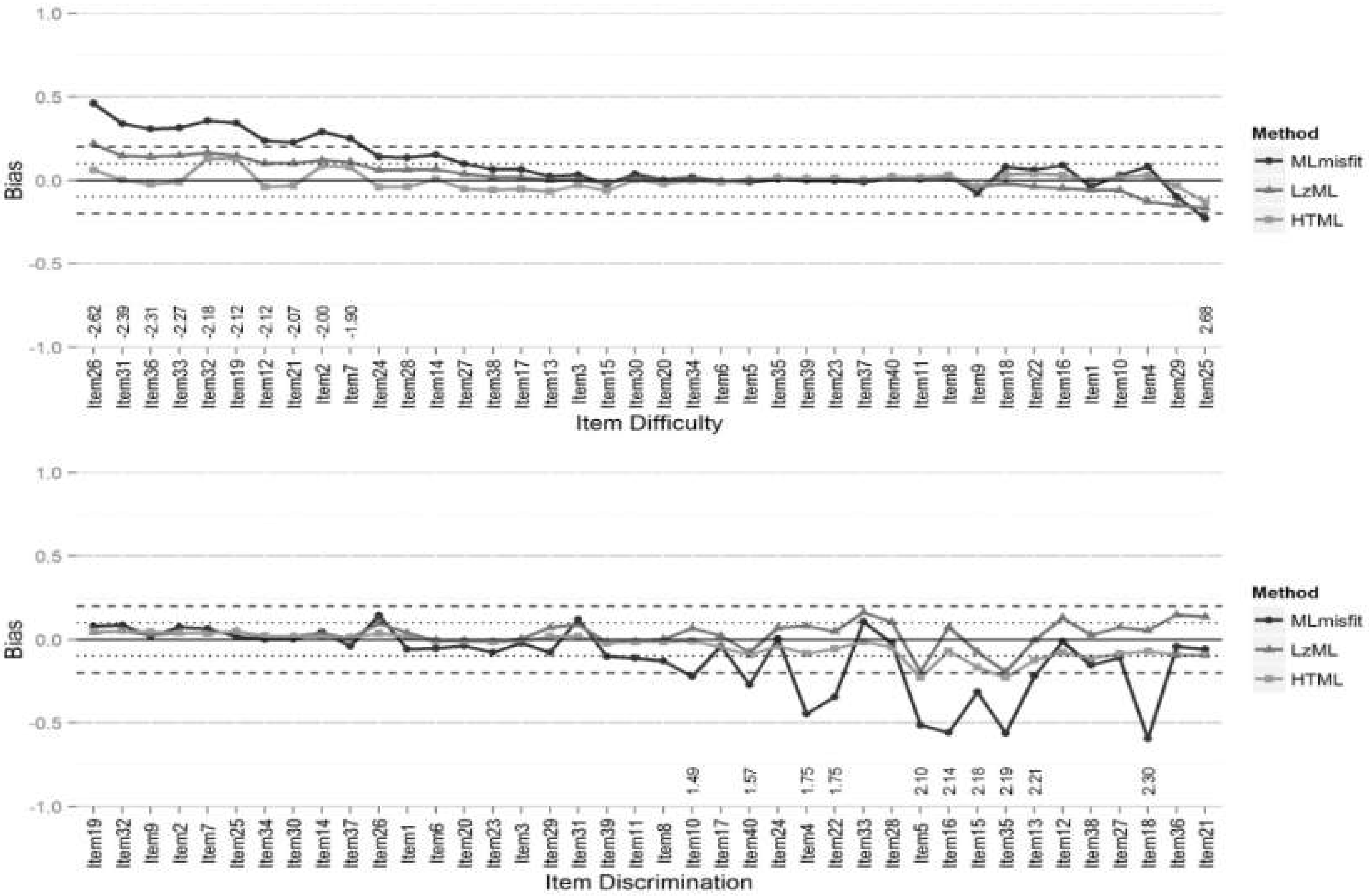

Figure 14: Bias of estimation for manipulating 50% of items and 10% of sample for test with 40 items

negative *b* parameters tended to be overestimated (e.g., Item 26), the b parameter for Item 25 tended to be underestimated.

a parameter: The *a* parameter was underestimated for the 10 items. In contrast to previous conditions, only one (Item 18) of the seven most discriminating items had bias less than -0.20 and only for the data set with all the misfitting response patterns. . Five consecutive items had bias less than -0.20 for the data set with all the misfitting response patterns. Of these five items, Items 5 and 35 also had high estimation bias for the data set with misfitting response patterns removed by H^T. The bias for the remaining four items (Items 10, 40, 4, and 22) had bias less than -0.20 for the data set with all the misfitting response patterns. Overall, the bias in *a* was largest for the data set with all the misfitting response patterns.

Both b and a parameters: There was no common items with estimation bias for both the *b* and *a* parameters.

Comparison a and b parameters. A greater number of items with marked bias in the *a* parameter than marked bias in the b parameter were found. Further, the bias was larger for the *a* parameter than for the *b* parameter.

The results in Table 21 reveal that there were similar numbers of highly biased estimates of item discrimination parameter in the misfitting data set (10) compared to difficulty parameter (11). The use of H^T to remove misfitting response patterns resulted in fewer, but essentially the same, biased item parameters than l_z for both the *b* and *a* parameters. For example, whereas the number of items with bias beyond ±0.20 in using l_z is 1 for the *b* parameter and 0 for the *a* parameter, the number of items with bias beyond ±0.20 using H^T is 0 for the *b* parameter and 2 for the *a* parameter.

Table 21
Bias of estimation across different methods (50% of items and 10% of sample)for test with 40 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	26	27	37
	±0.20 to ±0.10	3	12	3
	Beyond ±0.20	11	1	0
Item discrimination	-0.10 to 0.10	22	33	35
	±0.20 to ±0.10	8	7	3
	Beyond ±0.20	10	0	2

The values of the mean absolute deviation (MAD) for the b parameter were 0.02 for fitting data set, 0.12 for the data set with all the misfitting response patterns, 0.06 for the data set in which students with misfitting response patterns were removed using l_z, and 0.04 for the data set in which students with misfitting response patterns were removed using H^T (see Table A9, Appendix A). The corresponding values of MAD for the a parameter were 0.01, 0.15, 0.06, and 0.06. Overall, the data sets in which students with misfitting response patterns were removed using the H^T provided smaller values of MAD for the b parameter.

Classification accuracy. The values of the classification accuracy are reported in Table 22. For the fitting data, 96.8% of the 4,207 students classified as low performing using their true ability parameterswere classified as low performing using the item parameter estimates derived from the Fit data set and 3.2% were classified as false positives. Just over three-quarters (77.3%) of the 793 students classified as high performing using the true ability parameters classified as high performing using the item parameters derived from the Fit data set and just under a quarter (22.7%) were classified as false negative. While the overall classification percentage for this data set was 93.7%, the accuracy was greater for the low performing students than the high performing students.

The percentage of correct decisions for the students with true low ability varied slightly from 94.1% to 96.8% across the four data sets, respectively. In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to 68.5%, 69.2%, and 69.6% for the misfit, l_z and H^T data sets, respectively. Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (90.2% or 90.4%).

Table 22
Classification accuracy (50% of items and 10% of sample) for test with 40 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4073	134	3978	229	3962	245	3959	248
		(96.8)	(3.2)	(94.6)	(5.4)	(94.2)	(5.8)	(94.1)	(5.9)
	High	180	613	250	543	244	549	241	552
		(22.7)	(77.3)	(31.5)	(68.5)	(30.8)	(69.2)	(30.4)	(69.6)
% of total agreement		93.7		90.4		90.2		90.2	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

40 items, 20% of students with misfitting response patterns in 50% of items

The results for the 40 item test with 20% of the students with misfitting responses to 50% of the items are presented graphically in Figure 15. The full set of results are reported Table A10 in Appendix A.

b parameter: The bias of estimation for the *b* parameter occurred in the tails of the distribution of *b* parameters. The eight items with the highest *b* values had bias for the data set with all the misfitting response patterns. Further, the four most difficult items had bias for the data sets with misfitting response patterns removed by l_z and by H^T. The seven itemswith the lowest negative true *b* parameters had estimation bias beyond ± 0.20 for data set with all the misfitting response patterns. The five most difficult items had bias beyond ± 0.20 for the data set with misfitting response patterns removed by l_z and the most difficult item also had bias for the data set with misfitting response patterns removed by H^T. One other item with a negative true b

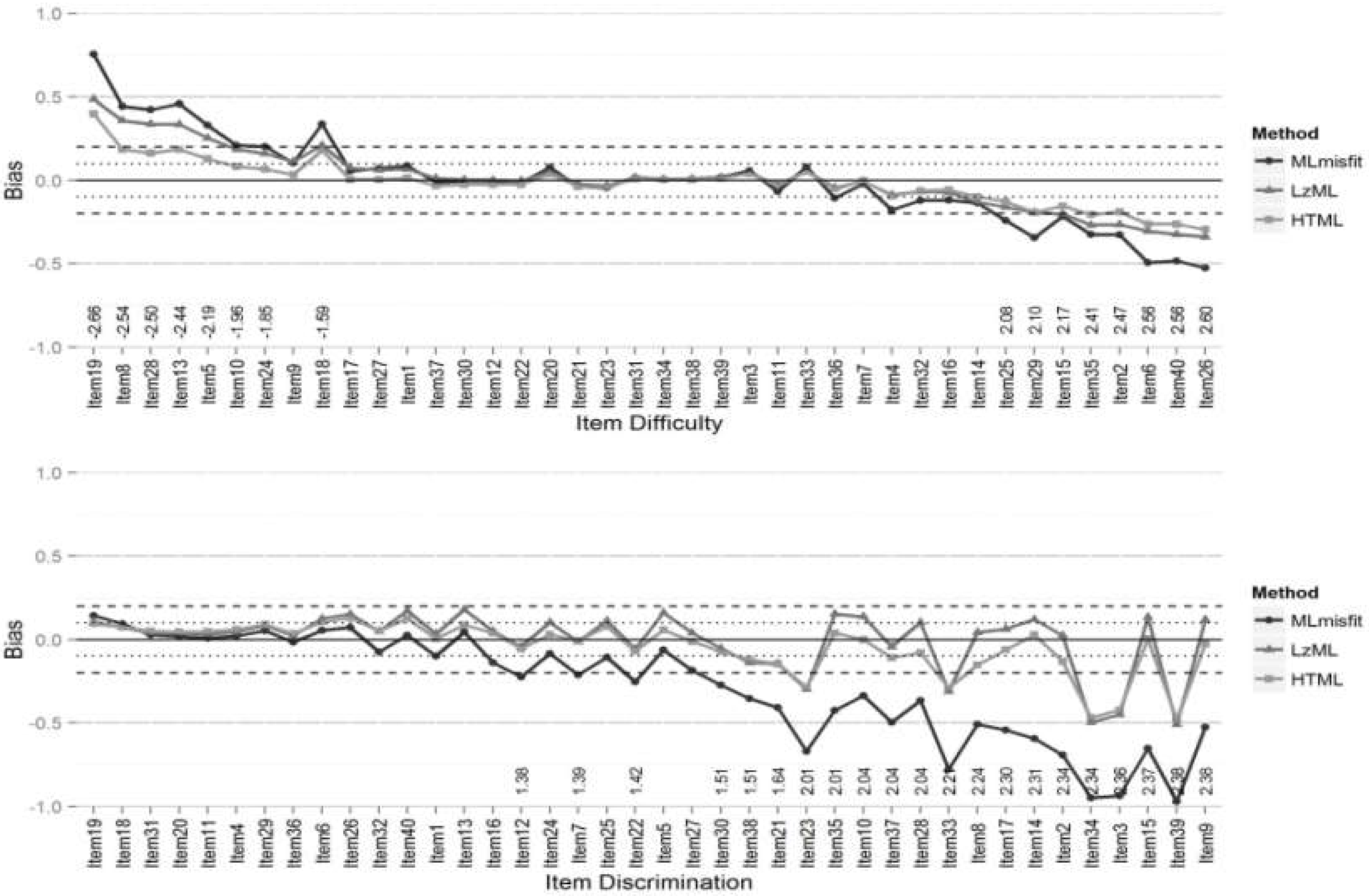

Figure 15: Bias of estimation for manipulating 50% of items and 20% of sample for test with 40 items

parameter, Item 18, had bias for the data set with all the mifitting response patterns. Again, whereas items with large negative *b* parameters tended to be overestimated (e.g., Item 19), items with large positive *b* parameters tended to be underestimated (e.g., Item 26).

a parameter: The *a* parameter was underestimated for the 18 items with the largest a value for at least one the misfitting data sets. For all 18 items, one of the data sets was the data set with all the missing response patterns. Of these 18 items, bias for the two data sets with misfitting response patterns occurred for Items 23, 33, 34, 3, and 39. In addition to the 18 items, bias was found for the data set with all misfittting response patterns for Items 12, 7, and 22.

Both b and a parameters: Of the 16 items identified with high bias for *b* and the 21 items identified with high bias for *a*, six items were in common. The items were Item 2 (true *b* of 2.47, true *a* of 2.34), Item 8 (true *b* of -2.54, true *a* of 2.24), Item 10 (true *b* of -1.96, true *a* of 2.04), Item 15 (true *b* of 2.17, true *a* of 2.37), Item 28 (true *b* of -2.50, true *a* of 2.04) and Item 35 (true *b* of 2.41, true *a* of 2.01). What is common to these items is that they were among the items that had relatively high (in absolute value) true *b* and true *a* values.

Comparison a and b parameters. In comparison to the *b* parameter, a greater number of items with marked bias in the *a* parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the *a* parameter than for the *b* parameter. The bias for the *b* and *a* parameters estimated using the data set with misfitting response patterns removed by H^T tended to be the smallest.

The results in Table 23 reveal that there were more highly biased estimates of item discrimination parameter in the misfitting data set (21) compared to difficulty parameter (16). The use of H^T to remove misfitting response patterns resulted in fewer biased item parameters than l_z for both the *b* and *a* parameters. For example, whereas the number of items with bias beyond ±0.20 in using l_z is 12 for the *b* parameter and 5 for the *a* parameter, the number of items with bias beyond ±0.20 using H^T is 5 for both the *b* and *a* parameters.

Table 23
Bias of estimation across different methods (50% of items and 20% of sample)for test with 40 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	18	22	26
	±0.20 to ±0.10	6	6	9
	Beyond ±0.20	16	12	5
Item discrimination	-0.10 to 0.10	15	19	27
	±0.20 to ±0.10	4	16	8
	Beyond ±0.20	21	5	5

The values of the mean absolute deviation (MAD) for the *b* parameter were 0.01 for fitting data set, 0.19 for the data set with all the misfitting response patterns, 0.13 for the data set in which students with misfitting response patterns were removed using l_z, and 0.10 for the data set in which students with misfitting response patterns were removed using H^T (see Table A10, Appendix A). The corresponding values of MAD for the *a* parameter were 0.01, 0.31, 0.13, and 0.11. Overall, the data sets in which students with misfitting response patterns were removed using the H^T provided smaller values of MAD for the *b* parameter.

Classification accuracy. The values of the classification accuracy are reported in Table 24. For the fitting data, of the 4,209 students classified as low performing using their true ability parameters, 97.1% were classified as low performing using the item parameter estimates derived from the Fit data set and 2.9% were classified as false positives. Of the 791 students classified as high performing using the true ability parameters, 77.1% were classified as high performing using the item parameters derived from the Fit data set and 22.9% were classified as false negative. While the overall classification percentage for this data set was 94.0%, the accuracy was greater for the low performing students than the high performing students.

The percentage of correct decisions for the students with true low ability dropped from the percentage for the fitting data by approximately 10% (97.2% vs. 88.4%, 87.8%, and 86.7%, respectively). In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data by a greater amount to 50.3%, 57.4% and 59.3% for the misfit, l_z and H^T data sets, respectively. Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (82.4% to 83.2%).

Table 24
Classification accuracy (50% of items and 20% of sample) for test with 40 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4088 (97.1)	121 (2.9)	3721 (88.4)	487 (11.6)	3696 (87.8)	512 (12.2)	3691 (87.7)	517 (12.3)
	High	181 (22.9)	610 (77.1)	393 (49.7)	398 (50.3)	337 (42.6)	455 (57.4)	322 (40.7)	469 (59.3)
% of total agreement		94.0		82.4		83.0		83.2	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

40 items, 30% of students with misfitting response patterns in 50% of items

The results for the 40 item test with 30% of the students with misfitting responses to 50% of the items are presented graphically in Figure 16. The full set of results are reported Table A11 in Appendix A.

*b parameter:*The nine items with the highest values of *b* had bias beyond ± 0.20 for the data set with all the misfitting response patterns and the data set with misfitting response patterns removed using l_z. Further the bias was beyond ± 0.20 for the data set with misfitting response patterns removed using H^T for the five most difficult items and the seventh most difficult item (Item 30)The eight items with the lowest negative true *b* parameters had estimation bias beyond

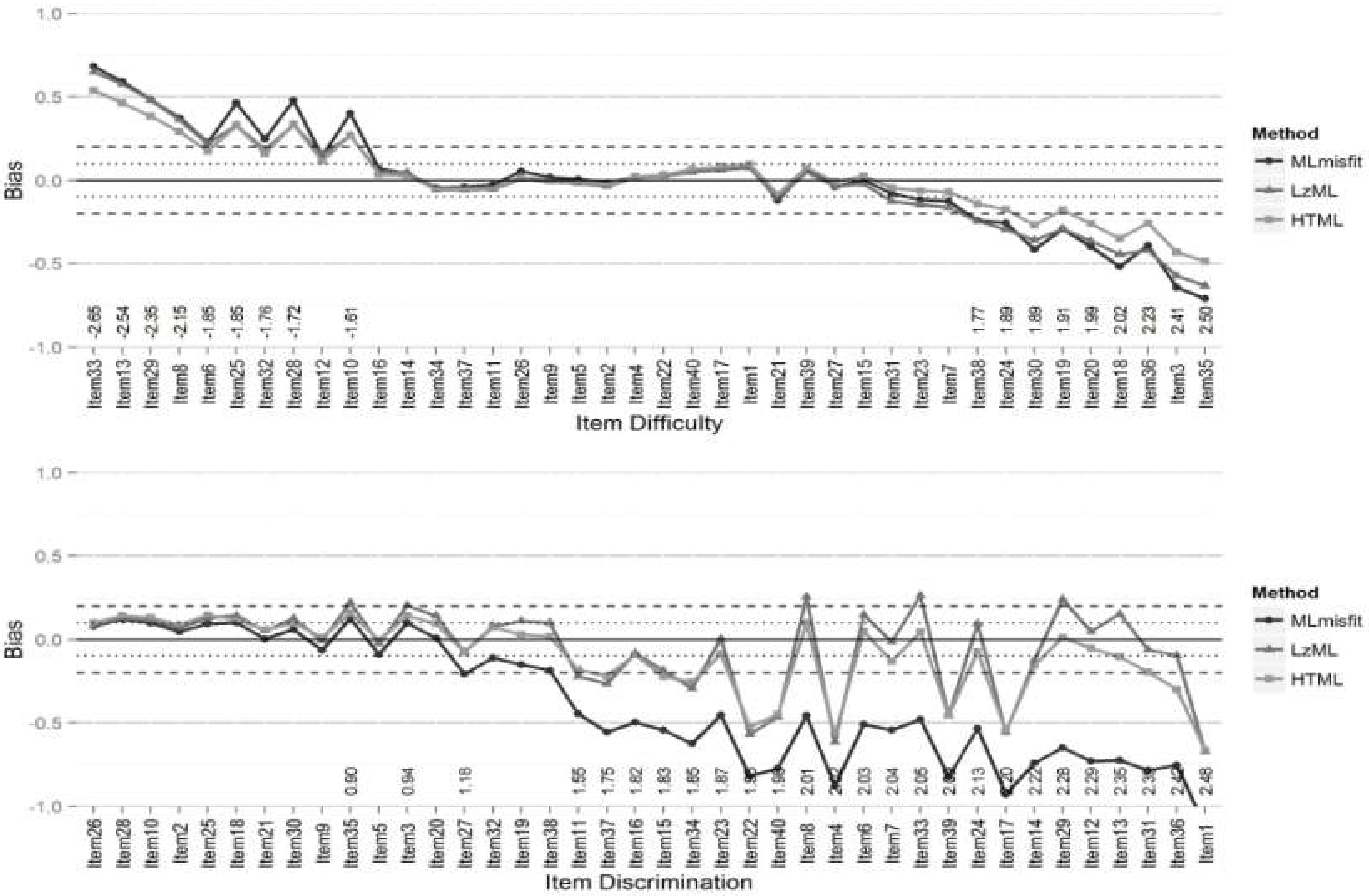

Figure 16: Bias of estimation for manipulating 50% of items and 30% of sample for test with 40 items

± 0.20 for data set with all the misfitting response patterns. Of the eight items, all but two items (Items 6 and 32) had bias beyond ± 0.20 for the data set with msifitting response patterns removed using H^T and all but one item (Item 32) had bias beyond ± 0.20 for the data set with misfitting response patterns removed using l_z. One other item, Item 10, had bias for all three misfitting data sets. Again, whereas items with large negative *b* parameters tended to be overestimated (e.g., Item 33), items with large positive *b* parameters tended to be underestimated (e.g., Item 35).

a parameter: The *a* parameter was underestimated for the 23 items for the data set with all misfitting response patterns. Further, of these items, seven had bias in *a* for both data sets for which misfitting response patterns were removed. In contrast, the bias for three items (Items 8, 33, and 28) had positive bias beyond 0.20 for the data set with misfitting response patterns removed using l_z. two additional items (Items 35 and 3) had positive bias for the data set with the misfitting response patterns removed using lz nd a third item (Item 27) had negative bias for the data set with all the misfitting response patterns.

Both b and a parameters: Of the 18 items identified with high bias for *b* and the 26 items identified with high bias for *a*, nine items were in common. The items were Item 3 (true *b* of 2.41, true *a* of 0.94), Item 6 (true *b* of -1.85, true *a* of 2.03), Item 8 (true *b* of -2.15, true *a* of 2.01), Item 13 (true *b* of -2.54, true *a* of 2.35), Item 24 (true *b* of 1.89, true *a* of 2.13) , Item 29 (true *b* of -2.35, true *a* of 2.28) , Item 33 (true *b* of -2.65, true *a* of 2.05) , Item 35 (true *b* of 2.50, true *a* of 0.90) and Item 36 (true *b* of 2.23, true *a* of 2.42). What is common to these items is that they were among the items that had relatively high (in absolute value) true *b* and true *a* values.

Comparison a and b parameters. A greater number of items with bias in the *a* parameter than items with bias in the *b* parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the *a* parameter than for the *b* parameter and for was positive for the data set with missing response patterns removed by l_z.. The bias for the *b* and *a* parameters estimated using the data set with misfitting response patterns removed by H^T tended to be the smallest.

The results in Table 25 reveal that there were more highly biased estimates of the item discrimination parameter in the misfitting data set (24) compared to the difficulty parameter (18).

Table 25
Bias of estimation across different methods (50% of items and 30% of sample)for test with 40 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	18	17	21
	±0.20 to ±0.10	4	6	6
	Beyond ±0.20	18	17	13
Item discrimination	-0.10 to 0.10	9	15	18
	±0.20 to ±0.10	7	11	12
	Beyond ±0.20	24	14	10

The use of H^T to remove misfitting response patterns resulted in fewer biased item parameters than l_z for both the *b* and *a* parameters. For example, whereas the number of items with bias beyond ±0.20 in using l_z was 17 for the *b* parameter and 14 for the *a* parameter, the number of items with bias beyond ±0.20 using H^T was 13 for the *b* parameter and 10 for the *a* parameter.

The values of the mean absolute deviation (MAD) for the *b* parameter were 0.01 for fitting data set, 0.23 for the data set with all the misfitting response patterns, 0.21 for the data set in which students with misfitting response patterns were removed using l_z, and 0.16 for the data set in which students with misfitting response patterns were removed using H^T (see Table A11, Appendix A). The corresponding values of MAD for the *a* parameter were 0.01, 0.43, 0.19, and 0.18. Overall, the data sets in which students with misfitting response patterns were removed using the H^T provided smaller values of MAD for the *b* parameter.

Classification accuracy. The values of the classification accuracy are reported in Table 26. For the fitting data with ML estimate of item parameters, of the 4,208 students classified as low performing using their true ability parameters, 97.1% were classified as low performing using the item parameter estimates derived from the Fit data set and 2.9% were classified as false positives. Of the 793 students classified as high performing using the true ability parameters, 84.5% were classified as high performing using the item parameters derived from the Fit data set and 15.5% were classified as false negative. While the overall classification percentage for this data set was 95.1%, clearly the accuracy was greater for the low performing students than the high performing students.

Table 26
Classification accuracy (50% of items and 30% of sample) for test with 40 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4086 (97.1)	122 (2.9)	3783 (89.9)	424 (10.1)	3780 (89.8)	428 (10.2)	3745 (89.0)	463 (11)
	High	123 (15.5)	670 (84.5)	357 (45.1)	435 (54.9)	353 (44.6)	439 (55.4)	334 (42.2)	458 (57.8)
% of total agreement		95.1		84.4		84.4		84.1	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

The percentage of correct decisions for the students dropped from the percentage for the fitting data set by approximately 9% for the three misfitting data sets (97.1% vs. 89.9%, 89.8%, and 89.0%). In contrast the percentage of correct decisions for the students with true high ability dropped by a greater from the percentage for the fitted data to 45.1%, 44.6% and 42.2% for the misfit, l_z and H^T data sets, respectively. Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (84.1% or 84.4%).

Summary of results for test with 40 items

A summary of findings for simulation study examining estimation bias of *b* and *a* parameters in a test with 40 items is provided as follows:

b parameter: The results showed that an increase in the percentage of misfitting response patterns in a sample led to larger biased estimates of the *b* parameter. Estimation bias occurred for large negative or positive *b* values whereas values in the middle of distribution were less affected by presence of misfitting response patterns. In the situation where 25% of items were susceptible to misfitting responses, the large negative *b* values were underestimated and the large positive *b* values were overestimated. In contrast, in the situations where 50% of items were susceptible to be answered misfittingly, the large negative *b* values were overestimated and large positive *b* values were underestimated due to presence of misfitting response patterns. The degree of estimation bias was more pronounced under the condition where 25% of items were susceptible to misfitting responses than under the condition where 50% of items were susceptible to misfitting responses.

a parameter: The number of items with bias in the *a* parameter and the size and direction of the bias in *a* increased as the percentage of misfitting response patterns increased. Further, the bias was not found in the items with lowest true *a* values. When bias was found, the true *a* parameters were consistently underestimated across all conditions when 25% of items were susceptible misfitting responses. In contrast, while the true *a* parameters were underestimated for the majority of items with bias for the conditions in which 50% of items were susceptible misfitting responses, there were a small number of items for which the bias was postive.

Comparison a and b parameters: The *a* parameter was more affected by presence of misfitting response patterns than the *b* parameter across all data sets for all conditions. There were items with estimation bias for both the *b* and *a* parameters. What was common to these items was that they were among the items that had relatively high (in absolute value) true *b* and true *a* values. Using l_z to remove misfitting response patterns resulted in less biased estimates for both the *b* and *a* parameters, particularly when the percentage of misfitting response patterns in sample were 10% or 20%. In conditions where the percentage of misfitting response patterns in the sample and the percentage of susceptible items was high, using H^T resulted in less bias.

Classification accuracy: The patterns of classification of the students into two categories were similar for each of the four data sets. The classification accuracy was higher and essentially equal (> 87%) across the four data sets for the low performing students. For the high performing students, the classification accuracy was lower than 90% for the fitting data set and even lower but essentially equal for the misfitting, l_z and H^T data sets. Taken together, despite the presence of bias in the *b* parameter and, to a greater extent, the *a* parameter, using the person fit indices l_z and H^T to remove misfitting response patterns did not improve classification accuracy over not removing misfitting response patterns.

Chapter Six: Results for the test with 60 items

This chapter presents the results of the analyses of the simulation for the test with 60 items. As with the previous two chapters, Chapter includes only the graph and tables for the summary of the graph and classification accuracy for each of the simulations. The tables for the bias of the item parameter estimates for each simulation condition are included in Appendix A.

60 items, 10% of students with misfitting response patterns in 25% of items

The results for the 60 item test with 10% of the students with misfitting responses to 25% of the items are presented graphically in Figure 17. The full set of results are reported Table A12 in Appendix A.

b parameter: Items 40 and 41, which had the highest positive true *b* parameter , had estimation bias beyond ± 0.20 for data set with all the misfitting response patterns. Item 40 also had estimation bias close to 0.20 for data set with all the misfitting response patterns removed by H^T. Four items with the lowest ngative true *b* parameters and four items with the sixth to ninth lowest negative true *b* parameters, had estimation bias beyond ± 0.20 for data set with all the misfitting response patterns. Item 22, which had he lowest true b parameter also had estimation bias below -0.20 for the data set with misfitting response patterns removed by H^T. In contrast to the 20 and 40 item tests, whereas items with large negative *b* parameters tended to be underestimated (e.g., Item 22), the two items with large positive *b* parameters tended to be overestimated (e.g., Item 40).

a parameter: The *a* parameter was underestimated for the six items with the highest true *a* valuesfor the data set with all misfitting response patterns. As well, Item 48 had bias for the data set with misfitting response patterns removed by H^T. As shown in Figure 17, 17 additional items with true *a* parameters that were more moderate in value had bias beyond ± 0.20 for the

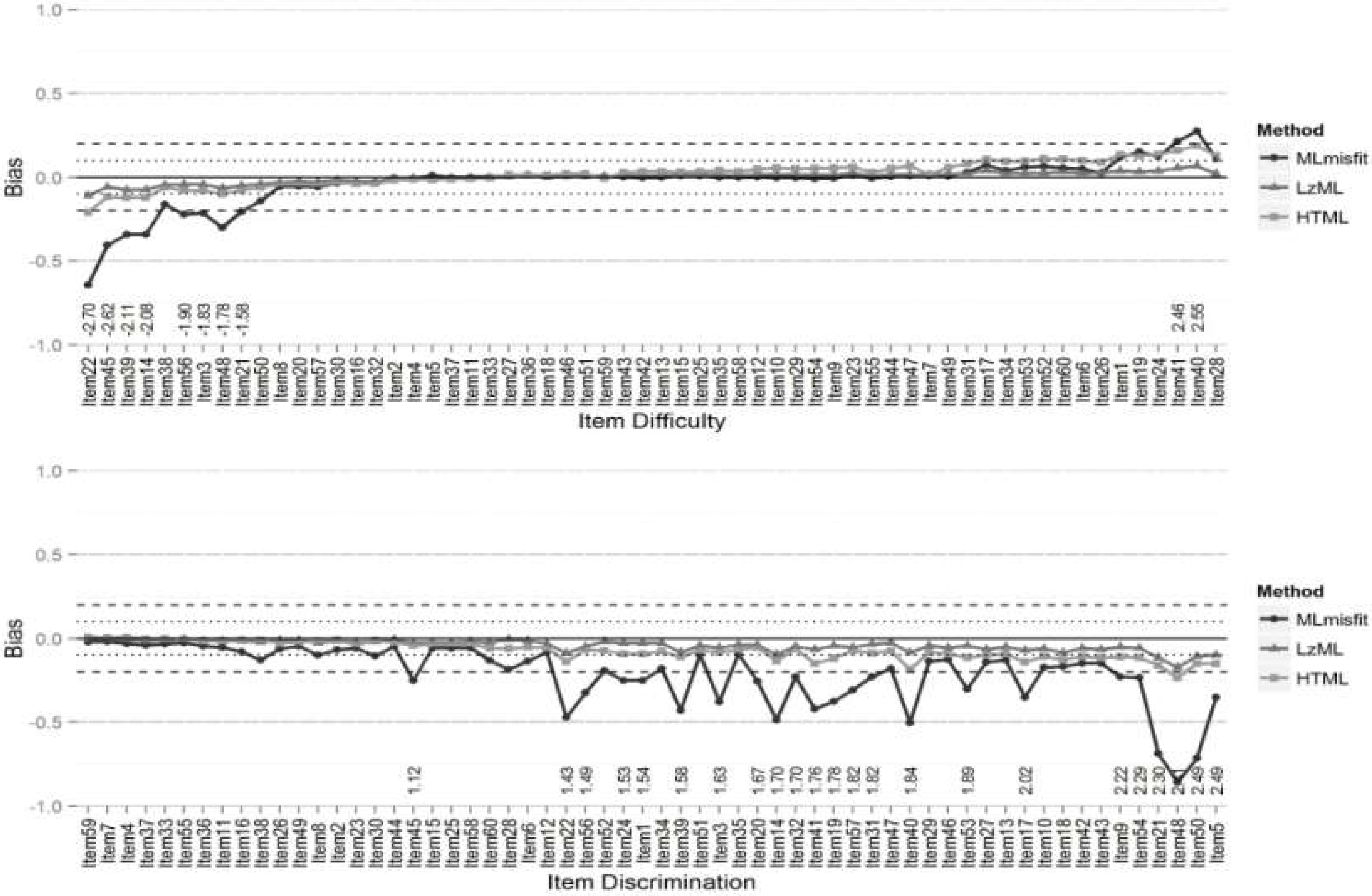

Figure 17: Bias of estimation for manipulating 25% of items and 10% of sample for test with 60 items

data set with all the misfitting response patterns.

Both b and a parameters: Of the 10 items identified with high bias for *b* and the 23 items identified with high bias for *a*, nine items were in common. The items were Item 3 (true *b* of -1.83, true *a* of 1.63), Item 14 (true *b* of -2.08, true *a* of 1.70), Item 22 (true *b* of -2.70, true *a* of 1.43), Item 39 (true *b* of -2.11, true *a* of 1.58), Item 40 (true *b* of 2.55, true *a* of 1.84), Item 41 (true *b* of 2.46, true *a* of 1.76), Item 45 (true *b* of -2.62, true *a* of 1.12) , Item 48 (true *b* of -1.78, true *a* of 2.41) and Item 56 (true *b* of -1.90, true *a* of 1.49). What is common to these items is that they were among the items that had relatively high (in absolute value) true *b* and true *a* values.

Comparison a and b parameters. In comparison to the *b* parameter, a greater number of items with marked bias in the *a* parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the *a* parameter than for the *b*.

The results in Table 27 reveal that there were more highly biased estimates of the item discrimination parameter in the misfitting data set (23) compared to the difficulty parameter (10). The use of l_z to remove misfitting response patterns resulted in fewer, but essentially the same, biased item parameters than the use of H^T for both the *b* and *a* parameters. For example, whereas the number of items with bias beyond ±0.20 in using l_z is 0 for both the *b* and *a* parameters, the number of items with bias beyond ±0.20 using H^T is 1 for both the *b* and *a* parameters.

Table 27
Bias of estimation across different methods (25% of items and 10% of sample)for test with 60 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	44	59	46
	±0.20 to ±0.10	6	1	13
	Beyond ±0.20	10	0	1
Item discrimination	-0.10 to 0.10	21	57	42
	±0.20 to ±0.10	16	3	17
	Beyond ±0.20	23	0	1

The values of the mean absolute deviation (MAD) for the *b* parameter were 0.01 for fitting data set, 0.08 for the data set with all the misfitting response patterns, 0.03 for the data set in which students with misfitting response patterns were removed using l_z, and 0.06 for the data set in which students with misfitting response patterns were removed using H^T (see Table A12, Appendix A). The corresponding values of MAD for the *a* parameter were 0.01, 0.21, 0.04, and 0.07. Overall, the data sets in which students with misfitting response patterns were removed using the l_z provided smaller values of MAD for the *b* and *a* parameters.

Classification accuracy. The values of the classification accuracy are reported in Table 28. For the fitting data, 97.0% of the 4,210 students classified as low performing using their true ability parameterswere classified as low performing using the item parameter estimates derived from the Fit data set and 3.0% were classified as false positives. A lower percentage, 85.3%, of the 791 students classified as high performing using the true ability parameterswere classified as high performing using the item parameters derived from the Fit data set and 14.7% were classified as false negative. While the overall classification percentage for this data set was 95.1%, the accuracy was greater for the low performing students than the high performing students.

Table 28
Classification accuracy (25% of items and 10% of sample) for test with 60 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4082 (97.0)	128 (3.0)	4087 (97.1)	123 (2.9)	4084 (97.0)	126 (3.0)	4085 (97.1)	124 (2.9)
	High	116 (14.7)	675 (85.3)	177 (22.4)	614 (77.6)	174 (22)	617 (78.0)	175 (22.1)	616 (77.9)
% of total agreement		95.1		94.0		94.0		94.0	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

The percentage of correct decisions for the students with true low ability was essentially the same (approximately 97%) across the four data sets. In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to 77.6%, 78.0% and 77.9% for the misfit, l_z and H^T data sets, respectively. Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and constant for the remaining data sets (94.0%).

60 items, 20% of students with misfitting response patterns in 25% of items

The results for the 60 item test with 20% of the students with misfitting responses to 25% of the items are presented graphically in Figure 18. The full set of results are reported Table A13 in Appendix A.

b parameter: Items 27, 40, and 46, which had the highest positive true *b* parameter, had estimation bias beyond ± 0.20 across all three misfitting data sets. Items 51 and 6, which were the sixth and seventh most difficult items, had estimation bias larger than 0.20 for data set with all the misfitting response patterns. The five items with the lowest negative true *b* parameters and the item with the sixth lowest true *b* value had estimation bias beyond ± 0.20 only for data set with all the misfitting response patterns. Item 59, which was the most difficult item, also had estimation bias below -0.20 for the data sets with misfitting response patterns removed by l_z and by H^T. As for the previous condition, whereas items with large negative *b* parameters tended to be underestimated (e.g., Item 59), items with large positive *b* parameters tended to be overestimated (e.g., Item 46).

a parameter: The *a* parameter was underestimatedfor the data set with all misfitting response patterns for the 27 items with the largest true *a* parameters. Further, items 2, 4, 6, 8, 11,

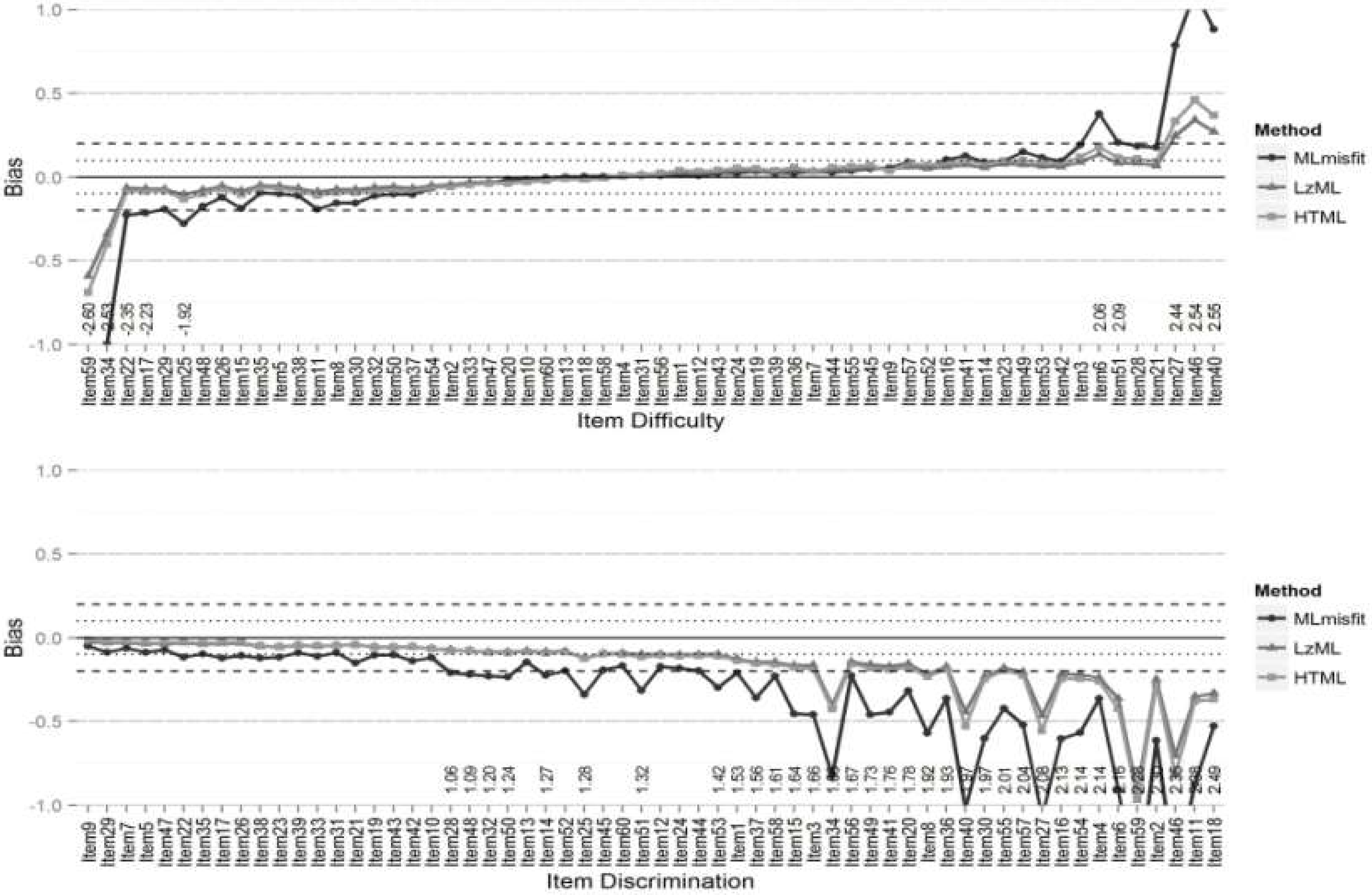

Figure 18: Bias of estimation for manipulating 25% of items and 20% of sample for test with 60 items

16, 18, 27, 30 , 34, 40, 46, 54, 57 and 59 had below -0.20 for the data sets with misfitting response patterns removed by l_z and H^T. Seven additional items with lower values of true a also had bias for the data set with all the misfitting response patterns.

Both b and a parameters: Of the 10 items identified with high bias for b and the 34 items identified with high bias for a, eight items were in common. The items were Item 6 (true b of 2.06, true a of 2.16), Item 25 (true b of -1.92, true a of 1.28), Item 27 (true b of 2.44, true a of 2.08), Item 34 (true b of -2.53, true a of 1.66), Item 40 (true b of 2.55, true a of 1.97), Item 46 (true b of 2.54, true a of 2.35), Item 51 (true b of 2.09, true a of 1.32) and Item 59 (true b of - 2.60, true a of 2.28). What is common to these items is that they were among the items that had relatively high (in absolute value) true b and true a values.

Comparison a and b parameters. In comparison to the b parameter, a greater number of items with marked bias in the a parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the a parameter than for the b parameter.

The results in Table 29 reveal that there were more highly biased estimates of the item discrimination parameter in the misfitting data set (34) compared to the difficulty parameter (10). The use of l_z to remove misfitting response patterns and the use of H^T to remove misfitting response patterns resulted in same biased items for both b and a parameters(5 for the b parameter and 15 for the a parameter).

Table 29
Bias of estimation across different methods (25% of items and 20% of sample)for test with 60 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	32	53	47
	±0.20 to ±0.10	18	2	8
	Beyond ±0.20	10	5	5
Item discrimination	-0.10 to 0.10	8	32	28
	±0.20 to ±0.10	18	13	17
	Beyond ±0.20	34	15	15

The values of the mean absolute deviation (MAD) for the b parameter were 0.01 for fitting data set, 0.18 for the data set with all the misfitting response patterns, 0.08 for the data set in which students with misfitting response patterns were removed using l_z, and 0.10 for the data set in which students with misfitting response patterns were removed using H^T (see Table A13, Appendix A). The corresponding values of MAD for the a parameter were 0.01, 0.35, 0.16, and 0.17. Overall, the data sets in which students with misfitting response patterns were removed using the l_z provided smaller values of MAD for the b and a parameters.

Classification accuracy. The values of the classification accuracy are reported in Table 30. For the fitting data, of the 4,205 students classified as low performing using their true ability parameters, 97.1% were classified as low performing using the item parameter estimates derived from the Fit data set and 2.9% were classified as false positives. Of the 795 students classified as high performing using the true ability parameters, 80.9% were classified as high performing using the item parameters derived from the Fit data set and 19.1% were classified as false negative. While the overall classification percentage for this data set was 94.6%, clearly the accuracy was greater for the low performing students than the high performing students.

Table 30
Classification accuracy (25% of items and 20% of sample) for test with 60 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4085 (97.1)	120 (2.9)	4074 (96.9)	131 (3.1)	4066 (96.7)	139 (3.3)	4068 (96.7)	137 (3.3)
	High	152 (19.1)	643 (80.9)	263 (33.1)	532 (66.9)	265 (33.3)	530 (66.7)	266 (33.5)	529 (66.5)
% of total agreement		94.6		92.1		91.9		91.9	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

The percentage of correct decisions for the students with true low ability was essentially the same (approximately 96.8%) across the four data sets. In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to

66.9%, 66.7% and 66.5% for the misfit, l_z and H^T data sets, respectively. Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (91.9% or 92.1%).

60 items, 30% of students with misfitting response patterns in 25% of items

The results for the 60 item test with 30% of the students with misfitting responses to 25% of the items are presented graphically in Figure 19. The full set of results are reported Table A14 in Appendix A.

b parameter: Three items with the highest b parameters with bias, the fifth, sixth, and eighth through tenth items with high b parameters had estimation bias beyond ± 0.20 for the data set with all the misfitting response patterns. Further, except of the second most difficult item (Item 10), these items had bias bias larger than 0.20 for the two data sets with misfitting response patterns removed. In the case of Item 10, only the data set with misfitting response patterns removed by H^T had bias. the data set with all the misfitting response patterns. The nine items with the lowest negative true *b* parameters had estimation bias beyond ± 0.20 only for data set with all the misfitting response patterns. Of these nine items, Items 21, 23, and 34 had estimation bias below -0.20 across all three data sets and Items 7 and 8 had estimation bias below -0.20 for the data set with misfitting response patterns removed by H^T. Again, as for the two previous conditions for the test with 60 items, items with large negative *b* parameters tended to be underestimated (e.g., Item 21) and items with large positive *b* parameters tended to be overestimated (e.g., Item 38).

a parameter: The *a* parameter was underestimatedfor the data set with all the misfitting response patterns for the 33 items with the largest true *a* parameters. Further, the bias in *a* was

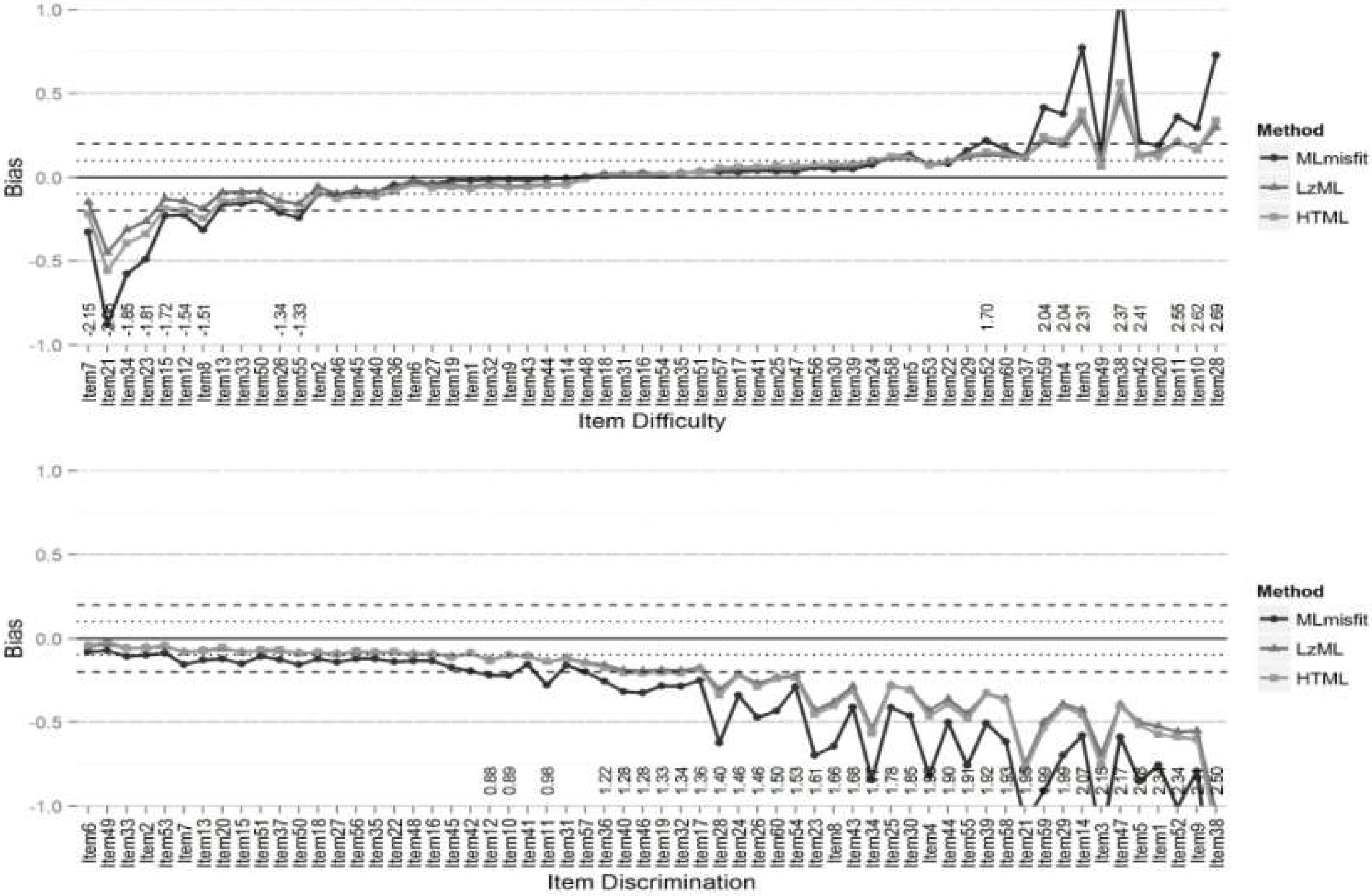

Figure 19: Bias of estimation for manipulating 25% of items and 30% of sample for test with 60 items

beyond ± 0.20 for the data set with misfitting response patterns removed by l_z and by H^T except for items 17 and 36 that showed bias below -0.20 only for the data set with all misfitting response patterns. Three additional items – Items 10, 11, and 12 – had bias in *a* the data set with all the misfitting response patterns.

Both b and a parameters: Of the 18 items identified with high bias for *b* and the 36 items identified with high bias for *a*, 15 items were in common. The items were Item 3 (true *b* of 2.31, true *a* of 2.15), Item 4 (true *b* of 2.04, true *a* of 1.90), Item 8 (true *b* of -1.51, true *a* of 1.66), Item 10 (true *b* of 2.62, true *a* of 0.89), Item 11 (true *b* of 2.55, true *a* of 0.98), Item 12 (true *b* of -1.54, true *a* of 0.88), Item 21 (true *b* of -2.06, true *a* of 1.95) , Item 23 (true *b* of -1.81, true *a* of 1.61) , Item 26 (true *b* of -1.34, true *a* of 1.46) , Item 28 (true *b* of 2.69, true *a* of 1.40) , Item 34 (true *b* of -1.85, true *a* of 1.77) , Item 38 (true *b* of 2.37, true *a* of 2.50) , Item 52 (true *b* of 1.70, true *a* of 2.34) , Item 55 (true *b* of -1.33, true *a* of 1.91) and Item 59 (true *b* of 2.04, true *a* of 1.99). Again common to these items is that they were among the items that had relatively high (in absolute value) true *b* and true *a* values.

Comparison a and b parameters. In comparison to the *b* parameter, a greater number of items with bias in the *a* parameter than items with bias in the *b* parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the *a* parameter than for the *b* parameter.

The results in Table 31 reveal that there were more biased estimates of item discrimination parameter in misfitting data set (36) compared to difficulty parameter (18). The use of l_z to remove misfitting response patterns resulted in slightly fewer biased item parameters than H^T specially for *b* parameter. For example, whereas the number of items with bias beyond

±0.20 in using l_z is 8 for the *b* parameter and 27 for the *a* parameter, the number of items with bias beyond ±0.20 using H^T is 12 for the *b* parameter and 31 for the *a* parameter.

Table 31
Bias of estimation across different methods (25% of items and 30% of sample)for test with 60 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	31	35	29
	±0.20 to ±0.10	11	17	19
	Beyond ±0.20	18	8	12
Item discrimination	-0.10 to 0.10	4	20	21
	±0.20 to ±0.10	20	13	8
	Beyond ±0.20	36	27	31

The values of the mean absolute deviation (MAD) for the *b* parameter were 0.01 for fitting data set, 0.18 for the data set with all the misfitting response patterns, 0.11 for the data set in which students with misfitting response patterns were removed using l_z, and 0.14 for the data set in which students with misfitting response patterns were removed using H^T (see Table A14, Appendix A). The corresponding values of MAD for the *a* parameter were 0.01, 0.41, 0.25, and 0.26. Overall, the data sets in which students with misfitting response patterns were removed using the l_z provided smaller values of MAD for the *b* and *a* parameters.

Classification accuracy. The values of the classification accuracy are reported in Table 32. For the fitting data, of the 4,208 students classified as

Table 32
Classification accuracy (25% of items and 30% of sample) for test with 60 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4081 (97.0)	127 (3.0)	4019 (95.5)	189 (4.5)	4052 (96.3)	157 (3.7)	4054 (96.3)	154 (3.7)
	High	169 (21.3)	623 (78.7)	310 (39.1)	482 (60.9)	327 (41.3)	465 (58.7)	329 (41.5)	463 (58.5)
% of total agreement		94.1		90.0		90.3		90.4	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

low performing using their true ability parameters, 97.0% were classified as low performing using the item parameter estimates derived from the Fit data set and 3.0% were classified as false positives. The percentage of high performing students classified as high performing, 78.7% , was again lower than the percentage of low performing students classified as low performing. Again, whie the overall classification percentage for this data set was 94.1%, the accuracy was greater for the low performing students than the high performing students.

The percentage of correct decisions for the students with true low ability was 97.0% for the fitted data set, 95.5% for the misfit data set, and 96.3% for the l_z and H^T data sets. In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to 60.9%, 58.7% and 58.5% for the misfit, l_z and H^T data sets, respectively. Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (90.0% to 90.4%).

60 items, 10% of students with misfitting response patterns in 50% of items

The results for the 60 item test with 10% of the students with misfitting responses to 50% of the items are presented graphically in Figure 20. The full set of results are reported Table A15 in Appendix A.

b parameter: The six items with the highest *b* parameter had estimation bias beyond ± 0.20 only for data set with all the misfitting response patterns. One other item with a high *b* parameter value (Item 55) had estimation bias beyond ± 0.20 and only for data set with all the misfitting response patterns Only Item 47 with the lowest negative true *b* parameter and Item 53 with the seventh lowest true b parameter had estimation bias at or beyond ± 0.20 and only for

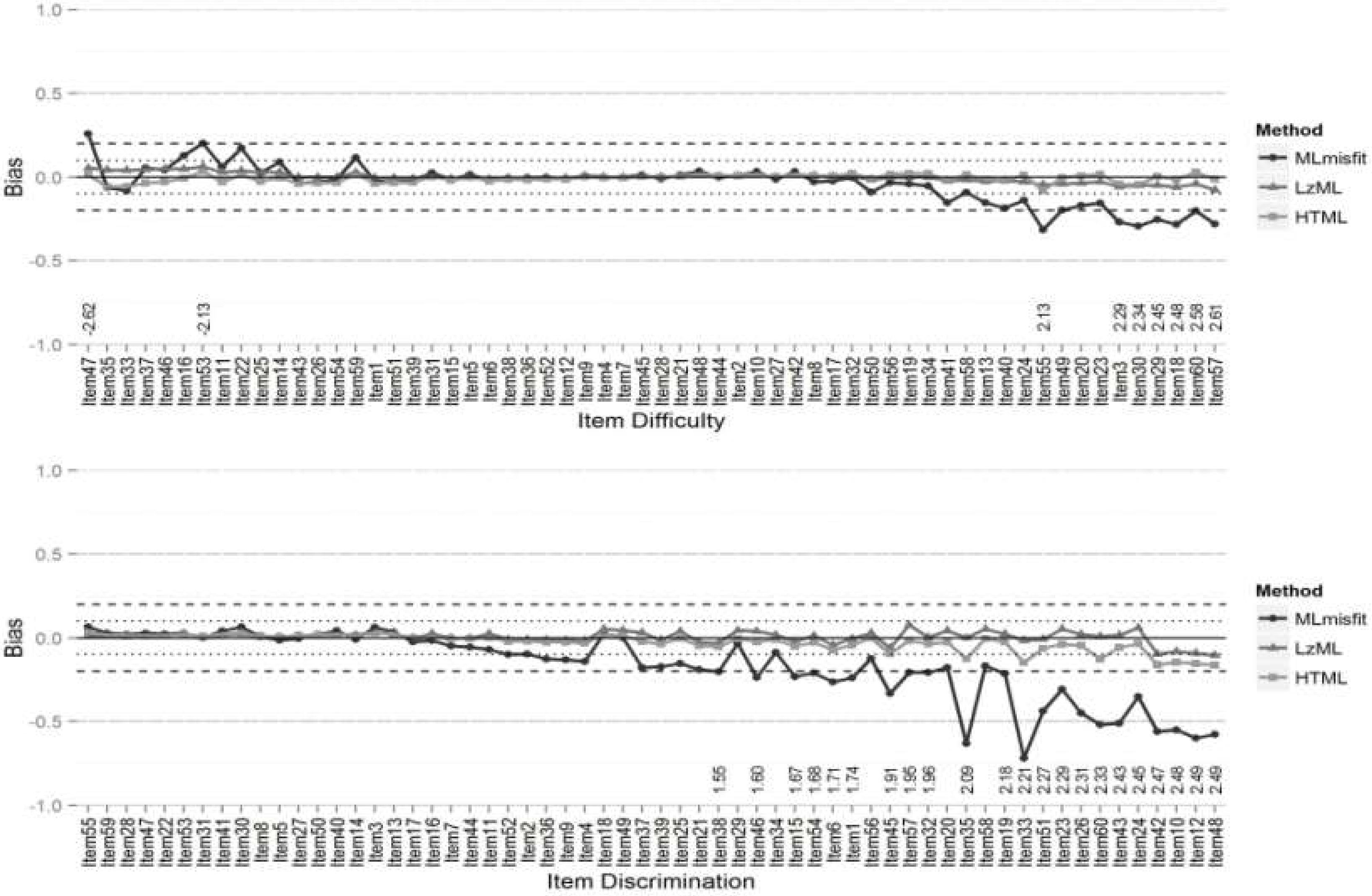

Figure 20: Bias of estimation for manipulating 50% of items and 10% of sample for test with 60 items

data set with all the misfitting response patterns. Again items with large negative *b* parameters tended to be overestimated (e.g., Item 47) and items with large positive *b* parameters tended to be underestimated (e.g., Item 57).

a parameter: The 12 items with the largest true *a* parameters had bias less than -0.20 in *a* but only of the data set with all misfitting response patterns. Ten additional items with lower values of *a* had estimation bias, but again only for the data set with all misfitting response patterns.

Both b and a parameters: Of the nine items identified with high bias for *b* and the 21 items identified with high bias for *a*, two items were in common. The items were Item 57 (true *b* of 2.61, true *a* of 1.95) and Item 60 (true *b* of 2.58, true *a* of 2.33). What is common to these items is that they were among the items that had relatively high (in absolute value) true *b* and true *a* values.

Comparison a and b parameters. In comparison to the *b* parameter, a greater number of items with marked bias in the *a* parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the *a* parameter than for the *b* parameter.

The results in Table 33 reveal that there were more biased estimates of item discrimination parameter in misfitting data set (22) compared to difficulty parameter (9). No

Table 33
Bias of estimation across different methods (50% of items and 10% of sample)for test with 60 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	41	60	60
	±0.20 to ±0.10	10	0	0
	Beyond ±0.20	9	0	0
Item discrimination	-0.10 to 0.10	28	59	53
	±0.20 to ±0.10	10	1	7
	Beyond ±0.20	22	0	0

items were found with bias in the b parameter or the a parameter when misfitting response patterns were removed using of l_z and using H^T.

The values of the mean absolute deviation (MAD) for the *b* parameter were 0.01 for fitting data set, 0.08 for the data set with all the misfitting response patterns, 0.02 for the data set in which students with misfitting response patterns were removed using l_z, and 0.02 for the data set in which students with misfitting response patterns were removed using H^T (see Table A15, Appendix A). The corresponding values of MAD for the *a* parameter were 0.01, 0.19, 0.03, and 0.04. Overall, data sets in which students with misfitting response patterns were removed using the l_z provided smaller values of MAD for the *b* and *a* parameters.

Classification accuracy. The values of the classification accuracy are reported in Table 34. For the fitting data, of the 4,208 students classified as low performing using their true ability parameters, 97.0% were classified as low performing using the item parameter estimates derived from the Fit data set and 3.0% were classified as false positives. Of the 792 students classified as high performing using the true ability parameters, 85.9% were classified as high performing using the item parameters derived from the Fit data set and 14.1% were classified as false negative. While the overall classification percentage for this data set was 95.2%, clearly the accuracy was greater for the low performing students than the high performing students.

The percentage of correct decisions for the students with true low ability was 94.4% for misfit data set, 93.8% for l_z and 93.9% for H^T data sets, which are approximately 3% less than for the fitted data. In contrast the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to 75.4%, 77.4% and 77.0% for the misfit, l_z and H^T data sets, respectively.

Table 34
Classification accuracy (50% of items and 10% of sample) for test with 60 items

		Fit		Misfit		l_z		H^T	
	True \ Estimated	Low	High	Low	High	Low	High	Low	High
ML	Low	4082 (97.0)	126 (3.0)	3972 (94.4)	237 (5.6)	3949 (93.8)	259 (6.2)	3951 (93.9)	257 (6.1)
	High	112 (14.1)	680 (85.9)	195 (24.6)	597 (75.4)	179 (22.6)	613 (77.4)	182 (23.0)	610 (77.0)
% of total agreement		95.2		91.4		91.2		91.2	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (91.2% or 91.4%).

60 items, 20% of students with misfitting response patterns in 50% of items

The results for the 60 item test with 20% of the students with misfitting responses to 50% of the items are presented graphically in Figure 21. The full set of results are reported Table A16 in Appendix A.

b parameter: The bias of estimation for the *b* parameter occurred in the tails of the distribution of *b* parameters. The two items with the highest true *b* values and the sixth, eighth, and eleventh most difficult items had had estimation bias beyond ± 0.20 only for data set with all the misfitting response patterns. The most difficult item, Item 11, also had estimation bias beyond ± 0.20 for the data set with misfitting response patterns removed by l_z. The 15 items with the lowest negative true *b* parameters had estimation bias at or beyond ± 0.20 for data set with all the misfitting response patterns. Further, the 11 most easy items had estimation bias at or beyond ± 0.20 for data set with misfitting response patterns removed by lz and the four most easy items

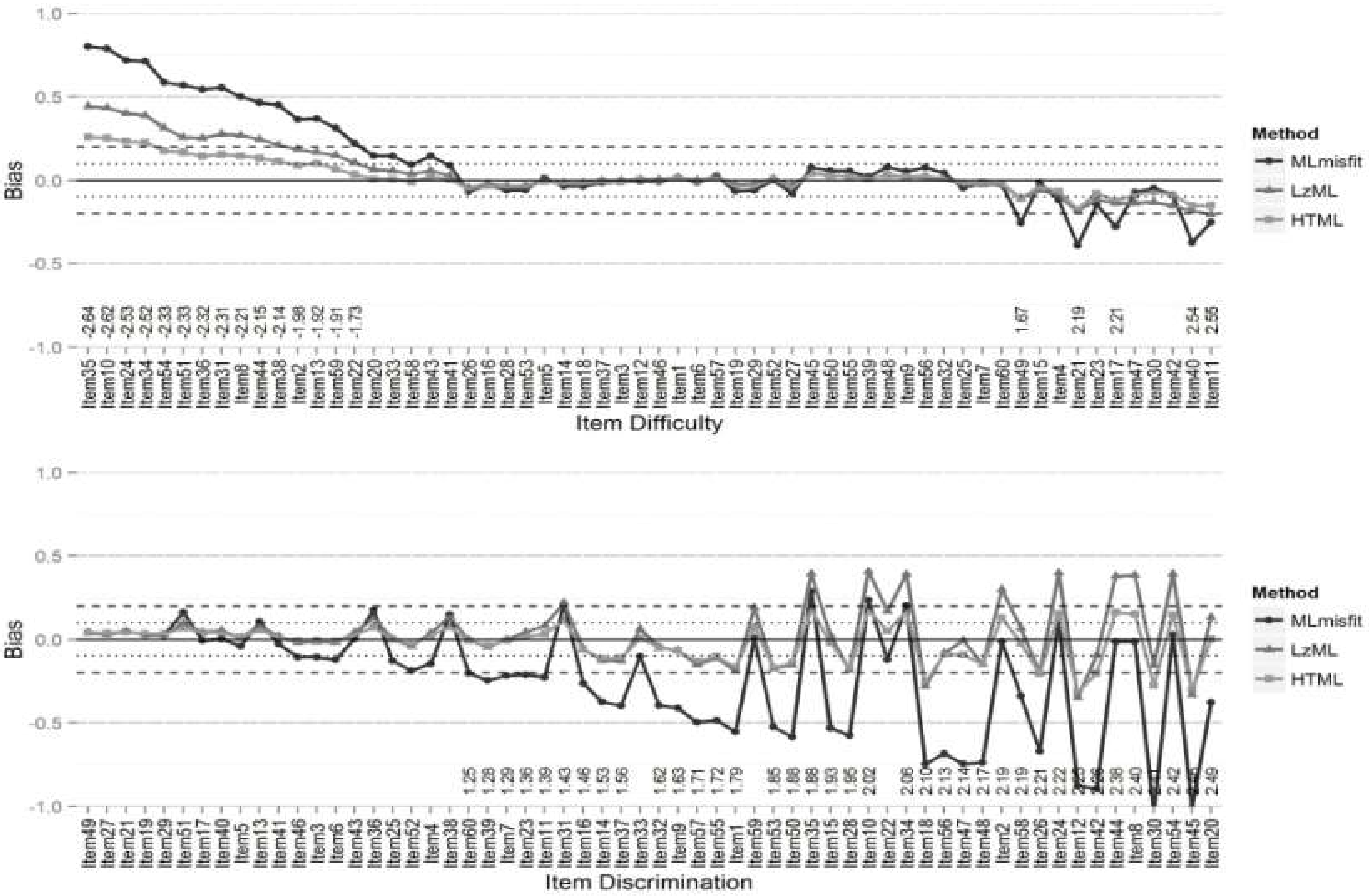

Figure 21: Bias of estimation for manipulating 50% of items and 20% of sample for test with 60 items

had estimation bias at or beyond ± 0.20 across all three data sets. Again, whereas items with large negative *b* parameters tended to be overestimated (e.g., Item 35), items with large positive *b* parameters tended to be underestimated (e.g., Item 40).

a parameter: The bias in the *a* parameter was beyond ± 0.20 beyond for at least one of the data sets for the 17 items with the largest true *a* parameters. For example, the bias for Items 20, the item with the highest true *a* value, was negative and occurred only of the data set with all the missing response patterns. The item with the second highest true value of *a* was underestimated for all three data sets. The item with the third highest true value of a was overestimated only for the data set with misfitting response patterns removed using l_z. And the item with the fourth highest true value of *a* was underestimated for the data set with all the misfitting response patterns and the data set with misfitting response patterns removed using H^T. These examples illustrate the complexity of the pattern of bias for the 17 items with the largest true a parameters and the other 20 items with at least one data set with bias beyond ± 0.20. Across the full set of item with bias, the bias tends to be negative for the data set with all the misfitting response patterns, positive for the data set with misfitting response patterns removed by lz, and negative for the data set with misfitting response patterns removed using H^T.

Both b and a parameters: Of the 20 items identified with high bias for *b* and the 38 items identified with high bias for *a*, five items were in common. The items were Item 10 (true *b* of -2.61, true *a* of 2.02), Item 11 (true *b* of 2.55, true *a* of 1.39), Item 31 (true *b* of -2.31, true *a* of 1.43), Item 34 (true *b* of -2.52, true *a* of 2.06) and Item 35 (true *b* of -2.64, true *a* of 1.88). What is common to these items is that they were among the items that had relatively high (in absolute value) true *b* and true *a* values.

Comparison a and b parameters. In comparison to the *b* parameter, a greater number of items with marked bias in the *a* parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the *a* parameter than for the *b*. The results in Table 35 reveal that there were more highly biased estimates of the item discrimination parameter in the misfitting data set (32) compared to the difficulty parameter (20). The use of H^T to remove misfitting response patterns resulted in fewer biased item parameters than l_z for both the *b* and *a* parameters.

Table 35
Bias of estimation across different methods (50% of items and 20% of sample)for test with 60 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	35	36	43
	±0.20 to ±0.10	5	12	13
	Beyond ±0.20	20	12	4
Item discrimination	-0.10 to 0.10	15	32	36
	±0.20 to ±0.10	13	16	18
	Beyond ±0.20	32	12	6

For example, whereas the number of items with bias beyond ±0.20 in using l_z is 12 for both the *b* and *a* parameters, the number of items with bias beyond ±0.20 using H^T is 4 for the *b* parameter and 6 for the *a* parameter.

The values of the mean absolute deviation (MAD) for the *b* parameter were 0.01 for fitting data set, 0.20 for the data set with all the misfitting response patterns, 0.11 for the data set in which students with misfitting response patterns were removed using l_z, and 0.07 for the data set in which students with misfitting response patterns were removed using H^T (see Table A16, Appendix A). The corresponding values of MAD for the *a* parameter were 0.01, 0.30, 0.14, and 0.09. Overall, data sets in which students with misfitting response patterns were removed using the H^T provided smaller values of MAD for the *b* and *a* parameters.

Classification accuracy. The values of the classification accuracy are reported in Table 36. For the fitting data with ML estimate of item parameters, 96.9% of the 4,209 students classified as low performing using their true ability parameterswere classified as low performing using the item parameter estimates derived from the Fit data set and 3.1% were classified as false positives. Again lower, 87.0% of the 790 students classified as high performing using the true ability parameterswere classified as high performing using the item parameters derived from the Fit data set and 13.0% were classified as false negative. While the overall classification percentage for this data set was 95.4%, the accuracy was greater for the low performing students than the high performing students.

Table 36
Classification accuracy (50% of items and 20% of sample) for test with 60 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4080 (96.9)	129 (3.1)	3903 (92.7)	306 (7.3)	3847 (91.4)	363 (8.6)	3844 (91.3)	366 (8.7)
	High	103 (13.0)	687 (87.0)	258 (32.6)	533 (67.4)	244 (30.9)	546 (69.1)	242 (30.6)	549 (69.4)
% of total agreement		95.4		88.7		87.9		87.8	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

The percentage of correct decisions for the students with true low ability was 92.7% for misfit data set, 91.4% for l_z and 91.3% for H^T data sets, which are approximately 4% lower than for the Fit data set. In contrast, the percentage of correct decisions for the students with true high ability dropped from the percentage for the fitted data to 67.4%, 69.1% and 69.4% for misfit, l_z and H^T data sets, respectively. Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (87.8% to 88.7%).

60 items, 30% of students with misfitting response patterns in 50% of items

The results for the 60 item test with 30% of the students with misfitting responses to 50% of the items are presented graphically in Figure 22. The full set of results are reported Table A17 in Appendix A.

b parameter: The six items with the highest true b values had estimation bias beyond ± 0.20 across all three data sets. Four other items – Items 24, 60, 23, and 58 – had estimation bias below -0.20 for data set with all the misfitting response patterns and for Items 24, also the data set with misfitting response patterns removed by H^T. The 13 items with the lowest true *b* parameters had estimation bias at or above 0.20 for the data set with all the misfitting response patterns; the 12 with the lowest true b values had bias for the data set with misfitting response patterns removed by l_z; and the 10 items had bias for the data set with misfitting response patterns removed by H^T. Three other items with negative true values of b – Items 50, 45, and 9 – had bias across all three data sets Again, items with large negative *b* parameters tended to be overestimated (e.g., Item 22) and items with large positive *b* parameters tended to be underestimated (e.g., Item 13).

a parameter: For the with the previous condition, both negative and positive bias in *a* were found, but in contrast to the previous example, the presence of negative bias increases with increasing *a*. For example, within the set of 12 items with the highest true *a* values, the bias is positive only for the data set with misfitting response patterns removed by *lz* for Item 15; the remaining biases for these 12 items are negative and occur for the data set with all the misfitting response patterns and the data set with misfitting response patterns removed by H^T. . In the next set of 12 items with large true a values, positive bias occurred data set with missing responses removed by lz for three items and one data set with misfitting responses removed by H^T for two

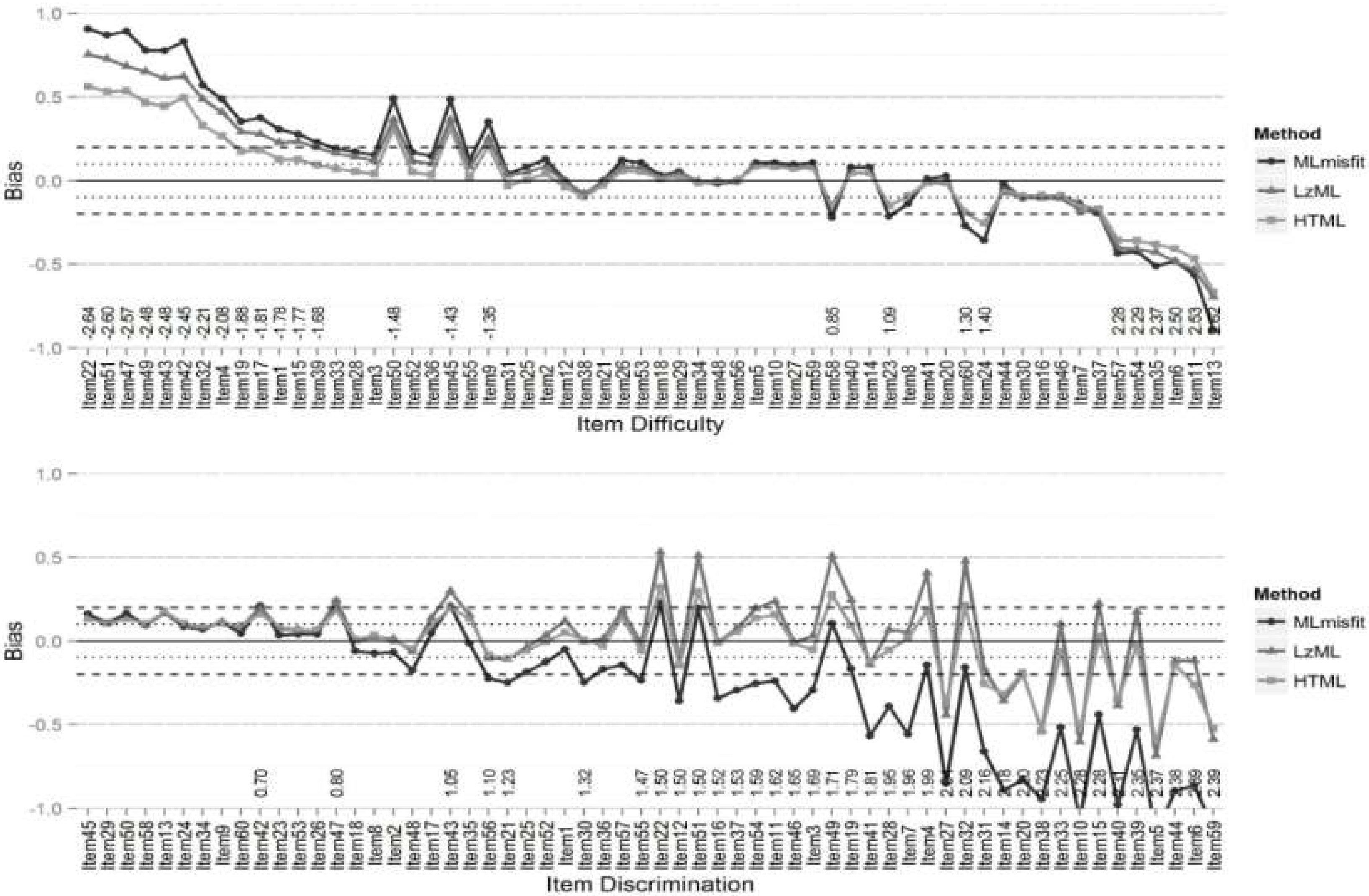

Figure 22: Bias of estimation for manipulating 50% of items and 30% of sample for test with 60 items

of these three items; the remaining biases, which occurred for the data set with all the missing response patterns, was negative. This pattern continued for the remaining 13 items that had bias for at least one data set.

Both b and a parameters: Of the 27 items identified with high bias for *b* and the 37 items identified with high bias for *a*, nine items were in common. The items were Item 6 (true *b* of -2.50, true *a* of 2.39), Item 11 (true *b* of 2.53, true *a* of 1.62), Item 15 (true *b* of -1.77, true *a* of 2.28), Item 22 (true *b* of -2.64, true *a* of 1.50), Item 39 (true *b* of -1.68, true *a* of 2.35) , Item 42 (true *b* of -2.45, true *a* of 0.70) , Item 43 (true *b* of -2.48, true *a* of 1.05) , Item 47 (true *b* of -2.57, true *a* of 0.80) and Item 54 (true *b* of 2.29, true *a* of 1.59). What is common to these items is that they were among the items that had relatively high (in absolute value) true *b* and true *a* values.

Comparison a and b parameters. In comparison to the *b* parameter, a greater number of items with marked bias in the *a* parameter due to presence of misfitting response patterns were found. Further, the bias was larger for the *a* parameter than for the *b* parameter and the pattern of bias was more variable for *a* than for *b*.

The results in Table 37 reveal that there were more highly biased estimates of item discrimination parameter in the misfitting data set (32) compared to the difficulty parameter (26). The use of H^T to remove misfitting response patterns resulted in slightly fewer biased item parameters than l_z for both the *b* and *a* parameters. For example, the number of items with bias beyond ±0.20 in using l_z is 22 for the *b* parameter and 18 for the *a* parameter and the number of items with bias beyond ±0.20 using H^T is 18 for the *b* parameter and 14 for the *a* parameter.

Table 37
Bias of estimation across different methods (50% of items and 30% of sample)for test with 60 items

		Misfit	l_z	H^T
Item difficulty	-0.10 to 0.10	18	25	33
	±0.20 to ±0.10	16	13	9
	Beyond ±0.20	26	22	18
Item discrimination	-0.10 to 0.10	13	22	27
	±0.20 to ±0.10	15	20	19
	Beyond ±0.20	32	18	14

The values of the mean absolute deviation (MAD) for the *b* parameter were 0.01 for fitting data set, 0.27 for the data set with all the misfitting response patterns, 0.22 for the data set in which students with misfitting response patterns were removed using l_z, and 0.17 for the data set in which students with misfitting response patterns were removed using H^T (see Table A17, Appendix A). The corresponding values of MAD for the *a* parameter were 0.01, 0.34, 0.19, and 0.15. Overall, data sets in which students with misfitting response patterns were removed using the H^T provided smaller values of MAD for the *b* and *a* parameters.

Classification accuracy. The values of the classification accuracy are reported in Table 38. For the fitting data, of the 4,204 students classified as low performing using their true ability parameters, 97.0% were classified as low performing using the item parameter estimates derived from the Fit data set and 3.0% were classified as false positives. Of the 796 students classified as high performing using the true ability parameters, 85.4% were classified as high performing using the item parameters derived from the Fit data set and 14.6% were classified as false negative. While the overall classification percentage for this data set was 95.2%, clearly the accuracy was greater for the low performing students than the high performing students.

The percentage of correct decisions for the students with true low ability was 89.0% for misfit data set, 86.8% for l_z and 87.5% for H^T data sets, which are approximately 10% less than for the fitted data. In contrast the percentage of correct decisions for the students with true high

Table 38
Classification accuracy (50% of items and 30% of sample) for test with 60 items

		Fit		Misfit		l_z		H^T	
	Estimated True	Low	High	Low	High	Low	High	Low	High
ML	Low	4078 (97.0)	126 (3.0)	3742 (89.0)	462 (11.0)	3649 (86.8)	555 (13.2)	3680 (87.5)	524 (12.5)
	High	116 (14.6)	680 (85.4)	345 (43.3)	451 (56.7)	324 (40.7)	472 (59.3)	329 (41.3)	467 (58.7)
% of total agreement		95.2		83.9		82.4		82.9	

Note: Fit= agreement between fitting data set and true classification; Misfit= agreement between misfitting data set and true classification; l_z= agreement between fitting data set based on l_z and true classification; H^T agreement between fitting data set based on H^T and true classification. Numbers in brackets are row percentages for each of Fit, Mistfit,, l_z, and H^T.

ability dropped from the percentage for the fitted data to 56.7%, 59.3% and 58.7% for misfit, l_z and H^T data sets, respectively. Consequently, the percentage of false negative misclassifications is higher than the percentage of false positive misclassifications across all conditions. Lastly, the overall classification percentage was highest for the Fit data set and essentially constant for the remaining data sets (82.4% to 83.9%).

Summary of results for test with 60 items

A summary of findings for simulation study examining estimation bias of *b* and *a* parameters in a test with 60 items is provided as follows:

b parameter: The results showed that an increase in the percentage of misfitting response patterns in a sample led to larger biased estimates of the *b* parameter. Estimation bias occurred for items with large negative or positive true *b* values whereas items with true values of *b* in the middle of distribution were less affected by presence of misfitting response patterns. In situation where 25% of items were susceptible to misfitting responses, the large negative *b* values were underestimated and large positive *b* values were overestimated due to presence of misfitting response patterns. But in situation where 50% of items were susceptible to be answered misfittingly, the large negative *b* values were overestimated and large positive *b* values were underestimated due to presence of misfitting response patterns. The degree of estimation bias was more pronounced under the condition when 25% of items were susceptible to misfitting responses than when 50% of items were susceptible to misfitting responses.

a parameter: The number of items with bias in the a parameter and the size of the bias in a increased as the percentage of misfitting response patterns increased. Further, the bias was not found in the items with lowest true a values. When bias was found, the true a parameters were consistently underestimated across all conditions when 25% of items were susceptible misfitting responses. In contrast, the true a parameters were less consistently underestimated for the conditions in which 50% of items were susceptible misfitting responses.

Comparison a and b parameters: The a parameter was more affected by presence of misfitting response patterns than the b parameter across all the data sets for all conditions. There were items with estimation bias for both the b and a parameters. What was common to these items was that they were among the items that had relatively high (in absolute value) true b and true a values. Using l_z to remove misfitting response patterns resulted in less biased estimates for both the b and a parameters, particularly when the percentage of misfitting response patterns in sample were 10% or 20%. In conditions where the percentage of misfitting response patterns in the sample and the percentage of susceptible items was high, using H^T resulted in less bias.

Classification accuracy: The patterns of classification of students into two categories were similar for each of the conditions. The classification accuracy was higher and essentially equal (> 86%) across the four data sets for the low performing students. For the high performing students, the classification accuracy was lower than 90% for the fitting data set and even lower but essentially equal for the misfitting, l_z and H^T sets. Taken together, despite the presence of bias in the b parameter and, to a greater extent, the a parameter, using the person fit indices l_z

and H^T to remove misfitting response patterns did not improve classification accuracy over not removing misfitting response patterns.

Chapter Seven: Conclusion

Introduction

The problem of misfitting responding has been studied by many researchers (Rupp, 2013; Meijer & Sijtsma, 2001). Misfitting refers to the mismatch between the observed response patterns and the expected response patterns of students who have responded to a test. Previous studies suggested that the validity and reliability of a test might be compromised because of the existence of misfitting responses in the test data (Meijer, 1997; Schmitt et al, 1993). This is mainly due to the effect of the presence of misfitting response patterns on the estimation of student's ability. The existence of misfitting response patterns can distort the shape of likelihood function and result in incorrect ability estimates.

The procedure to assess the fit of a student's response pattern to the measurement model is referred to as "person fit analysis". To date, the majority of the person fit literature has been heavily focused on creating and evaluating new indices (Rupp, 2013). Only five studies (Levine & Drasgow, 1982; Philips, 1986; Runder, Bracey, & Skaggs, 1996; Hendrawan et al., 2005; Sotaridona et al., 2005) have been conducted to investigate the effect of misfitting response patterns on the estimates of the item parameters and the findings have been mixed. If the item parameters are affected, this in turn might affect the ability estimates.

Therefore, the purpose of the present study was to investigate the effect of the inclusion and exclusion of misfitting response patterns on item parameter estimates under using simulated data. The factors considered included test length, item parameter estimation method, percentage of students who responded misfittingly in the sample, and percentage of items susceptible to misfitting responses in a 3 x 2 x 3 x 2 fully crossed design. Two indices for detecting misfitting response patterns, namely l_z (Drasgow et al., 1985) and H^T(Sijtsma, 1986) were used. The 2-

parameter IRT model was used for all analyses. The number of students was 5,000 and the number of replications was 100 for each condition. The bias in the *b* and *a* item parameters was used to assess the effect of misfitting response patterns on the two item parameter estimates. The effect of biased item parameter estimates due to the presence of misfitting response patterns on classification accuracy was also considered as a third dependent variable.

Summary of Results

The findings of the simulation study showed almost consistent patterns of results across the three test lengths. Those findings can be summarized as follows:

- No difference was found between Maximum Likelihood and Bayes Modal estimates of items parameter in terms of the effect of misfitting response patterns on the estimated item parameters.
- The bias in the *b* parameter occurred for items in the tails of the distribution and increased as the percentage of misfitting response patterns increased.
- Whereas the large negative *b* values were underestimated and large positive *b* values were overestimated for the conditions in which 25% of the items were susceptible to misfitting responses, the large negative *b* values were overestimated and large positive *b* values were underestimated more so in the data set with all misfitting response patterns than in the data sets in which misfitting response patterns were removed using l_z and H^T for the conditions in which 50% of items were susceptible to misfitting responses.
- The bias in the *a* parameter tended to occur for items with large true *a* parameters.
- The *a* parameter was consistently underestimated for all conditions in which 25% of the items were susceptible to misfitting responses. For the conditions in which 50%

of the items were susceptible to misfitting responses, both underrepresentation and overestimation were found, with the overestimation occurring in the data sets with misfitting responses removed by l_z.

- The number of items with bias in a and the size of the bias in a increased as the percentage of misfitting response patterns increased.
- The a parameter was more affected by presence of misfitting response patterns than the b parameter both in terms of the number of items and the size of the bias for all conditions.
- Using l_z to remove misfitting response patterns resulted in less biased estimates for both the b and a parameters, particularly when the percentage of misfitting response patterns in sample were 10% or 20%. In conditions where the percentage of misfitting response patterns in the sample and the percentage of susceptible items was high, using H^T resulted in less bias.
- The patterns of classification of students into two categories were similar for all conditions.
- The classification accuracy was high (> 0.90) and essentially equal across the four data sets (i.e., fitting, misfitting, l_z , and H^T) for the low performing students. For the high performing students, the classification accuracy was lower (< 0.90) for the fit data set and dropped even lower for the misfitting, l_z and H^T data sets for which the classification accuracy was essentially the same.
- The use of the person fit indices l_z and H^T to remove misfitting response patterns did not improve classification accuracy over not removing misfitting response patterns.

Discussion

Levine and Drasgow (1982) used the 3-PL IRT model to study 1) the effect of using estimated item parameters versus true item parameters on the detection rate of l_0 and 2) the effect of misfitting response patterns on the detection rate of l_0 and item parameter estimation. Estimated item parameters from a previous calibration study of verbal section of the Scholastic Aptitude Test were used to generate simulated data. The spuriously low responding was simulated using about 7% of the sample in which 20% of the item scores of the selected students were changed to be correct with a probability of 0.20 and incorrect with a probability of 0.80. Levine and Drasgow concluded that the presence of misfitting response patterns had no effect on the detection rate of l_0 and item parameter estimation. They argued that different misfitting responses tended to have different incorrect response patterns and that, consequently a large number of the misfitting response patterns would have opposing effects on estimated item parameters. Their findings are comparable with the findings of the present study: when the percentage of misfitting response patterns was low (e.g., 10%), the effect of misfitting response patterns on the estimated item parameters were small. But the findings of the present study contradict the finding of Leven and Drasgow in that an increase in the percentage of misfitting response patterns led to a greater number of biased item parameters and a greater size of bias.

Philips (1986) investigated the effect of misfitting response vectors on the fit of the Rasch model, the value of the estimated *b* parameters, and the results of equipercentile equating using 1980 national standardization of the 3-R's K-12 multilevel achievement test battery at 4th and 8th grades for reading and mathematics sub-tests. The percentages of misfitting response patterns varied from 1.9% to 11.3% across the two sub-tests. He found that deletion of misfitting response patterns can improve the model-data fit, had small effect on the estimated *b* parameters, and did not influence the equating results. His findings are partly consistent with the findings of the present study in that the effect of misfitting response patterns on the *b* parameter was less than on the *a* parameter. Runder, Bracey, and Skaggs (1996) analyzed the 1990 NAEP Trail State Assessment data and found almost no misfiitting response vectors. Consequently, removing misfitting patterns did not result in significant difference in the mean of the test before and after deleting the misfitting response vectors.

Sotaridona et al. (2005) studied the effect of misfitting response patterns on item calibration and performance classification. In their study, they used l_z^* (Snijders, 2001) and *U3* (van der Flier, 1982) as the person fit measures and worked with a random sample of 10,000 students from a statewide assessment comprising three subject areas (i.e., Mathematics, Science and Language Arts). Two types of misfitting responding (i.e., copying and guessing) were simulated by manipulating data of selected students (10% of students). The data sets were calibrated independently using the 3-PL IRT model. Results of their simulation study showed that the *b* and *a* parameters were, with one exception, overestimated in the presence of person misfit and the standard errors of estimation were higher for the data with misfitting response patterns. The exception was for the guessing data set where the item parameters were underestimated on one occasion. Since the range of item parameter values were not reported by Sotaridona et al., it is hard to make a direct comparison with the present study. But based on the graphs provided in Sotaridona et al. (2005), it seems that the differences between difficulty parameters estimated from the original data sets and manipulated data sets were much higher than the differences for the discrimination parameter. This finding is contradictory to finding of the present study in that the *a* parameter was affected more by the presence of misfitting response patterns than the *b* parameter.

In addition to their original data set, Sotaridona et al. (2005) created two additional data sets: a data set comprising fitting responses flagged by l_z^* and another data set containing fitting responses flagged by *U3*. Each data set was calibrated and equated and standardized scale scores were converted into three levels (below proficiency, proficient, and advanced). The total classification agreement between original data set and fitting data set using l_z^* varied from 96% to 100% and the total classification agreement between original data set and fitting data set using *U3* varied from 94% to 100%, across subject matters. Sotaridona et al. concluded that inclusion of misfitting response patterns reduced the accuracy of the parameter estimates, but at the test level the effect was minimal. This finding is consistent with the finding of the present study where the total classification agreement was essential the same for data set with all misfitting response patterns and the data sets with misfitting response patterns removed by l_z and by H^T . However, the total agreement was lower for these three data sets and the data set with no misfitting responses had the highest total agreement.

Hendrawan et al. (2005) investigated the effect of misfitting response vectors on classification decisions. They used three true values of discrimination parameter – 0.5, 1.0, and 1.5 – crossed with true difficulty parameters that ranged from -2.0 to 1.6 for a test with 30 items and from -2.0 to 1.8 for a test with 60 items. They used three ability estimation methods – MLE, EAP, and Markov Chain Monte Carlo (MCMC) – with the three parameter normal ogive model and marginal maximum likelihood (MML) and Bayes Modal to estimate the item parameters. Five person fit indices were utilized in the study in addition to the l_z. The simulation factors included two test lengths, two misfitting response types, three item discrimination values, two sample sizes and three cutoff values for determining proficiency levels on the test. For generating misfitting response patterns, 10% of the samples were manipulated. Their results showed that the presence of misfitting response patterns resulted in biased estimates of item parameters and inaccurate mastery classification decisions, especially for guessing behavior which lowered the mean of distribution of estimated abilities. However the degree and direction of bias were not reported. The lower mean resulted in artificially higher classification accuracy for students with low ability. All person fit indices performed well and removal of students with misfitting response patterns resulted in increased classification decisions. As an overall conclusion, authors argued that person fit statistics are useful in finding fitting subsamples and are appropriate for using in mastery testing. The study by Hendrawan et al. (2005) was heavily focused on classification accuracy. They used three true values of the discrimination parameter and a restricted range of difficulty parameters and only mentioned that estimated parameters were biased but did not give the degree and direction of bias. Consequently, it is not possible to compare the findings for the estimated parameters of Hendrawan et al.'s study with the findings of the present study. However, the findings for classification accuracy are consistent with findings of classification accuracy in the present study.

The percentage of misfitting response patterns used in the previous study were essentially lower than or equal to 10% and this could be one of the reasons for not finding more and larger effects of removing misfitting response patterns on the estimated item parameters. Further, there was little information about details of the behavior of estimated item parameters before and after removing misfitting response patterns. The results of the present study revealed that the effect of misfitting response patterns on the estimation of b and a parameters is dependent on the number of items, percentage of students with misfitting response patterns, and the percent of item susceptible to misfitting responses. However, despite the presence of the effects, which were complex, on the b and a parameters, removal of misfitting response patterns by l_z and by H^T did not change the results of classification from when all the misfitting response patterns were included.

Limitations of the Study

The findings of the present study are bounded by the factors and design considered. This restricts generalization of the findings to similar conditions. Further, some of the conditions simulated may not be realistic. For example, it is unlikely that 30% of the students will produce misfitting response patterns or that 50% of items in a test will be susceptible to misfitting responses. Additionally, given the factors and associated levels considered in this study any claim about the effect of factors on the estimation bias, standard error of estimation and classification accuracy is limited to the design of this study.

It was assumed that the data fit the 2-PL IRT model across all simulation conditions. However, the assumption of data fit with the 2- PL IRT model may not be met, especially in the presence of misfitting response patterns in the data. Furthermore, simulation studies are based on some presumed assumptions that may not completely reflect the real-world situations and this study is not an exception. Thus, generalizability of results of this study to real-world conditions is constrained.

Conclusion

Based the results of this study and in light of the limitations of the study, the presence of misfitting response patterns created bias in the *b* and *a* parameters at the item level which in turn affected the classification of students, particularly high performing students, into performance categories regardless of whether students with misfitting response patterns were present in the data or were removed using l_z or H^T. Clearly the results differed by test length, the percentage of students with misfitting response patterns, and the percentage of items susceptible to misfitting responses.

Implication for Practice

The results of this study imply that if the difference between estimated item parameters before and after removing misfitting response patterns is considerable, then test should be investigated by test developers and psychometricians for potential causes. The effect of misfitting response patterns should be considered at the item level, especially when it is suspected the percentage of misfitting response patterns in the sample is high. If bias is found, then the students with misfitting response patterns should be removed from the data set using person fit indices.

One good strategy on utilizing person fit indices for refining item parameter estimation could be employing an iterative process of removal. Only one round of removing misfitting response patterns was applied in this study. The process of removal of misfitting response patterns and re-estimation of item parameters can be continued until some evidence of item parameter estimation stability is seen.

Although, we may not be able to see substantial effect of misfitting response patterns on the estimated item parameters in the context of large scale assessments but in other contexts like psychological testing person fit analysis can provide informative results. one example could be Fernando (2012).

Recommendation for Future Research

The person fit assessment is a growing field of research. Based on the findings of this study, the following topics can be considered for future research:

1- Using the 3PL IRT model is recommended in another study since the pseudo-guessing parameter may account for some degree of person misfit.

2- Other factors such as different combinations of item parameters can provide more information about the effect of misfitting response patterns on estimated item parameters.

3- The effect of misfitting response patterns on the item and model-data fit using 2PL and 3PLIRT models need to be studied. Previous studies such as Philips (1986) found improvement in Rasch model item fit after removing misfitting response patterns.

4- Using the recently developed corrected version of the l_z for mixed item format tests (Sinharay, 2015) needs to be studied.

References

Armstrong, R. D., & Shi, M. (2009). A IRT-based cumulative sum statistic for person fit. *Applied Psychological Measurement*, 33, 391-410.

Armstrong, R. D., Stoumbos, Z. G., Kung, M. T., & Shi, M. (2007). On the Performance of the lZ Person-Fit Statistic. *Practical Assessment, Research & Evaluation*, 12(16).

Baker, F., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (Vol. 176). CRC Press.

Choi, H.-J., & Cohen, A. S. (2008, March).*A Bayesian approach to the estimation of person-fit in the testlet model*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New York, NY.

Clark, J. M. III. (2010). *Aberrant response patterns as a multidimensional phenomenon: Using factor-analytic model comparison to detect cheating* (Unpublished doctoral dissertation). Lawrence, KA; University of Kansas.

D'Costa, A. (1993a, April). *Extending the Sato caution index to define the within and beyond ability caution indexes*. Paper presented at convention of National Council for Measurement in Education, Atlanta, GA.

de Ayala, R.J. (2008). *The Theory and Practice of Item Response Theory*, New York, NY: The Guilford Press.

De la Torre. J., & Deng, W. (2008).Improving Person-Fit Assessment by Correcting the Ability Estimate and Its Reference Distribution. *Journal of Educational Measurement*,45, 2, 159-177

Dimitrov, D. M., & Smith, R. M. (2006).Adjusted Rasch person-fit statistics. *Journal of Applied Measurement*,7, 170-183.

Dodeen, H., &Darabi, M. (2009). Person-fit: Relationship with four personality tests in mathematics. *Research Papers in Education*, 24, 115–126.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985).Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67-86.

Drasgow, F., Levine, M.V. & McLaughlin, M. E. (1987).Detecting inappropriate test scores with optimal andpractical appropriateness indices. *Applied Psychological Measurement*, 11, 59-79.

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

Emons, W. H. M. (2009). Detection and diagnosis of person misfit from patterns of summed polytomous item. *Applied Psychological Measurement*, 33, 599-619.

Emons, W. H. M., Glas, C. A. W., Meijer, R. R., &Sijtsma, K. (2003). Person fit in order restricted latent class models. *Applied Psychological Measurement*, 27, 459-478.

Emons, W. H. M., Meijer, R. R., &Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the U3 person fit statistic. *Applied Psychological Measurement*,26, 88-108.

Emons, W. H. M., Sijtsma, K., Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods*, 10, 101-119.

Ferrando, P. J. (2012). Assessing inconsistent responding in E and N measures: An application of person-fit analysis in personality. *Personality and Individual Differences*, 52(6), 718-722.

Finkelman, M., & Kim, W. (2007).*Using Person Fit in a Body of Work Standard Setting*. Paper presented at the American Educational Research Association, Chicago, IL,USA.

Gerald, B. & Lawrence, M., R. (1992). Person-fit statistics: high potential and many unanswered questions. *Practical Assessment, Research & Evaluation*, 3(7). Retrieved June 4, 2014 from http://PAREonline.net/getvn.asp?v=3&n=7 . This paper has been viewed 21,894 times since 11/13/1999.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principals and Applications*. Boston: Kluwer- Nijhof

Han, T., Kolen, M., &Pohlmann, J. (1997). A comparison among IRT true- and observed-score equating and traditional equpercentileequating. *Applied Measurement in Education*, 10(2), 105-121.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns.Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18(3), 133-146.

Harwell, M., Stone, C. A., Hsu, T. C., &Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement*, 20(2), 101-125.

Hendrawan, I., Glas, C. A., & Meijer, R. R. (2005).The effect of person misfit on classification decisions. *Applied psychological measurement*, 29(1), 26-44.

Kane, M. T., & Brennan, R. L. (1980).Agreement coefficients as indices of dependability for domainreferencedtests.*Applied Psychological Measurement*, 4, 105-126.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics,*Applied Measurement in Education*, 16, 277-298.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. British Journal of Mathematical and Statistical Psychology, 35(1), 42-56.

Levine, M.V. & Rubin, D.B. (1979).Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, 4, 269-290.

Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). The bookmark standard setting procedure. Monterey, CA: McGraw-Hill

Li, M. F. &Olejnik, S. (1997).The power of Rasch person-fit statistics in detecting unusual response patterns.*Applied Psychological Measurement*, 21, 215-231.

Linacre, J. M., & Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, 8, 360-361.

Liu, M.-T., & Yu, P.-T.(2011). Aberrant Learning Achievement Detection Based on Person-fit Statistics in Personalized e-Learning Systems. *Educational Technology & Society*, 14 (1), 107–120.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, 45, 507-530.

Magis, D., Raiche, G., &Beland, S. (2012). A didactic presentation of snijders'slz* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, 37, 1, 57-81.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina, and Whitney Study. *Applied Psychological Measurement*, 21, 99-113.

Meijer, R. R., &Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, 8, 261–272.

Meijer, R. R., &Sijtsma, K. (2001). Methodology review: Evaluating person fit. Applied Psychological Measurement, 25, 107-135.

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research* (Vol. 1). Walter de Gruyter.

Molenaar, I.W.,&Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors.*Applied Measurement in Education*, 9, 27–45.

Muraki, E. & Bock, R.D. (2003). PARSCALE: IRT item analysis and test scoring for rating scale data. (Version 4.1) [Computer Software]. Chicago, IL: Scientific Software International.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, 19, 121–129.

Nering, M. L. (1997). The distribution of indexes of person-fit within the computerized adaptive testing environment. *Applied Psychological Measurement*,21, 115–127.

Nering, M. L. (1998). *The influence of Nonmodel-Fitting Examinees in Estimating Person Parameters*. Paper presented at the Annual AERA conference. San Diego, CA.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement*,22(1), 53-69.

Petridou, A. and Williams, J. (2007). Accounting for Aberrant Test Response Patterns Using Multilevel Models,*Journal of Educational Measurement*,Fall 2007, Vol. 44, No. 3, pp. 227–247

Petridou, A. and Williams, J. (2010). Accounting for unexpected test responses through examinees' and their teachers' explanations. *Assessment in Education:Principles, Policy & Practice*, 17:4, 357-382

Phillips, S. E. (1986). The effects of the deletion of misfitting persons on vertical equating via the Rasch model. *Journal of Educational Measurement*,23(2), 107-118.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Reise, S. P., &Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods*, 4(1), 3.

Reise, S.(1995). Scoring method and the detection of person misfit in a personality assessment context.*Applied Psychological Measurement*, 19,213–229.

Reise, S.P. & Flannery, W.P. (1996).Assessing person-fit on measures of typical performance.*Applied Measurement in Education*, 9, 9-26.

Rudner, L. M., Bracey, G., & Skaggs, G. (1996).The use of a person-fit statistic with one high-quality achievement test. *Applied Measurement in Education*, 9(1), 91-109.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in Item Response Theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, Volume 55.

Sato, T. (1975).*The construction and interpretation of S-P tables*. Tokyo: Meiji Tosho (in Japanese).

Schmitt, N. S., Cortina, J. M., & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement*, 17, 143-150.

Seo, D. G., & Weiss, D. J. (2013). lz Person-Fit Index to Identify Misfit Students With Achievement Test Data. *Educational and Psychological Measurement*, 73(6), 994-1016.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, 7, 131–145.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's non-IRT-based IRT model. *Applied Psychological Measurement*, 16, 149–157.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person fit research. *Psychometrika*, 66, 191-208.

Sinharay, S. (2015). Asymptotically Correct Standardization of Person-Fit Statistics Beyond Dichotomous Items. Psychometrika, 1-22.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement*, 45, 433-444

Smith, R. M. (1986). Person fit in the Rasch model. *Education and Psychological Measurement*, 46, 359-372.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51,541-565.

Snijders, T. (2001).Asymptotic distribution of person-fit statistics with estimated person parameter.*Psychometrika*, 66, 331-342.

Sotaridona, L. S., Choi, S. W., & Meijer, R. R. (2005).*The Effect of Misfitting Response Vectors on Item Calibration and Performance Classification*. Retrieved May 2013, from CTB/McGraw-Hill: http://www.ctb.com/img/pdfs/raMisfittingResponseVectors.pdf

Stocking, M. L., & Lord, F. M. (1983).Developing a common metric in item response theory.*Applied Psychological Measurement*, 7, 201-210.

St-Onge, C., Valois, P., Abdous, B., &Germain, S. (2009). A Monte Carlo study of the effect of item characteristic curve estimation on the accuracy of three person fit statistics.*Applied Psychological Measurement*, 33, 307-324.

St-Onge, C., Valois, P., Abdous, B., &Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement*, 35, 419-432.

Stricker, L. J., &Emmerich, W. (1999). Possible determinants of differential item functioning: familiarity, interest, and emotional reaction. *Journal of Educational Measurement*, 36, 347-366.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, 49, 95-110.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement*, 7, 81-96.

Tatsuoka, K. K., &Tatsuoka, M. M. (1982).Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics*, 7, 215-231.

Tatsuoka, K. K., &Tatsuoka, M. M. (1983).Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, 20, 221-230.

Tendeiro, J. N., & Meijer, R. R. (2014). Detection of Invalid Test Scores: The Usefulness of Simple NonIRT-based Statistics. *Journal of Educational Measurement*, 51(3), 239-259.

Tendeiro, J. N., Meijer, R. R., Schakel, L., &Maij-de Meij, A. M. (2013).Using cumulative sum statistics to detect inconsistencies in unproctored Internet testing. *Educational and Psychological Measurement*, 73(1), 143-161.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item responsetheory models. In D. J. Weiss (Ed.), *New horizons in testing*. New York: Academic Press.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Crosscultural Psychology*, 13, 267-298.

van, K.-S. E. M. L. A., & Meijer, R. (1999).The Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests. *Applied Psychological Measurement*,23, 4, 327-45.

vanKrimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomousitems.*Applied Psychological Measurement*, 26,164-180.

Woods, C. M., Oltmanns, T. F., &Turkheimer, E. (2008). Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological assessment*, 20(2), 159.

Wright, B. D., & Masters, G. N. (1982).*Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Rasch measurement. Chicago: Mesa Press.

Zhang, B., & Walker, C. M. (2008).Impact of missing data on person-model fit and person trait estimation.*Applied Psychological Measurement*, 32, 466-479.

Zickar , M. J. , &Robie, C. (1999). Modeling faking good on personality items: An item level analysis. *Journal of Applied Psychology*, 84, 551–563.

Appendix

Table A**39**
Bias of estimated item parameters for manipulation of the 25% of items and 20% of sample (20 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4745	4749	4643	4640
Item difficulty	Item1	0.20	0.00	0.00	0.00	0.01	0.02	0.02	0.01	0.01
	Item2	0.18	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
	Item3	-1.51	-0.01	-0.02	-0.14	-0.14	-0.12	-0.12	-0.34	-0.35
	Item4	-0.54	0.00	0.01	-0.02	-0.02	-0.02	-0.02	-0.06	-0.06
	Item5	1.08	0.02	0.02	0.10	0.10	0.10	0.10	0.14	0.14
	Item6	0.54	0.00	0.00	0.03	0.03	0.05	0.05	0.06	0.06
	Item7	-0.44	0.00	0.00	-0.02	-0.02	-0.02	-0.02	-0.06	-0.06
	Item8	-1.72	-0.02	-0.02	-0.21	-0.21	-0.16	-0.16	-0.47	-0.47
	Item9	0.13	0.00	0.00	0.01	0.01	0.02	0.02	0.01	0.01
	Item10	1.92	0.02	0.03	0.81	0.81	0.37	0.38	0.61	0.62
	Item11	-0.62	-0.02	-0.01	-0.05	-0.04	-0.05	-0.04	-0.09	-0.08
	Item12	-0.44	0.00	0.00	-0.01	-0.01	-0.02	-0.02	-0.07	-0.07
	Item13	-0.79	-0.01	-0.01	-0.02	-0.02	-0.04	-0.04	-0.12	-0.12
	Item14	2.55	0.04	0.01	0.46	0.40	0.27	0.23	0.36	0.32
	Item15	-1.34	-0.01	-0.01	-0.07	-0.08	-0.09	-0.09	-0.26	-0.26
	Item16	-1.11	-0.01	-0.01	-0.05	-0.05	-0.07	-0.07	-0.20	-0.20
	Item17	-1.66	-0.02	-0.02	-0.17	-0.17	-0.14	-0.14	-0.41	-0.42
	Item18	-0.95	0.00	0.00	-0.05	-0.04	-0.06	-0.05	-0.13	-0.12
	Item19	-1.24	-0.01	-0.01	-0.07	-0.07	-0.08	-0.08	-0.22	-0.22
	Item20	2.16	0.02	0.02	0.64	0.63	0.29	0.29	0.54	0.54
MAD	NA	NA	0.01	0.01	0.15	0.14	0.10	0.10	0.21	0.21
Item discrimination	Item1	2.15	-0.01	-0.02	-0.45	-0.46	-0.35	-0.36	-0.42	-0.43
	Item2	1.04	0.00	0.00	-0.13	-0.13	-0.10	-0.10	-0.12	-0.12
	Item3	2.13	-0.02	-0.03	-0.64	-0.65	-0.42	-0.43	-0.65	-0.66
	Item4	0.71	0.01	0.01	-0.07	-0.07	-0.06	-0.05	-0.07	-0.07
	Item5	1.07	-0.01	-0.01	-0.20	-0.20	-0.13	-0.13	-0.14	-0.14
	Item6	2.16	-0.01	-0.01	-0.54	-0.54	-0.35	-0.36	-0.44	-0.44
	Item7	1.04	0.00	0.00	-0.12	-0.12	-0.10	-0.10	-0.14	-0.14
	Item8	1.97	-0.02	-0.04	-0.63	-0.64	-0.40	-0.41	-0.65	-0.66
	Item9	1.97	-0.01	-0.01	-0.37	-0.38	-0.29	-0.30	-0.36	-0.36
	Item10	2.44	-0.03	-0.06	-1.41	-1.41	-0.88	-0.90	-1.07	-1.08
	Item11	0.57	-0.01	0.00	-0.07	-0.06	-0.05	-0.05	-0.06	-0.06
	Item12	1.93	0.00	-0.01	-0.30	-0.31	-0.25	-0.26	-0.34	-0.34
	Item13	2.18	-0.02	-0.03	-0.43	-0.43	-0.33	-0.34	-0.46	-0.47
	Item14	0.63	-0.01	0.00	-0.15	-0.14	-0.08	-0.07	-0.07	-0.06
	Item15	1.45	0.00	-0.01	-0.27	-0.27	-0.18	-0.19	-0.31	-0.31
	Item16	1.14	-0.01	-0.01	-0.16	-0.16	-0.12	-0.12	-0.20	-0.20
	Item17	1.81	-0.01	-0.02	-0.51	-0.51	-0.32	-0.33	-0.53	-0.54
	Item18	0.62	0.00	0.00	-0.07	-0.06	-0.06	-0.05	-0.07	-0.07
	Item19	2.09	-0.01	-0.02	-0.51	-0.52	-0.34	-0.35	-0.52	-0.52
	Item20	1.41	-0.01	-0.01	-0.56	-0.56	-0.28	-0.29	-0.38	-0.38
MAD	NA	NA	0.01	0.02	0.38	0.38	0.26	0.26	0.35	0.35

Table A**39**
Bias of estimated item parameters for manipulation of the 25% of items and 20% of sample (20 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4745	4749	4643	4640

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**40**
Bias of estimated item parameters for manipulation of the 25% of items and 30% of sample (20 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4698	4698	4591	4589
Item difficulty	Item1	2.40	0.00	0.02	1.14	1.13	0.76	0.77	0.98	0.99
	Item2	1.48	0.02	0.02	0.14	0.14	0.16	0.15	0.20	0.19
	Item3	-1.90	-0.02	-0.02	-1.14	-1.12	-0.56	-0.57	-0.59	-0.60
	Item4	-0.57	0.00	0.00	-0.06	-0.06	-0.07	-0.07	-0.09	-0.09
	Item5	2.21	0.01	0.02	0.41	0.41	0.30	0.30	0.40	0.40
	Item6	0.80	0.01	0.01	0.07	0.06	0.07	0.07	0.09	0.09
	Item7	-1.87	-0.02	-0.01	-0.51	-0.49	-0.31	-0.31	-0.30	-0.30
	Item8	0.72	0.01	0.01	0.04	0.04	0.05	0.06	0.08	0.08
	Item9	-2.18	-0.02	-0.04	-1.92	-1.86	-0.88	-0.88	-0.94	-0.95
	Item10	0.48	0.00	0.00	0.03	0.03	0.04	0.04	0.06	0.06
	Item11	-0.45	0.00	0.00	-0.05	-0.05	-0.06	-0.06	-0.08	-0.08
	Item12	-2.64	-0.04	-0.05	-2.68	-2.47	-1.15	-1.12	-1.27	-1.26
	Item13	-0.26	-0.01	-0.01	-0.04	-0.04	-0.05	-0.05	-0.07	-0.07
	Item14	-0.18	0.00	0.00	-0.02	-0.02	-0.03	-0.03	-0.04	-0.04
	Item15	2.59	0.02	0.03	0.68	0.67	0.48	0.48	0.62	0.61
	Item16	1.24	0.01	0.01	0.09	0.09	0.10	0.10	0.14	0.15
	Item17	0.13	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01
	Item18	2.27	0.02	0.03	0.53	0.53	0.39	0.39	0.50	0.50
	Item19	1.39	0.02	0.02	0.12	0.12	0.13	0.13	0.18	0.18
	Item20	1.04	0.01	0.01	0.07	0.07	0.09	0.09	0.12	0.12
MAD	NA	NA	0.01	0.02	0.49	0.47	0.28	0.28	0.34	0.34
Item discrimination	Item1	2.20	0.02	-0.02	-1.37	-1.37	-1.12	-1.12	-1.19	-1.20
	Item2	0.95	-0.01	-0.01	-0.20	-0.20	-0.16	-0.16	-0.17	-0.16
	Item3	2.33	-0.01	-0.04	-1.51	-1.51	-1.09	-1.09	-1.05	-1.06
	Item4	1.05	0.00	0.00	-0.24	-0.24	-0.18	-0.18	-0.17	-0.17
	Item5	1.39	0.00	-0.01	-0.51	-0.51	-0.37	-0.37	-0.39	-0.39
	Item6	0.91	-0.01	-0.01	-0.17	-0.17	-0.14	-0.14	-0.14	-0.14
	Item7	1.04	0.00	0.00	-0.37	-0.37	-0.23	-0.23	-0.19	-0.19
	Item8	1.59	-0.01	-0.02	-0.36	-0.36	-0.31	-0.32	-0.32	-0.33
	Item9	2.30	-0.02	-0.06	-1.65	-1.64	-1.23	-1.24	-1.21	-1.22
	Item10	2.23	-0.02	-0.03	-0.65	-0.66	-0.59	-0.59	-0.61	-0.61
	Item11	1.14	0.00	0.00	-0.27	-0.26	-0.21	-0.21	-0.20	-0.20
	Item12	1.56	0.00	-0.02	-1.07	-1.05	-0.75	-0.74	-0.74	-0.74
	Item13	0.70	0.00	0.00	-0.13	-0.13	-0.10	-0.10	-0.09	-0.08
	Item14	0.96	0.00	0.01	-0.18	-0.18	-0.15	-0.14	-0.14	-0.14
	Item15	1.28	-0.01	-0.02	-0.53	-0.53	-0.39	-0.39	-0.42	-0.42
	Item16	1.48	0.00	-0.01	-0.35	-0.35	-0.28	-0.29	-0.29	-0.29
	Item17	2.47	-0.01	-0.02	-0.86	-0.87	-0.78	-0.79	-0.80	-0.81

Table A**40**

Bias of estimated item parameters for manipulation of the 25% of items and 30% of sample (20 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4698	4698	4591	4589
	Item18	1.52	-0.01	-0.02	-0.64	-0.64	-0.48	-0.48	-0.50	-0.51
	Item19	1.50	-0.01	-0.02	-0.39	-0.39	-0.31	-0.31	-0.32	-0.32
	Item20	1.34	-0.01	-0.01	-0.29	-0.29	-0.24	-0.24	-0.25	-0.25
MAD	NA	NA	0.01	0.02	0.59	0.58	0.46	0.46	0.46	0.46

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**41**

Bias of estimated item parameters for manipulation of the 50% of items and 10% of sample (20 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4778	4778	4658	4658
Item difficulty	Item1	-1.36	-0.01	-0.01	0.05	0.05	0.03	0.02	-0.09	-0.09
	Item2	-1.19	-0.01	-0.02	0.03	0.02	0.00	0.00	-0.09	-0.09
	Item3	1.53	0.00	0.00	0.11	0.11	-0.02	-0.02	0.11	0.12
	Item4	-1.27	0.01	0.01	0.12	0.12	0.08	0.09	0.03	0.04
	Item5	1.36	0.01	0.01	0.10	0.10	0.00	0.00	0.10	0.10
	Item6	1.49	0.01	0.01	0.01	0.01	-0.03	-0.04	0.04	0.04
	Item7	-2.23	-0.02	-0.04	0.18	0.17	0.12	0.11	-0.20	-0.22
	Item8	-0.86	-0.01	-0.01	0.00	0.00	-0.01	-0.01	-0.07	-0.07
	Item9	1.99	0.01	0.01	0.09	0.08	-0.08	-0.08	0.14	0.14
	Item10	0.09	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.02
	Item11	-0.07	0.01	0.01	0.01	0.01	0.02	0.02	0.04	0.04
	Item12	-0.40	0.00	0.00	-0.02	-0.02	-0.01	-0.01	-0.03	-0.03
	Item13	-1.02	-0.01	-0.01	0.03	0.03	0.01	0.01	-0.05	-0.05
	Item14	-1.77	-0.03	-0.02	0.18	0.19	0.12	0.12	0.00	0.00
	Item15	-1.68	-0.01	-0.02	0.11	0.10	0.07	0.06	-0.10	-0.11
	Item16	-1.13	-0.01	-0.01	0.04	0.04	0.02	0.02	-0.05	-0.05
	Item17	-1.44	-0.01	-0.02	0.06	0.06	0.03	0.03	-0.10	-0.10
	Item18	-1.18	0.00	-0.01	0.04	0.03	0.02	0.01	-0.08	-0.08
	Item19	-1.08	-0.01	-0.01	0.01	0.01	0.00	0.00	-0.09	-0.09
	Item20	1.52	0.00	0.00	0.08	0.08	-0.02	-0.02	0.09	0.09
MAD	NA	NA	0.01	0.01	0.06	0.06	0.04	0.04	0.08	0.08
Item discrimination	Item1	2.41	-0.02	-0.04	-0.20	-0.22	-0.03	-0.06	-0.26	-0.28
	Item2	1.80	-0.01	-0.02	-0.11	-0.12	-0.04	-0.05	-0.16	-0.17
	Item3	1.45	0.00	0.00	-0.30	-0.30	0.02	0.02	-0.11	-0.11
	Item4	0.67	0.00	0.01	0.04	0.04	0.04	0.04	0.02	0.02
	Item5	1.44	0.00	-0.01	-0.28	-0.28	0.00	0.00	-0.09	-0.10
	Item6	0.81	0.00	0.00	-0.06	-0.06	0.02	0.02	0.02	0.02
	Item7	2.44	-0.02	-0.07	-0.36	-0.38	-0.10	-0.14	-0.54	-0.57
	Item8	1.49	-0.01	-0.02	-0.10	-0.11	-0.06	-0.06	-0.13	-0.14
	Item9	1.08	0.01	0.01	-0.15	-0.15	0.06	0.06	-0.05	-0.05
	Item10	0.70	-0.01	-0.01	-0.01	-0.01	0.01	0.01	0.02	0.03
	Item11	0.53	0.00	0.00	0.02	0.03	0.03	0.04	0.05	0.06
	Item12	1.65	0.00	-0.01	-0.21	-0.21	-0.12	-0.13	-0.17	-0.18
	Item13	1.16	-0.01	-0.01	-0.03	-0.04	-0.01	-0.01	-0.06	-0.07

Table A**41**

Bias of estimated item parameters for manipulation of the 50% of items and 10% of sample (20 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4778	4778	4658	4658
	Item14	0.70	-0.01	0.00	0.05	0.05	0.04	0.04	0.01	0.01
	Item15	2.28	-0.03	-0.05	-0.13	-0.15	0.03	0.00	-0.25	-0.27
	Item16	1.12	-0.01	-0.01	-0.02	-0.02	0.00	0.00	-0.05	-0.06
	Item17	2.26	-0.02	-0.04	-0.15	-0.16	0.00	-0.02	-0.23	-0.25
	Item18	1.71	0.01	0.00	-0.08	-0.09	-0.02	-0.03	-0.13	-0.14
	Item19	2.44	-0.02	-0.03	-0.27	-0.28	-0.12	-0.13	-0.30	-0.32
	Item20	1.20	0.00	0.00	-0.18	-0.18	0.02	0.02	-0.05	-0.05
MAD	NA	NA	0.01	0.02	0.14	0.14	0.04	0.04	0.14	0.14

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**42**

Bias of estimated item parameters for manipulation of the 50% of items and 20% of sample (20 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4656	4655	4530	4531
Item difficulty	Item1	-2.03	-0.02	-0.02	0.35	0.35	0.28	0.28	0.18	0.18
	Item2	-2.31	-0.05	-0.04	0.50	0.50	0.41	0.41	0.30	0.31
	Item3	-0.20	-0.01	-0.01	-0.02	-0.02	-0.01	-0.01	-0.01	-0.01
	Item4	-1.49	-0.02	-0.02	0.09	0.09	0.08	0.08	0.01	0.00
	Item5	1.57	0.02	0.02	0.02	0.02	-0.07	-0.06	0.04	0.04
	Item6	-2.57	-0.03	-0.03	0.62	0.62	0.53	0.53	0.36	0.35
	Item7	-0.11	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.02
	Item8	0.40	0.00	-0.01	-0.08	-0.08	-0.06	-0.06	-0.04	-0.05
	Item9	2.05	0.02	0.03	0.01	0.02	-0.18	-0.18	0.02	0.04
	Item10	-1.08	-0.01	-0.01	0.08	0.08	0.05	0.05	0.03	0.03
	Item11	-1.25	-0.01	-0.01	-0.01	-0.02	0.00	0.00	-0.06	-0.06
	Item12	-1.12	-0.01	0.00	0.20	0.20	0.15	0.16	0.14	0.15
	Item13	2.46	0.03	0.01	-0.53	-0.54	-0.50	-0.51	-0.38	-0.39
	Item14	-2.39	-0.02	-0.04	0.45	0.44	0.42	0.41	0.22	0.21
	Item15	-2.10	-0.02	-0.03	0.31	0.30	0.29	0.28	0.13	0.12
	Item16	0.67	0.00	0.00	0.02	0.02	0.02	0.02	0.05	0.05
	Item17	1.95	0.02	0.02	-0.03	-0.03	-0.17	-0.17	0.00	0.01
	Item18	-0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
	Item19	0.67	0.01	0.01	0.03	0.03	0.03	0.03	0.05	0.05
	Item20	1.46	0.01	0.01	0.01	0.01	-0.06	-0.06	0.02	0.02
MAD	NA	NA	0.02	0.02	0.17	0.17	0.17	0.16	0.10	0.10
Item discrimination	Item1	0.99	-0.01	-0.01	0.07	0.07	0.13	0.12	0.09	0.08
	Item2	0.81	-0.01	-0.01	0.12	0.12	0.14	0.14	0.12	0.12
	Item3	1.33	-0.01	-0.01	-0.28	-0.28	-0.19	-0.19	-0.17	-0.17
	Item4	1.57	-0.02	-0.02	-0.18	-0.19	-0.01	-0.01	-0.07	-0.08
	Item5	1.56	-0.02	-0.02	-0.36	-0.37	-0.02	-0.03	-0.11	-0.12
	Item6	0.95	-0.01	-0.01	0.15	0.15	0.21	0.20	0.12	0.12
	Item7	0.84	0.00	0.00	-0.04	-0.04	0.00	0.00	0.02	0.02
	Item8	0.56	0.00	0.00	0.06	0.06	0.07	0.08	0.09	0.09
	Item9	2.40	-0.02	-0.06	-0.94	-0.95	-0.16	-0.20	-0.49	-0.52

Table A**42**
Bias of estimated item parameters for manipulation of the 50% of items and 20% of sample (20 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4656	4655	4530	4531
	Item10	1.00	-0.01	-0.01	-0.04	-0.04	0.00	0.00	0.01	0.01
	Item11	2.50	-0.01	-0.03	-0.76	-0.77	-0.33	-0.35	-0.45	-0.47
	Item12	0.63	-0.01	0.00	0.07	0.07	0.07	0.07	0.08	0.09
	Item13	0.60	0.00	0.00	0.08	0.09	0.13	0.14	0.12	0.12
	Item14	1.94	0.00	-0.03	-0.19	-0.20	0.19	0.16	-0.07	-0.10
	Item15	2.02	-0.02	-0.04	-0.23	-0.25	0.16	0.13	-0.08	-0.10
	Item16	1.22	0.00	0.00	-0.22	-0.22	-0.10	-0.10	-0.09	-0.09
	Item17	1.72	-0.01	-0.03	-0.44	-0.44	0.04	0.03	-0.15	-0.16
	Item18	1.38	-0.02	-0.02	-0.32	-0.32	-0.22	-0.22	-0.20	-0.20
	Item19	1.22	-0.02	-0.02	-0.23	-0.23	-0.12	-0.12	-0.11	-0.11
	Item20	1.45	0.00	-0.01	-0.30	-0.30	-0.02	-0.02	-0.07	-0.08
MAD	NA	NA	0.01	0.02	0.25	0.26	0.12	0.12	0.14	0.14

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**43**
Bias of estimated item parameters for manipulation of the 50% of items and 30% of sample (20 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4602	4603	4491	4490
Item difficulty	Item1	1.73	0.01	0.01	-0.18	-0.17	-0.15	-0.15	-0.04	-0.04
	Item2	0.51	0.00	0.00	-0.02	-0.02	0.00	0.00	0.01	0.01
	Item3	1.31	0.01	0.01	-0.11	-0.11	-0.08	-0.07	-0.01	0.00
	Item4	0.99	0.00	0.00	-0.01	-0.01	0.02	0.02	0.06	0.06
	Item5	1.63	0.01	0.02	-0.07	-0.07	-0.08	-0.08	0.03	0.04
	Item6	-0.11	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.02	-0.02
	Item7	-1.04	-0.01	-0.01	-0.04	-0.04	-0.03	-0.03	-0.10	-0.10
	Item8	-1.57	0.00	-0.01	-0.02	-0.02	0.05	0.05	-0.10	-0.10
	Item9	-1.67	-0.02	-0.03	-0.09	-0.09	0.03	0.03	-0.15	-0.16
	Item10	-1.62	-0.02	-0.01	0.29	0.30	0.26	0.26	0.11	0.12
	Item11	0.62	0.00	0.00	-0.01	-0.01	0.02	0.02	0.04	0.04
	Item12	0.24	0.00	0.00	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02
	Item13	-1.47	-0.01	-0.02	-0.10	-0.10	0.00	0.00	-0.14	-0.14
	Item14	-0.32	0.00	0.00	-0.01	-0.01	-0.03	-0.03	-0.05	-0.05
	Item15	0.03	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.01
	Item16	0.19	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.02
	Item17	1.80	0.01	0.01	-0.15	-0.15	-0.16	-0.15	-0.02	-0.02
	Item18	2.08	0.03	0.03	-0.41	-0.41	-0.35	-0.35	-0.21	-0.22
	Item19	0.64	0.00	0.01	0.00	0.00	0.03	0.03	0.05	0.05
	Item20	-1.17	-0.01	-0.01	0.05	0.05	0.04	0.04	-0.06	-0.06
MAD	NA	NA	0.01	0.01	0.08	0.08	0.07	0.07	0.06	0.06
Item discriminatio n	Item1	1.37	0.00	-0.01	-0.23	-0.23	-0.07	-0.08	-0.15	-0.15
	Item2	1.30	0.00	-0.01	-0.20	-0.21	-0.14	-0.14	-0.16	-0.16
	Item3	1.40	0.00	-0.01	-0.22	-0.23	-0.10	-0.10	-0.14	-0.15
	Item4	2.24	0.01	0.00	-0.72	-0.73	-0.47	-0.48	-0.54	-0.54
	Item5	2.19	-0.01	-0.03	-0.77	-0.78	-0.40	-0.42	-0.55	-0.56

Table A**43**
Bias of estimated item parameters for manipulation of the 50% of items and 30% of sample (20 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4602	4603	4491	4490
	Item6	2.17	-0.01	-0.02	-0.78	-0.78	-0.65	-0.65	-0.66	-0.67
	Item7	2.05	-0.01	-0.02	-0.76	-0.76	-0.42	-0.43	-0.50	-0.50
	Item8	1.65	0.00	-0.01	-0.54	-0.55	-0.22	-0.23	-0.33	-0.33
	Item9	2.04	-0.03	-0.04	-0.88	-0.89	-0.44	-0.45	-0.60	-0.60
	Item10	0.64	0.00	0.00	0.02	0.02	0.06	0.06	0.03	0.03
	Item11	1.74	-0.01	-0.01	-0.42	-0.43	-0.32	-0.32	-0.34	-0.34
	Item12	0.92	-0.02	-0.02	-0.06	-0.06	-0.04	-0.03	-0.05	-0.05
	Item13	2.37	-0.02	-0.04	-1.11	-1.11	-0.59	-0.60	-0.75	-0.76
	Item14	1.96	0.00	0.00	-0.63	-0.63	-0.49	-0.49	-0.51	-0.52
	Item15	1.62	-0.01	-0.02	-0.40	-0.40	-0.31	-0.32	-0.33	-0.33
	Item16	2.41	0.00	-0.01	-0.92	-0.93	-0.79	-0.80	-0.81	-0.81
	Item17	1.70	0.01	0.00	-0.43	-0.44	-0.18	-0.18	-0.29	-0.29
	Item18	0.83	-0.01	-0.01	0.01	0.01	0.06	0.06	0.02	0.02
	Item19	1.98	-0.02	-0.02	-0.57	-0.57	-0.43	-0.44	-0.46	-0.47
	Item20	1.20	-0.01	-0.01	-0.23	-0.22	-0.09	-0.10	-0.13	-0.13
MAD	NA	NA	0.01	0.01	0.50	0.50	0.32	0.32	0.37	0.37

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**44**
Bias of estimated item parameters for manipulation of the 25% of items and 10% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4657	4658	4589	4591
Item difficulty	Item1	-2.07	-0.02	-0.02	-0.18	-0.17	-0.06	-0.06	-0.04	-0.04
	Item2	2.52	0.04	0.06	0.62	0.64	0.17	0.20	0.37	0.40
	Item3	1.80	0.02	0.02	0.10	0.10	0.05	0.05	0.13	0.14
	Item4	2.07	0.02	0.04	0.22	0.23	0.08	0.09	0.18	0.19
	Item5	-0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
	Item6	-0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
	Item7	-1.89	-0.02	-0.02	-0.28	-0.28	-0.07	-0.08	-0.08	-0.08
	Item8	0.30	0.00	0.00	0.01	0.01	0.01	0.01	0.03	0.03
	Item9	-0.26	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	0.00
	Item10	-0.68	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01
	Item11	-2.23	-0.02	0.00	-0.12	-0.09	-0.05	-0.03	0.02	0.04
	Item12	2.62	0.03	0.03	0.20	0.20	0.06	0.07	0.20	0.20
	Item13	2.63	0.03	0.06	0.50	0.51	0.14	0.16	0.33	0.36
	Item14	-1.75	-0.03	-0.02	-0.09	-0.08	-0.05	-0.04	-0.02	0.00
	Item15	-2.37	-0.03	-0.01	-0.15	-0.12	-0.06	-0.04	0.02	0.03
	Item16	-1.15	-0.01	-0.01	-0.06	-0.06	-0.04	-0.04	-0.03	-0.04
	Item17	-0.39	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	0.00
	Item18	-1.98	-0.04	-0.02	-0.11	-0.09	-0.06	-0.04	0.01	0.02
	Item19	0.95	0.01	0.01	0.00	0.00	0.02	0.02	0.06	0.06
	Item20	-2.46	-0.03	-0.04	-0.63	-0.63	-0.14	-0.16	-0.17	-0.18
	Item21	0.57	0.01	0.01	0.00	0.01	0.02	0.02	0.05	0.05
	Item22	1.47	0.03	0.03	0.05	0.05	0.04	0.05	0.10	0.11

Table A**44**
Bias of estimated item parameters for manipulation of the 25% of items and 10% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4657	4658	4589	4591
	Item23	-1.52	-0.02	-0.02	-0.10	-0.10	-0.05	-0.05	-0.04	-0.04
	Item24	1.82	0.03	0.04	0.08	0.09	0.05	0.05	0.13	0.13
	Item25	0.64	0.00	0.01	0.00	0.00	0.02	0.02	0.05	0.05
	Item26	2.11	0.02	0.03	0.21	0.22	0.07	0.08	0.18	0.19
	Item27	-0.13	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	0.00
	Item28	-2.27	-0.03	-0.03	-0.25	-0.24	-0.08	-0.08	-0.06	-0.06
	Item29	2.65	0.01	0.01	0.10	0.10	0.02	0.02	0.13	0.12
	Item30	0.88	0.01	0.02	0.01	0.01	0.03	0.03	0.06	0.07
	Item31	0.06	0.00	0.00	0.01	0.01	0.00	0.00	0.02	0.02
	Item32	1.42	0.01	0.02	0.05	0.05	0.03	0.04	0.09	0.09
	Item33	-1.83	-0.02	-0.02	-0.11	-0.10	-0.06	-0.05	-0.02	-0.02
	Item34	0.92	0.01	0.01	0.00	0.01	0.02	0.02	0.06	0.06
	Item35	-0.58	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02
	Item36	-2.38	-0.04	-0.04	-0.56	-0.57	-0.14	-0.15	-0.16	-0.17
	Item37	-1.37	-0.02	-0.02	-0.11	-0.12	-0.05	-0.05	-0.04	-0.05
	Item38	0.96	0.01	0.01	0.01	0.02	0.02	0.02	0.06	0.06
	Item39	-2.36	-0.04	-0.02	-0.17	-0.14	-0.07	-0.05	0.00	0.02
	Item40	0.99	0.01	0.01	0.01	0.01	0.02	0.02	0.06	0.06
MAD	NA	NA	0.02	0.02	0.13	0.13	0.05	0.05	0.08	0.08
Item discrimination	Item1	1.06	-0.01	-0.01	-0.15	-0.15	-0.03	-0.03	-0.01	-0.01
	Item2	2.38	-0.04	-0.10	-1.07	-1.09	-0.38	-0.43	-0.58	-0.62
	Item3	1.83	-0.02	-0.03	-0.32	-0.32	-0.07	-0.08	-0.14	-0.15
	Item4	2.22	-0.01	-0.04	-0.64	-0.65	-0.16	-0.19	-0.28	-0.30
	Item5	0.79	-0.01	-0.01	-0.04	-0.04	-0.02	-0.02	-0.01	-0.01
	Item6	0.94	-0.01	-0.01	-0.05	-0.05	-0.02	-0.02	-0.02	-0.02
	Item7	2.23	0.00	-0.02	-0.72	-0.73	-0.14	-0.16	-0.18	-0.20
	Item8	1.74	-0.01	-0.02	-0.12	-0.13	-0.06	-0.06	-0.08	-0.09
	Item9	1.11	-0.02	-0.02	-0.08	-0.08	-0.04	-0.04	-0.04	-0.04
	Item10	0.87	-0.01	-0.01	-0.06	-0.06	-0.03	-0.02	-0.01	-0.01
	Item11	0.60	0.00	0.00	-0.05	-0.04	-0.01	0.00	0.02	0.03
	Item12	1.17	0.00	-0.01	-0.18	-0.19	-0.03	-0.03	-0.08	-0.08
	Item13	1.86	-0.03	-0.06	-0.66	-0.67	-0.19	-0.22	-0.33	-0.36
	Item14	0.69	-0.01	0.00	-0.06	-0.05	-0.01	-0.01	0.01	0.02
	Item15	0.64	0.00	0.00	-0.06	-0.06	-0.01	0.00	0.02	0.03
	Item16	2.06	-0.01	-0.01	-0.37	-0.37	-0.07	-0.08	-0.09	-0.09
	Item17	2.00	-0.02	-0.02	-0.22	-0.23	-0.09	-0.09	-0.12	-0.12
	Item18	0.62	-0.01	0.00	-0.05	-0.04	-0.01	0.00	0.02	0.03
	Item19	2.42	-0.02	-0.03	-0.29	-0.30	-0.09	-0.10	-0.15	-0.16
	Item20	1.89	0.00	-0.03	-0.73	-0.74	-0.18	-0.20	-0.23	-0.25
	Item21	1.70	-0.02	-0.02	-0.12	-0.13	-0.06	-0.06	-0.09	-0.09
	Item22	1.52	-0.02	-0.03	-0.17	-0.17	-0.05	-0.06	-0.08	-0.09
	Item23	1.31	-0.02	-0.02	-0.18	-0.18	-0.04	-0.04	-0.03	-0.03
	Item24	1.40	-0.03	-0.03	-0.18	-0.18	-0.05	-0.05	-0.08	-0.09
	Item25	2.39	-0.01	-0.02	-0.23	-0.24	-0.09	-0.10	-0.15	-0.16
	Item26	2.03	-0.02	-0.04	-0.53	-0.55	-0.13	-0.15	-0.24	-0.26

Table A**44**

Bias of estimated item parameters for manipulation of the 25% of items and 10% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4657	4658	4589	4591
	Item27	1.49	-0.01	-0.01	-0.10	-0.11	-0.05	-0.05	-0.06	-0.06
	Item28	1.15	-0.01	-0.02	-0.21	-0.21	-0.04	-0.04	-0.02	-0.03
	Item29	0.89	0.00	0.00	-0.08	-0.08	0.00	0.00	-0.02	-0.02
	Item30	2.16	-0.03	-0.04	-0.23	-0.24	-0.08	-0.09	-0.13	-0.14
	Item31	1.42	-0.02	-0.02	-0.10	-0.10	-0.05	-0.05	-0.06	-0.06
	Item32	2.36	-0.02	-0.03	-0.41	-0.42	-0.09	-0.11	-0.16	-0.18
	Item33	0.81	0.00	0.00	-0.08	-0.08	-0.02	-0.01	0.01	0.01
	Item34	2.17	0.00	-0.01	-0.21	-0.22	-0.05	-0.06	-0.11	-0.12
	Item35	1.89	-0.01	-0.02	-0.22	-0.22	-0.07	-0.08	-0.09	-0.09
	Item36	1.89	-0.02	-0.04	-0.71	-0.72	-0.18	-0.19	-0.22	-0.24
	Item37	2.35	-0.02	-0.03	-0.58	-0.58	-0.12	-0.14	-0.15	-0.17
	Item38	1.31	-0.01	-0.01	-0.09	-0.09	-0.03	-0.03	-0.05	-0.05
	Item39	0.66	-0.01	0.00	-0.07	-0.06	-0.02	-0.01	0.02	0.02
	Item40	1.33	-0.02	-0.02	-0.09	-0.09	-0.03	-0.04	-0.05	-0.06
MAD	NA	NA	0.01	0.02	0.26	0.27	0.07	0.08	0.11	0.12

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**45**

Bias of estimated item parameters for manipulation of the 25% of items and 20% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4534	4538	4497	4498
Item difficulty	Item1	1.22	0.01	0.02	0.04	0.04	0.05	0.06	0.12	0.13
	Item2	-2.68	-0.02	-0.06	-2.30	-2.22	-0.56	-0.60	-0.57	-0.60
	Item3	-0.64	0.00	0.00	-0.03	-0.02	-0.03	-0.02	-0.01	0.00
	Item4	-2.65	-0.03	-0.02	-0.49	-0.46	-0.15	-0.14	-0.05	-0.04
	Item5	0.05	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.02
	Item6	-2.51	-0.01	0.00	-0.33	-0.30	-0.12	-0.10	0.00	0.01
	Item7	-2.09	-0.01	-0.01	-0.41	-0.40	-0.13	-0.13	-0.08	-0.09
	Item8	-2.61	-0.02	-0.03	-0.98	-0.96	-0.24	-0.25	-0.20	-0.21
	Item9	2.19	0.02	0.03	0.15	0.15	0.08	0.08	0.24	0.24
	Item10	-2.06	-0.01	-0.02	-0.63	-0.63	-0.18	-0.19	-0.15	-0.16
	Item11	-1.67	-0.01	0.00	-0.16	-0.15	-0.09	-0.08	-0.04	-0.04
	Item12	-1.86	-0.02	-0.02	-0.22	-0.20	-0.12	-0.11	-0.05	-0.04
	Item13	-1.11	-0.03	-0.02	-0.09	-0.08	-0.08	-0.07	-0.04	-0.03
	Item14	1.77	0.02	0.03	0.14	0.14	0.09	0.10	0.20	0.21
	Item15	1.35	0.01	0.02	0.05	0.05	0.05	0.05	0.13	0.14
	Item16	1.20	0.01	0.01	0.04	0.04	0.04	0.03	0.09	0.08
	Item17	-2.08	-0.02	-0.02	-0.51	-0.50	-0.16	-0.16	-0.12	-0.12
	Item18	1.36	0.01	0.01	0.04	0.05	0.05	0.05	0.13	0.13
	Item19	-2.19	-0.01	-0.03	-0.84	-0.84	-0.23	-0.24	-0.21	-0.22
	Item20	-1.59	-0.02	-0.01	-0.16	-0.15	-0.09	-0.09	-0.05	-0.04
	Item21	-0.24	0.00	0.00	-0.01	-0.01	-0.02	-0.02	0.00	0.00
	Item22	2.35	0.03	0.01	0.07	0.04	0.03	0.01	0.14	0.12

Table A**45**
Bias of estimated item parameters for manipulation of the 25% of items and 20% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4534	4538	4497	4498
	Item23	1.44	0.01	0.02	0.06	0.07	0.06	0.06	0.14	0.15
	Item24	0.92	0.01	0.01	0.02	0.02	0.04	0.04	0.10	0.10
	Item25	2.66	0.02	0.05	0.83	0.84	0.41	0.43	0.72	0.74
	Item26	1.28	0.00	-0.01	0.02	0.01	0.01	0.00	0.06	0.05
	Item27	-0.35	0.00	0.00	-0.02	-0.02	-0.03	-0.03	-0.02	-0.02
	Item28	2.52	0.03	0.04	0.24	0.24	0.11	0.12	0.30	0.31
	Item29	-2.67	-0.03	-0.03	-0.57	-0.55	-0.17	-0.16	-0.08	-0.07
	Item30	0.08	0.00	0.00	0.01	0.01	0.00	0.00	0.02	0.02
	Item31	-0.97	-0.01	0.00	-0.06	-0.05	-0.06	-0.04	-0.02	-0.01
	Item32	0.64	0.01	0.01	0.03	0.03	0.04	0.04	0.08	0.08
	Item33	-1.62	-0.01	-0.01	-0.30	-0.30	-0.12	-0.12	-0.09	-0.10
	Item34	1.88	0.03	0.03	0.10	0.11	0.07	0.07	0.19	0.19
	Item35	-1.33	-0.02	-0.01	-0.15	-0.15	-0.10	-0.10	-0.07	-0.07
	Item36	0.60	0.01	0.01	0.03	0.03	0.02	0.02	0.05	0.05
	Item37	0.86	0.01	0.01	0.02	0.02	0.04	0.04	0.09	0.09
	Item38	0.52	0.01	0.01	0.02	0.02	0.03	0.03	0.06	0.07
	Item39	1.68	0.02	0.02	0.11	0.11	0.08	0.08	0.18	0.19
	Item40	2.67	0.03	0.03	0.20	0.20	0.08	0.08	0.28	0.28
MAD	NA	NA	0.01	0.02	0.26	0.26	0.10	0.10	0.13	0.13
Item discrimination	Item1	2.12	-0.02	-0.03	-0.44	-0.45	-0.23	-0.24	-0.32	-0.33
	Item2	2.40	0.01	-0.08	-1.71	-1.70	-0.93	-0.96	-1.00	-1.03
	Item3	0.61	-0.01	0.00	-0.07	-0.06	-0.04	-0.03	-0.02	-0.01
	Item4	0.85	-0.01	0.00	-0.22	-0.21	-0.06	-0.06	-0.03	-0.03
	Item5	2.13	-0.02	-0.02	-0.50	-0.50	-0.36	-0.37	-0.41	-0.42
	Item6	0.70	0.00	0.01	-0.14	-0.13	-0.04	-0.03	0.00	0.00
	Item7	1.29	0.00	0.00	-0.41	-0.40	-0.13	-0.13	-0.11	-0.11
	Item8	1.51	-0.01	-0.02	-0.71	-0.70	-0.25	-0.26	-0.25	-0.26
	Item9	1.24	-0.01	-0.01	-0.25	-0.25	-0.11	-0.12	-0.18	-0.18
	Item10	1.98	0.01	-0.01	-0.93	-0.93	-0.33	-0.35	-0.35	-0.36
	Item11	0.81	0.00	0.00	-0.14	-0.14	-0.06	-0.05	-0.03	-0.03
	Item12	0.78	-0.01	0.00	-0.15	-0.14	-0.06	-0.05	-0.02	-0.02
	Item13	0.63	-0.01	-0.01	-0.09	-0.08	-0.04	-0.04	-0.02	-0.02
	Item14	2.02	-0.02	-0.04	-0.58	-0.59	-0.30	-0.32	-0.40	-0.41
	Item15	2.06	-0.01	-0.02	-0.45	-0.46	-0.22	-0.23	-0.31	-0.32
	Item16	0.78	-0.01	-0.01	-0.08	-0.08	-0.04	-0.04	-0.06	-0.06
	Item17	1.60	-0.01	-0.02	-0.61	-0.61	-0.21	-0.22	-0.20	-0.21
	Item18	1.91	-0.01	-0.02	-0.38	-0.39	-0.20	-0.20	-0.28	-0.29
	Item19	2.24	0.00	-0.04	-1.22	-1.22	-0.49	-0.52	-0.54	-0.56
	Item20	0.84	0.00	0.00	-0.15	-0.14	-0.06	-0.06	-0.03	-0.03
	Item21	0.93	0.00	0.00	-0.11	-0.11	-0.07	-0.06	-0.06	-0.06
	Item22	0.61	0.00	0.00	-0.06	-0.05	-0.02	-0.02	-0.04	-0.03
	Item23	2.29	-0.02	-0.03	-0.59	-0.60	-0.30	-0.32	-0.41	-0.42
	Item24	2.22	-0.02	-0.03	-0.44	-0.45	-0.26	-0.27	-0.35	-0.36
	Item25	2.06	-0.01	-0.05	-1.08	-1.08	-0.70	-0.72	-0.85	-0.87
	Item26	0.59	0.00	0.00	-0.05	-0.04	-0.02	-0.02	-0.02	-0.02

Table A**45**
Bias of estimated item parameters for manipulation of the 25% of items and 20% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4534	4538	4497	4498
	Item27	1.89	0.00	-0.01	-0.44	-0.44	-0.29	-0.29	-0.31	-0.32
	Item28	1.18	-0.01	-0.01	-0.27	-0.27	-0.12	-0.13	-0.19	-0.19
	Item29	0.96	-0.01	-0.01	-0.29	-0.28	-0.08	-0.08	-0.05	-0.05
	Item30	0.84	0.00	0.00	-0.09	-0.09	-0.06	-0.05	-0.05	-0.05
	Item31	0.53	-0.01	0.00	-0.07	-0.06	-0.03	-0.02	-0.01	0.00
	Item32	1.80	-0.01	-0.02	-0.30	-0.31	-0.20	-0.21	-0.26	-0.26
	Item33	1.92	-0.01	-0.02	-0.72	-0.72	-0.26	-0.27	-0.26	-0.27
	Item34	1.26	-0.01	-0.02	-0.22	-0.22	-0.11	-0.11	-0.16	-0.16
	Item35	1.10	-0.01	-0.01	-0.22	-0.22	-0.10	-0.10	-0.08	-0.08
	Item36	0.82	-0.01	-0.01	-0.09	-0.09	-0.06	-0.05	-0.06	-0.06
	Item37	1.86	-0.02	-0.02	-0.31	-0.32	-0.18	-0.19	-0.25	-0.26
	Item38	1.38	-0.01	-0.01	-0.19	-0.19	-0.13	-0.13	-0.16	-0.16
	Item39	2.06	-0.01	-0.02	-0.56	-0.57	-0.29	-0.30	-0.38	-0.40
	Item40	1.01	-0.01	-0.01	-0.20	-0.20	-0.08	-0.08	-0.14	-0.14
MAD	NA	NA	0.01	0.02	0.39	0.39	0.19	0.19	0.22	0.22

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**46**
Bias of estimated item parameters for manipulation of the 25% of items and 30% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4506	4508	4434	4433
Item difficulty	Item1	-1.09	0.00	-0.01	-0.09	-0.09	-0.10	-0.10	-0.12	-0.12
	Item2	0.25	0.01	0.01	0.02	0.02	0.04	0.04	0.04	0.04
	Item3	2.10	0.02	0.03	0.89	0.89	0.45	0.46	0.50	0.50
	Item4	-1.43	-0.01	-0.01	-0.19	-0.20	-0.14	-0.15	-0.18	-0.18
	Item5	-1.45	0.00	-0.01	-0.18	-0.18	-0.13	-0.14	-0.17	-0.17
	Item6	0.12	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.02
	Item7	-1.98	-0.02	-0.03	-0.34	-0.34	-0.21	-0.22	-0.26	-0.27
	Item8	1.11	0.03	0.02	0.12	0.10	0.10	0.08	0.10	0.09
	Item9	0.99	0.02	0.01	0.09	0.08	0.09	0.09	0.10	0.10
	Item10	1.57	0.02	0.03	0.35	0.35	0.21	0.22	0.24	0.24
	Item11	-2.46	-0.02	-0.02	-0.40	-0.38	-0.23	-0.22	-0.28	-0.27
	Item12	-1.19	-0.01	-0.01	-0.10	-0.10	-0.11	-0.11	-0.13	-0.13
	Item13	-0.06	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	Item14	-2.40	-0.03	-0.02	-0.28	-0.25	-0.17	-0.15	-0.19	-0.17
	Item15	-2.53	-0.03	-0.05	-1.87	-1.83	-0.87	-0.89	-1.08	-1.09
	Item16	1.39	0.01	0.00	0.14	0.12	0.11	0.10	0.12	0.11
	Item17	2.51	0.03	0.03	0.62	0.59	0.30	0.29	0.32	0.31
	Item18	2.32	0.03	0.04	1.00	0.99	0.48	0.49	0.54	0.54
	Item19	2.68	0.05	0.06	1.22	1.18	0.55	0.54	0.62	0.61
	Item20	-0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Item21	1.40	0.03	0.02	0.15	0.13	0.12	0.10	0.12	0.11
	Item22	-0.75	0.00	0.00	-0.04	-0.04	-0.05	-0.05	-0.06	-0.06
	Item23	-1.03	-0.01	-0.01	-0.07	-0.07	-0.08	-0.08	-0.11	-0.11

Table A**46**
Bias of estimated item parameters for manipulation of the 25% of items and 30% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4506	4508	4434	4433
	Item24	-2.17	-0.01	-0.03	-0.83	-0.83	-0.43	-0.44	-0.52	-0.53
	Item25	0.85	0.01	0.01	0.08	0.08	0.09	0.09	0.10	0.11
	Item26	0.14	0.01	0.01	0.01	0.01	0.03	0.03	0.03	0.03
	Item27	1.98	0.03	0.02	0.27	0.24	0.18	0.16	0.18	0.16
	Item28	1.11	0.01	0.01	0.12	0.12	0.11	0.11	0.13	0.13
	Item29	0.80	0.02	0.02	0.07	0.08	0.09	0.09	0.10	0.10
	Item30	1.18	0.02	0.01	0.12	0.10	0.10	0.09	0.11	0.10
	Item31	0.20	0.00	0.00	0.01	0.01	0.03	0.03	0.03	0.03
	Item32	1.62	0.02	0.02	0.21	0.20	0.16	0.16	0.17	0.17
	Item33	1.15	0.01	0.02	0.15	0.15	0.13	0.13	0.15	0.15
	Item34	-2.45	-0.01	0.00	-0.28	-0.25	-0.16	-0.14	-0.19	-0.17
	Item35	-1.16	-0.01	-0.01	-0.10	-0.10	-0.10	-0.10	-0.13	-0.13
	Item36	-0.90	-0.01	-0.01	-0.06	-0.06	-0.08	-0.08	-0.10	-0.10
	Item37	-0.54	0.00	0.00	-0.03	-0.03	-0.04	-0.04	-0.06	-0.06
	Item38	-1.27	-0.01	-0.01	-0.14	-0.14	-0.12	-0.12	-0.15	-0.15
	Item39	-2.29	-0.04	0.00	-0.21	-0.15	-0.14	-0.08	-0.13	-0.09
	Item40	-1.25	-0.01	-0.01	-0.10	-0.09	-0.10	-0.10	-0.12	-0.12
MAD	NA	NA	0.02	0.02	0.27	0.26	0.17	0.16	0.19	0.19
Item discrimination	Item1	1.77	-0.01	-0.02	-0.50	-0.51	-0.35	-0.35	-0.36	-0.36
	Item2	1.54	-0.02	-0.02	-0.35	-0.35	-0.28	-0.28	-0.29	-0.29
	Item3	2.16	-0.02	-0.04	-1.26	-1.26	-0.86	-0.87	-0.88	-0.89
	Item4	2.39	-0.02	-0.03	-1.00	-1.01	-0.64	-0.65	-0.68	-0.68
	Item5	2.05	0.00	-0.01	-0.76	-0.76	-0.47	-0.48	-0.49	-0.50
	Item6	1.80	-0.02	-0.02	-0.45	-0.45	-0.37	-0.38	-0.39	-0.39
	Item7	1.34	-0.02	-0.02	-0.46	-0.46	-0.27	-0.28	-0.29	-0.29
	Item8	0.57	-0.01	0.00	-0.10	-0.09	-0.06	-0.06	-0.05	-0.04
	Item9	0.78	0.00	0.00	-0.14	-0.14	-0.09	-0.09	-0.08	-0.08
	Item10	2.13	-0.02	-0.03	-0.97	-0.97	-0.60	-0.61	-0.62	-0.62
	Item11	0.92	-0.01	-0.01	-0.26	-0.26	-0.15	-0.14	-0.15	-0.15
	Item12	1.44	0.00	0.00	-0.36	-0.36	-0.24	-0.24	-0.25	-0.25
	Item13	1.05	-0.01	-0.01	-0.18	-0.18	-0.14	-0.14	-0.14	-0.14
	Item14	0.71	-0.01	0.00	-0.16	-0.15	-0.08	-0.08	-0.08	-0.08
	Item15	2.46	0.00	-0.07	-1.77	-1.76	-1.33	-1.34	-1.41	-1.42
	Item16	0.68	-0.01	0.00	-0.13	-0.12	-0.08	-0.08	-0.07	-0.07
	Item17	0.96	0.00	0.00	-0.33	-0.33	-0.18	-0.18	-0.17	-0.17
	Item18	1.71	-0.01	-0.03	-0.92	-0.92	-0.59	-0.60	-0.61	-0.62
	Item19	1.31	-0.01	-0.02	-0.67	-0.66	-0.40	-0.40	-0.42	-0.42
	Item20	0.58	-0.01	0.00	-0.09	-0.08	-0.06	-0.05	-0.05	-0.04
	Item21	0.60	-0.01	-0.01	-0.11	-0.10	-0.07	-0.06	-0.06	-0.05
	Item22	0.77	-0.01	0.00	-0.12	-0.12	-0.09	-0.08	-0.08	-0.08
	Item23	1.45	-0.01	-0.01	-0.34	-0.34	-0.24	-0.24	-0.24	-0.24
	Item24	2.32	-0.02	-0.06	-1.38	-1.38	-0.95	-0.96	-1.01	-1.02
	Item25	2.47	-0.01	-0.02	-0.92	-0.92	-0.62	-0.63	-0.64	-0.65
	Item26	1.53	-0.03	-0.03	-0.34	-0.34	-0.28	-0.28	-0.29	-0.29
	Item27	0.70	-0.01	0.00	-0.16	-0.15	-0.09	-0.09	-0.08	-0.07

Table A**46**
Bias of estimated item parameters for manipulation of the 25% of items and 30% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4506	4508	4434	4433
	Item28	1.79	-0.01	-0.01	-0.57	-0.57	-0.36	-0.36	-0.36	-0.37
	Item29	1.63	-0.01	-0.02	-0.43	-0.43	-0.30	-0.30	-0.30	-0.30
	Item30	0.70	-0.01	0.00	-0.13	-0.12	-0.08	-0.08	-0.07	-0.07
	Item31	2.04	-0.02	-0.03	-0.58	-0.58	-0.47	-0.48	-0.49	-0.49
	Item32	0.96	-0.01	-0.01	-0.24	-0.23	-0.14	-0.14	-0.14	-0.13
	Item33	1.89	-0.01	-0.02	-0.65	-0.65	-0.41	-0.42	-0.42	-0.43
	Item34	0.73	0.00	0.00	-0.16	-0.16	-0.08	-0.08	-0.08	-0.08
	Item35	1.96	0.01	0.00	-0.61	-0.61	-0.39	-0.40	-0.41	-0.42
	Item36	1.03	-0.02	-0.02	-0.20	-0.20	-0.15	-0.15	-0.15	-0.15
	Item37	1.20	-0.01	-0.01	-0.22	-0.22	-0.17	-0.17	-0.17	-0.17
	Item38	2.42	-0.03	-0.04	-0.95	-0.96	-0.62	-0.63	-0.65	-0.66
	Item39	0.51	-0.01	0.00	-0.09	-0.08	-0.05	-0.04	-0.04	-0.03
	Item40	0.95	-0.01	-0.01	-0.18	-0.18	-0.12	-0.12	-0.12	-0.12
MAD	NA	NA	0.01	0.02	0.48	0.48	0.32	0.32	0.33	0.33

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**47**
Bias of estimated item parameters for manipulation of the 50% of items and 10% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4639	4639	4561	4564
Item difficulty	Item1	1.92	0.02	0.02	-0.04	-0.04	-0.06	-0.06	0.00	0.00
	Item2	-2.00	-0.04	-0.02	0.29	0.30	0.12	0.13	0.09	0.10
	Item3	-1.11	-0.02	-0.02	0.03	0.03	0.00	0.00	-0.03	-0.03
	Item4	2.56	0.02	0.03	0.08	0.09	-0.13	-0.12	0.03	0.04
	Item5	-0.02	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00
	Item6	-0.33	-0.01	-0.01	-0.01	-0.01	0.00	0.00	-0.01	-0.01
	Item7	-1.90	-0.02	-0.01	0.25	0.26	0.11	0.12	0.08	0.09
	Item8	0.97	0.01	0.01	0.02	0.02	0.01	0.01	0.03	0.03
	Item9	1.01	0.01	0.01	-0.08	-0.08	-0.04	-0.05	-0.04	-0.05
	Item10	1.96	0.01	0.01	0.03	0.03	-0.06	-0.06	0.02	0.02
	Item11	0.57	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.02
	Item12	-2.12	-0.03	-0.05	0.24	0.23	0.10	0.09	-0.04	-0.05
	Item13	-1.35	-0.02	-0.02	0.02	0.02	0.00	0.00	-0.07	-0.07
	Item14	-1.67	-0.01	-0.01	0.15	0.15	0.06	0.07	0.01	0.01
	Item15	-1.04	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02	-0.06	-0.06
	Item16	1.88	0.01	0.02	0.09	0.09	-0.05	-0.04	0.03	0.04
	Item17	-1.50	-0.02	-0.02	0.07	0.06	0.02	0.02	-0.05	-0.05
	Item18	1.53	0.00	0.01	0.08	0.08	-0.02	-0.02	0.03	0.04
	Item19	-2.12	-0.05	-0.02	0.34	0.36	0.15	0.16	0.13	0.15
	Item20	-0.69	-0.01	-0.01	0.00	0.00	0.00	0.00	-0.02	-0.02
	Item21	-2.07	-0.03	-0.04	0.23	0.22	0.10	0.09	-0.03	-0.05
	Item22	1.86	0.02	0.02	0.06	0.07	-0.04	-0.04	0.04	0.04
	Item23	0.22	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01

Table A**47**
Bias of estimated item parameters for manipulation of the 50% of items and 10% of sample (40 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4639	4639	4561	4564
	Item24	-1.80	-0.02	-0.02	0.14	0.14	0.06	0.05	-0.04	-0.05
	Item25	2.68	0.04	0.02	-0.23	-0.25	-0.17	-0.18	-0.13	-0.15
	Item26	-2.62	-0.02	-0.02	0.46	0.46	0.22	0.22	0.06	0.06
	Item27	-1.65	-0.02	-0.02	0.10	0.09	0.04	0.03	-0.05	-0.06
	Item28	-1.76	-0.01	-0.02	0.14	0.13	0.06	0.05	-0.04	-0.04
	Item29	2.59	0.02	0.02	-0.10	-0.10	-0.15	-0.15	-0.03	-0.03
	Item30	-0.82	-0.01	0.00	0.04	0.04	0.02	0.02	0.00	0.00
	Item31	-2.39	-0.04	-0.04	0.34	0.34	0.15	0.14	0.00	0.00
	Item32	-2.18	-0.02	0.00	0.36	0.37	0.17	0.18	0.13	0.15
	Item33	-2.27	-0.02	-0.03	0.32	0.30	0.15	0.14	-0.01	-0.02
	Item34	-0.57	-0.01	-0.01	0.02	0.02	0.01	0.01	0.00	0.00
	Item35	0.13	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.02
	Item36	-2.31	-0.03	-0.05	0.31	0.30	0.14	0.12	-0.02	-0.04
	Item37	0.32	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00
	Item38	-1.54	-0.02	-0.02	0.07	0.06	0.02	0.02	-0.06	-0.06
	Item39	0.19	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
	Item40	0.35	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.02
MAD	NA	NA	0.02	0.02	0.12	0.12	0.06	0.06	0.04	0.04
Item discrimination	Item1	0.96	-0.01	0.00	-0.06	-0.06	0.04	0.04	0.02	0.02
	Item2	0.58	-0.01	0.00	0.08	0.08	0.04	0.05	0.04	0.05
	Item3	1.01	-0.02	-0.02	-0.02	-0.02	0.00	0.00	-0.01	-0.01
	Item4	1.75	0.00	-0.02	-0.45	-0.45	0.08	0.06	-0.08	-0.10
	Item5	2.10	-0.02	-0.03	-0.51	-0.52	-0.19	-0.20	-0.23	-0.23
	Item6	0.99	-0.01	-0.01	-0.05	-0.05	-0.01	-0.01	-0.01	-0.01
	Item7	0.61	-0.01	0.00	0.07	0.07	0.04	0.04	0.04	0.04
	Item8	1.21	0.00	0.00	-0.13	-0.13	0.00	0.00	-0.01	-0.01
	Item9	0.58	0.00	0.00	0.02	0.03	0.03	0.03	0.05	0.05
	Item10	1.49	0.00	0.00	-0.22	-0.22	0.07	0.06	-0.01	-0.01
	Item11	1.13	-0.01	-0.01	-0.11	-0.11	-0.01	-0.01	-0.01	-0.01
	Item12	2.22	-0.03	-0.06	-0.01	-0.04	0.13	0.09	-0.08	-0.11
	Item13	2.21	-0.02	-0.03	-0.22	-0.23	0.00	-0.02	-0.12	-0.14
	Item14	0.88	0.00	0.00	0.04	0.04	0.04	0.04	0.02	0.02
	Item15	2.18	-0.02	-0.03	-0.32	-0.32	-0.07	-0.09	-0.17	-0.18
	Item16	2.14	0.00	-0.02	-0.56	-0.56	0.08	0.05	-0.07	-0.09
	Item17	1.50	-0.02	-0.03	-0.04	-0.04	0.02	0.02	-0.04	-0.05
	Item18	2.30	0.00	-0.01	-0.60	-0.60	0.06	0.04	-0.07	-0.08
	Item19	0.53	-0.01	0.00	0.08	0.08	0.04	0.05	0.05	0.05
	Item20	1.00	-0.01	-0.01	-0.04	-0.04	0.00	0.00	-0.01	-0.01
	Item21	2.43	-0.04	-0.08	-0.06	-0.09	0.14	0.09	-0.09	-0.14
	Item22	1.75	-0.02	-0.03	-0.34	-0.35	0.05	0.04	-0.05	-0.06
	Item23	1.00	-0.01	-0.01	-0.08	-0.08	-0.01	-0.01	-0.01	-0.01
	Item24	1.65	-0.01	-0.02	0.01	0.00	0.07	0.06	-0.04	-0.05
	Item25	0.63	-0.01	0.00	0.02	0.02	0.05	0.05	0.05	0.06
	Item26	0.93	0.00	0.00	0.15	0.14	0.10	0.09	0.04	0.04
	Item27	2.27	-0.03	-0.05	-0.11	-0.13	0.07	0.05	-0.08	-0.11

Table A**47**
Bias of estimated item parameters for manipulation of the 50% of items and 10% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4639	4639	4561	4564
	Item28	2.02	-0.01	-0.03	-0.02	-0.04	0.10	0.08	-0.04	-0.06
	Item29	1.05	-0.01	-0.01	-0.08	-0.08	0.07	0.07	0.02	0.02
	Item30	0.82	0.00	0.00	0.00	0.01	0.02	0.02	0.02	0.02
	Item31	1.09	-0.01	-0.02	0.12	0.12	0.09	0.08	0.02	0.02
	Item32	0.54	0.00	0.01	0.09	0.09	0.05	0.06	0.05	0.06
	Item33	1.84	-0.01	-0.03	0.11	0.08	0.16	0.14	-0.01	-0.04
	Item34	0.78	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.02
	Item35	2.19	-0.01	-0.01	-0.56	-0.56	-0.19	-0.20	-0.23	-0.23
	Item36	2.33	-0.01	-0.06	-0.04	-0.08	0.15	0.10	-0.09	-0.14
	Item37	0.92	0.00	0.00	-0.04	-0.04	0.01	0.01	0.01	0.01
	Item38	2.23	-0.04	-0.06	-0.15	-0.17	0.03	0.00	-0.11	-0.13
	Item39	1.13	-0.01	-0.01	-0.10	-0.10	-0.02	-0.02	-0.02	-0.02
	Item40	1.57	-0.02	-0.02	-0.27	-0.27	-0.08	-0.08	-0.09	-0.09
MAD	NA	NA	0.01	0.02	0.15	0.15	0.06	0.05	0.06	0.06

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using l_z; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**48**
Bias of estimated item parameters for manipulation of the 50% of items and 20% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4386	4386	4313	4311
Item difficulty	Item1	-1.43	-0.01	-0.02	0.09	0.09	0.07	0.07	0.01	0.01
	Item2	2.47	0.04	0.06	-0.33	-0.32	-0.26	-0.25	-0.19	-0.17
	Item3	0.40	0.00	0.00	0.06	0.06	0.04	0.04	0.04	0.04
	Item4	1.39	0.01	0.01	-0.18	-0.18	-0.09	-0.09	-0.10	-0.10
	Item5	-2.19	-0.02	-0.02	0.33	0.32	0.25	0.24	0.13	0.12
	Item6	2.56	0.04	0.04	-0.49	-0.49	-0.31	-0.30	-0.26	-0.26
	Item7	1.25	0.01	0.01	-0.02	-0.02	-0.01	0.00	0.00	0.00
	Item8	-2.54	0.00	-0.02	0.44	0.44	0.36	0.34	0.19	0.16
	Item9	-1.68	-0.01	-0.02	0.11	0.10	0.11	0.10	0.03	0.03
	Item10	-1.96	-0.02	-0.03	0.21	0.20	0.19	0.18	0.08	0.07
	Item11	0.61	0.00	0.00	-0.07	-0.07	-0.03	-0.03	-0.04	-0.04
	Item12	-0.93	-0.01	-0.01	0.00	0.00	0.00	0.00	-0.03	-0.03
	Item13	-2.44	-0.03	-0.04	0.46	0.46	0.34	0.33	0.18	0.18
	Item14	1.92	0.02	0.02	-0.14	-0.13	-0.13	-0.12	-0.10	-0.09
	Item15	2.17	0.02	0.03	-0.22	-0.21	-0.20	-0.19	-0.15	-0.14
	Item16	1.69	0.02	0.02	-0.12	-0.12	-0.07	-0.07	-0.06	-0.05
	Item17	-1.54	-0.02	-0.02	0.06	0.05	0.07	0.07	0.00	0.00
	Item18	-1.59	-0.02	-0.01	0.34	0.34	0.21	0.22	0.18	0.19
	Item19	-2.66	-0.03	0.00	0.76	0.76	0.49	0.50	0.40	0.42
	Item20	-0.71	0.00	0.00	0.08	0.08	0.05	0.05	0.04	0.04
	Item21	-0.60	0.00	0.00	-0.04	-0.04	-0.02	-0.02	-0.04	-0.04
	Item22	-0.85	0.00	0.00	-0.01	-0.01	0.00	0.00	-0.03	-0.03
	Item23	-0.35	0.00	0.00	-0.04	-0.04	-0.03	-0.03	-0.04	-0.04

Table A**48**
Bias of estimated item parameters for manipulation of the 50% of items and 20% of sample (40 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4386	4386	4313	4311
	Item24	-1.85	-0.02	-0.02	0.20	0.20	0.16	0.16	0.07	0.06
	Item25	2.08	0.03	0.04	-0.24	-0.24	-0.16	-0.16	-0.13	-0.12
	Item26	2.60	0.03	0.03	-0.53	-0.52	-0.34	-0.34	-0.30	-0.29
	Item27	-1.48	-0.02	-0.02	0.07	0.07	0.06	0.06	0.00	0.00
	Item28	-2.50	-0.03	-0.05	0.42	0.42	0.34	0.32	0.16	0.14
	Item29	2.10	0.01	0.01	-0.34	-0.35	-0.20	-0.20	-0.19	-0.20
	Item30	-1.02	-0.01	-0.01	0.00	0.00	0.00	0.00	-0.03	-0.02
	Item31	-0.08	0.00	0.00	0.02	0.02	0.02	0.02	0.01	0.01
	Item32	1.53	0.00	0.00	-0.12	-0.12	-0.07	-0.07	-0.06	-0.06
	Item33	0.72	0.00	0.01	0.08	0.08	0.06	0.06	0.05	0.06
	Item34	0.04	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00
	Item35	2.41	0.02	0.03	-0.33	-0.32	-0.27	-0.25	-0.21	-0.19
	Item36	1.14	0.02	0.01	-0.11	-0.11	-0.04	-0.05	-0.05	-0.05
	Item37	-1.22	-0.01	-0.01	-0.01	-0.01	0.01	0.01	-0.03	-0.03
	Item38	0.09	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00
	Item39	0.13	0.00	0.00	0.02	0.02	0.02	0.02	0.01	0.01
	Item40	2.56	0.02	0.02	-0.48	-0.48	-0.32	-0.32	-0.26	-0.26
MAD	NA	NA	0.01	0.02	0.19	0.19	0.14	0.13	0.10	0.09
Item discrimination	Item1	1.22	-0.01	-0.01	-0.10	-0.10	0.03	0.03	0.01	0.00
	Item2	2.34	-0.04	-0.09	-0.69	-0.71	0.02	-0.03	-0.13	-0.18
	Item3	2.36	-0.02	-0.02	-0.94	-0.94	-0.45	-0.46	-0.42	-0.43
	Item4	0.78	-0.01	-0.01	0.02	0.02	0.05	0.05	0.06	0.06
	Item5	1.49	-0.01	-0.02	-0.06	-0.07	0.16	0.15	0.06	0.05
	Item6	1.03	-0.02	-0.02	0.06	0.05	0.13	0.12	0.11	0.10
	Item7	1.39	-0.01	-0.01	-0.21	-0.21	-0.01	-0.02	-0.01	-0.01
	Item8	2.24	0.03	-0.02	-0.51	-0.53	0.04	0.00	-0.15	-0.19
	Item9	2.38	0.00	-0.02	-0.53	-0.54	0.12	0.09	-0.03	-0.05
	Item10	2.04	-0.03	-0.05	-0.34	-0.35	0.14	0.11	0.00	-0.02
	Item11	0.75	0.00	0.00	0.01	0.01	0.03	0.04	0.05	0.05
	Item12	1.38	-0.02	-0.02	-0.22	-0.22	-0.04	-0.04	-0.05	-0.06
	Item13	1.25	0.00	-0.01	0.05	0.04	0.18	0.17	0.09	0.08
	Item14	2.31	0.00	-0.03	-0.60	-0.60	0.12	0.09	0.03	0.00
	Item15	2.37	-0.01	-0.04	-0.65	-0.67	0.13	0.09	0.00	-0.04
	Item16	1.36	-0.01	-0.01	-0.14	-0.14	0.05	0.05	0.04	0.04
	Item17	2.30	-0.02	-0.04	-0.54	-0.55	0.06	0.04	-0.06	-0.08
	Item18	0.55	-0.01	0.00	0.09	0.10	0.07	0.08	0.08	0.08
	Item19	0.55	0.00	0.00	0.14	0.15	0.11	0.11	0.10	0.10
	Item20	0.73	0.00	0.00	0.02	0.02	0.04	0.04	0.05	0.05
	Item21	1.64	-0.01	-0.01	-0.41	-0.41	-0.14	-0.15	-0.15	-0.15
	Item22	1.42	-0.01	-0.01	-0.25	-0.25	-0.06	-0.06	-0.07	-0.07
	Item23	2.01	0.00	-0.01	-0.67	-0.67	-0.30	-0.30	-0.29	-0.29
	Item24	1.38	-0.01	-0.02	-0.08	-0.09	0.10	0.09	0.03	0.03
	Item25	1.41	-0.02	-0.02	-0.11	-0.11	0.11	0.10	0.08	0.07
	Item26	1.04	0.00	0.00	0.07	0.07	0.15	0.15	0.13	0.13
	Item27	1.49	-0.02	-0.02	-0.18	-0.19	0.04	0.03	-0.01	-0.02

Table A**48**
Bias of estimated item parameters for manipulation of the 50% of items and 20% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4386	4386	4313	4311
	Item28	2.04	-0.02	-0.05	-0.37	-0.38	0.10	0.06	-0.08	-0.11
	Item29	0.82	0.00	0.00	0.05	0.05	0.08	0.08	0.09	0.09
	Item30	1.51	-0.02	-0.02	-0.27	-0.27	-0.05	-0.06	-0.07	-0.07
	Item31	0.67	0.00	0.00	0.03	0.03	0.04	0.04	0.05	0.06
	Item32	1.14	0.01	0.01	-0.08	-0.08	0.05	0.05	0.05	0.05
	Item33	2.21	-0.01	-0.02	-0.78	-0.78	-0.31	-0.32	-0.29	-0.30
	Item34	2.34	-0.02	-0.03	-0.95	-0.95	-0.50	-0.50	-0.47	-0.47
	Item35	2.01	-0.01	-0.04	-0.42	-0.44	0.15	0.12	0.04	0.01
	Item36	0.84	-0.01	-0.01	-0.02	-0.01	0.02	0.02	0.04	0.04
	Item37	2.04	-0.01	-0.02	-0.50	-0.50	-0.04	-0.05	-0.11	-0.12
	Item38	1.51	-0.01	-0.01	-0.36	-0.36	-0.14	-0.14	-0.12	-0.12
	Item39	2.38	-0.02	-0.02	-0.97	-0.97	-0.51	-0.51	-0.48	-0.48
	Item40	1.21	0.00	0.00	0.03	0.02	0.17	0.16	0.13	0.12
MAD	NA	NA	0.01	0.02	0.31	0.32	0.13	0.12	0.11	0.11

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using l_z; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**49**
Bias of estimated item parameters for manipulation of the 50% of items and 30% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4319	4321	4197	4198
Item difficulty	Item1	0.75	0.01	0.01	0.09	0.09	0.08	0.08	0.10	0.10
	Item2	0.06	0.00	0.00	-0.02	-0.02	-0.04	-0.04	-0.02	-0.02
	Item3	2.41	0.04	0.04	-0.64	-0.65	-0.57	-0.58	-0.43	-0.43
	Item4	0.17	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.02
	Item5	-0.09	0.00	0.00	0.01	0.01	-0.02	-0.01	-0.01	-0.01
	Item6	-1.85	-0.01	-0.02	0.23	0.22	0.22	0.22	0.18	0.17
	Item7	1.62	0.02	0.02	-0.13	-0.12	-0.16	-0.16	-0.07	-0.07
	Item8	-2.15	-0.02	-0.03	0.38	0.37	0.36	0.35	0.29	0.28
	Item9	-0.19	0.00	0.00	0.02	0.02	-0.01	-0.01	0.00	0.00
	Item10	-1.61	-0.02	-0.01	0.40	0.40	0.27	0.28	0.27	0.28
	Item11	-0.93	-0.01	-0.01	-0.03	-0.03	-0.05	-0.05	-0.04	-0.05
	Item12	-1.70	-0.01	-0.02	0.15	0.14	0.15	0.15	0.12	0.11
	Item13	-2.54	-0.02	-0.04	0.59	0.58	0.58	0.56	0.46	0.45
	Item14	-1.44	0.00	-0.01	0.04	0.04	0.05	0.04	0.03	0.03
	Item15	1.13	0.01	0.01	0.00	0.00	-0.02	-0.02	0.03	0.03
	Item16	-1.46	-0.01	-0.01	0.07	0.07	0.05	0.05	0.04	0.03
	Item17	0.69	0.00	0.00	0.07	0.07	0.06	0.06	0.08	0.08
	Item18	2.02	0.03	0.02	-0.52	-0.52	-0.44	-0.45	-0.35	-0.35
	Item19	1.91	0.02	0.02	-0.29	-0.29	-0.29	-0.29	-0.18	-0.17
	Item20	1.99	0.02	0.02	-0.40	-0.40	-0.36	-0.36	-0.26	-0.26
	Item21	0.76	0.00	0.00	-0.12	-0.12	-0.11	-0.11	-0.08	-0.08
	Item22	0.25	0.00	0.00	0.03	0.03	0.03	0.03	0.03	0.03
	Item23	1.53	0.00	0.01	-0.12	-0.11	-0.15	-0.14	-0.06	-0.06

Table A**49**
Bias of estimated item parameters for manipulation of the 50% of items and 30% of sample (40 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4319	4321	4197	4198
	Item24	1.89	0.01	0.02	-0.26	-0.25	-0.30	-0.29	-0.17	-0.17
	Item25	-1.85	-0.01	0.00	0.46	0.47	0.33	0.33	0.33	0.33
	Item26	-0.24	0.00	0.00	0.06	0.06	0.01	0.01	0.02	0.02
	Item27	0.81	0.00	0.00	-0.04	-0.04	-0.04	-0.04	-0.01	-0.01
	Item28	-1.72	-0.02	0.00	0.48	0.48	0.33	0.34	0.34	0.34
	Item29	-2.35	0.00	-0.02	0.48	0.48	0.48	0.47	0.38	0.37
	Item30	1.89	0.01	0.00	-0.42	-0.42	-0.36	-0.36	-0.27	-0.27
	Item31	1.50	0.01	0.01	-0.08	-0.08	-0.13	-0.12	-0.04	-0.04
	Item32	-1.76	0.00	0.00	0.25	0.25	0.18	0.18	0.16	0.16
	Item33	-2.65	-0.02	-0.04	0.68	0.67	0.65	0.63	0.54	0.52
	Item34	-1.03	0.00	0.00	-0.04	-0.04	-0.06	-0.06	-0.05	-0.06
	Item35	2.50	0.03	0.02	-0.71	-0.71	-0.64	-0.64	-0.49	-0.49
	Item36	2.23	0.01	0.03	-0.39	-0.39	-0.42	-0.41	-0.26	-0.25
	Item37	-0.98	0.00	0.00	-0.04	-0.04	-0.06	-0.06	-0.05	-0.05
	Item38	1.77	0.02	0.02	-0.24	-0.24	-0.25	-0.24	-0.14	-0.14
	Item39	0.78	0.00	0.00	0.07	0.07	0.05	0.05	0.08	0.08
	Item40	0.63	0.00	0.00	0.06	0.06	0.05	0.05	0.07	0.07
MAD	NA	NA	0.01	0.01	0.23	0.23	0.21	0.21	0.16	0.16
Item discrimination	Item1	2.48	0.00	-0.01	-1.13	-1.13	-0.67	-0.68	-0.66	-0.67
	Item2	0.65	0.00	0.00	0.05	0.05	0.07	0.07	0.09	0.09
	Item3	0.94	-0.01	-0.01	0.10	0.10	0.20	0.20	0.14	0.14
	Item4	2.02	-0.02	-0.02	-0.88	-0.88	-0.62	-0.62	-0.57	-0.57
	Item5	0.91	-0.01	0.00	-0.09	-0.09	-0.02	-0.02	0.00	0.00
	Item6	2.03	0.00	-0.01	-0.51	-0.52	0.15	0.12	0.05	0.02
	Item7	2.04	-0.02	-0.03	-0.54	-0.55	-0.01	-0.03	-0.13	-0.14
	Item8	2.01	0.00	-0.03	-0.46	-0.47	0.26	0.22	0.10	0.07
	Item9	0.88	0.00	0.00	-0.06	-0.06	0.00	0.00	0.01	0.01
	Item10	0.64	0.00	0.00	0.10	0.10	0.12	0.12	0.13	0.13
	Item11	1.55	-0.02	-0.02	-0.44	-0.44	-0.22	-0.22	-0.18	-0.19
	Item12	2.29	-0.01	-0.03	-0.73	-0.74	0.05	0.02	-0.05	-0.07
	Item13	2.35	0.00	-0.06	-0.72	-0.74	0.15	0.09	-0.10	-0.15
	Item14	2.22	0.01	0.00	-0.74	-0.75	-0.12	-0.14	-0.16	-0.17
	Item15	1.83	-0.01	-0.02	-0.54	-0.54	-0.19	-0.19	-0.22	-0.23
	Item16	1.82	-0.01	-0.02	-0.50	-0.50	-0.08	-0.09	-0.09	-0.10
	Item17	2.20	-0.01	-0.02	-0.93	-0.93	-0.56	-0.56	-0.55	-0.55
	Item18	0.74	-0.01	-0.01	0.10	0.10	0.15	0.15	0.12	0.12
	Item19	1.40	-0.01	-0.02	-0.15	-0.15	0.11	0.10	0.03	0.02
	Item20	1.04	-0.01	-0.01	0.01	0.01	0.14	0.14	0.10	0.09
	Item21	0.78	0.00	0.00	0.00	0.01	0.05	0.05	0.06	0.06
	Item22	1.95	-0.02	-0.02	-0.82	-0.82	-0.57	-0.57	-0.52	-0.52
	Item23	1.87	0.00	-0.01	-0.45	-0.46	0.01	0.00	-0.09	-0.10
	Item24	2.13	0.00	-0.02	-0.53	-0.54	0.10	0.07	-0.07	-0.10
	Item25	0.72	0.00	0.00	0.10	0.10	0.13	0.13	0.15	0.15
	Item26	0.58	0.00	0.00	0.08	0.08	0.09	0.09	0.10	0.10
	Item27	1.18	0.00	0.00	-0.21	-0.21	-0.07	-0.07	-0.08	-0.07

Table A**49**
Bias of estimated item parameters for manipulation of the 50% of items and 30% of sample (40 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4319	4321	4197	4198
	Item28	0.61	0.00	0.00	0.12	0.12	0.13	0.13	0.14	0.15
	Item29	2.28	0.02	-0.03	-0.65	-0.66	0.24	0.19	0.01	-0.03
	Item30	0.84	0.00	0.00	0.06	0.06	0.13	0.13	0.10	0.10
	Item31	2.39	0.00	-0.01	-0.79	-0.79	-0.06	-0.08	-0.20	-0.21
	Item32	1.23	0.00	-0.01	-0.11	-0.12	0.08	0.07	0.08	0.07
	Item33	2.05	0.00	-0.05	-0.48	-0.49	0.26	0.21	0.05	0.01
	Item34	1.85	0.00	0.00	-0.62	-0.62	-0.29	-0.30	-0.26	-0.26
	Item35	0.90	-0.01	0.00	0.13	0.13	0.22	0.22	0.16	0.16
	Item36	2.42	0.02	-0.02	-0.75	-0.76	-0.10	-0.13	-0.30	-0.32
	Item37	1.75	0.00	0.00	-0.56	-0.56	-0.27	-0.27	-0.22	-0.22
	Item38	1.46	-0.01	-0.01	-0.19	-0.19	0.10	0.09	0.02	0.01
	Item39	2.09	-0.01	-0.02	-0.83	-0.83	-0.46	-0.46	-0.45	-0.46
	Item40	1.98	0.00	0.00	-0.77	-0.77	-0.46	-0.47	-0.45	-0.45
MAD	NA	NA	0.01	0.01	0.43	0.43	0.19	0.19	0.18	0.18

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**50**
Bias of estimated item parameters for manipulation of the 25% of items and 10% of sample (60 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4619	4620	4549	4550
Item difficulty	Item1	2.35	0.01	0.02	0.12	0.13	0.04	0.05	0.14	0.14
	Item2	-0.41	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.02	-0.02
	Item3	-1.83	0.00	-0.01	-0.22	-0.22	-0.04	-0.04	-0.08	-0.08
	Item4	-0.29	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01
	Item5	-0.25	0.00	0.00	0.01	0.01	-0.01	-0.01	-0.01	-0.01
	Item6	2.21	0.01	0.02	0.05	0.06	0.02	0.03	0.10	0.11
	Item7	1.61	0.00	-0.01	0.01	-0.01	0.01	0.00	0.02	0.00
	Item8	-1.09	-0.02	-0.01	-0.05	-0.05	-0.03	-0.03	-0.04	-0.04
	Item9	1.28	0.01	0.01	0.00	0.00	0.02	0.02	0.06	0.06
	Item10	1.09	0.01	0.01	0.00	0.00	0.02	0.02	0.06	0.06
	Item11	-0.01	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.01
	Item12	1.06	0.01	0.01	0.00	0.00	0.02	0.02	0.05	0.05
	Item13	0.73	0.01	0.01	0.00	0.00	0.01	0.01	0.04	0.04
	Item14	-2.08	-0.02	-0.02	-0.34	-0.34	-0.07	-0.08	-0.12	-0.13
	Item15	0.77	0.01	0.01	0.01	0.01	0.02	0.01	0.03	0.03
	Item16	-0.92	-0.01	-0.01	-0.04	-0.03	-0.02	-0.02	-0.04	-0.04
	Item17	1.85	0.02	0.03	0.07	0.08	0.04	0.05	0.11	0.12
	Item18	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02
	Item19	2.36	0.00	0.01	0.16	0.17	0.03	0.05	0.13	0.14
	Item20	-1.04	-0.01	-0.01	-0.05	-0.05	-0.02	-0.02	-0.04	-0.04
	Item21	-1.58	-0.02	-0.02	-0.20	-0.21	-0.05	-0.05	-0.08	-0.08
	Item22	-2.70	-0.02	-0.03	-0.64	-0.64	-0.11	-0.12	-0.21	-0.22
	Item23	1.35	0.02	0.02	0.01	0.01	0.03	0.03	0.06	0.06

Table A**50**
Bias of estimated item parameters for manipulation of the 25% of items and 10% of sample (60 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4619	4620	4549	4550
	Item24	2.39	0.01	0.02	0.12	0.13	0.04	0.05	0.14	0.15
	Item25	0.79	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.03
	Item26	2.28	0.01	0.01	0.03	0.02	0.03	0.03	0.09	0.09
	Item27	0.30	0.00	0.00	0.01	0.01	0.00	0.00	0.02	0.02
	Item28	2.61	0.00	0.01	0.11	0.12	0.02	0.03	0.13	0.14
	Item29	1.13	0.01	0.01	0.00	0.00	0.02	0.02	0.05	0.05
	Item30	-1.00	0.00	0.00	-0.04	-0.04	-0.02	-0.02	-0.03	-0.03
	Item31	1.78	0.01	0.01	0.03	0.04	0.02	0.03	0.08	0.09
	Item32	-0.78	-0.01	-0.01	-0.03	-0.03	-0.03	-0.03	-0.04	-0.04
	Item33	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Item34	1.91	0.01	0.02	0.04	0.05	0.03	0.03	0.10	0.10
	Item35	0.89	0.01	0.01	0.00	0.00	0.01	0.02	0.04	0.04
	Item36	0.32	0.01	0.01	0.02	0.02	0.01	0.01	0.01	0.01
	Item37	-0.08	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01
	Item38	-1.93	-0.02	-0.02	-0.16	-0.16	-0.05	-0.04	-0.06	-0.06
	Item39	-2.11	-0.02	-0.02	-0.34	-0.34	-0.08	-0.08	-0.12	-0.13
	Item40	2.55	0.02	0.04	0.28	0.29	0.07	0.09	0.19	0.21
	Item41	2.46	0.02	0.03	0.22	0.23	0.06	0.07	0.16	0.18
	Item42	0.70	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.03
	Item43	0.60	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.03
	Item44	1.39	0.01	0.01	0.00	0.00	0.02	0.02	0.05	0.05
	Item45	-2.62	0.00	0.00	-0.41	-0.40	-0.06	-0.06	-0.12	-0.12
	Item46	0.38	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02
	Item47	1.58	0.00	0.01	0.01	0.01	0.02	0.02	0.07	0.07
	Item48	-1.78	-0.01	-0.02	-0.30	-0.30	-0.06	-0.07	-0.10	-0.10
	Item49	1.61	0.00	0.00	0.01	0.00	0.02	0.02	0.06	0.06
	Item50	-1.36	-0.01	-0.01	-0.14	-0.14	-0.04	-0.04	-0.06	-0.06
	Item51	0.43	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.02
	Item52	2.19	0.01	0.02	0.07	0.07	0.03	0.04	0.11	0.12
	Item53	1.96	0.00	0.01	0.06	0.07	0.02	0.03	0.10	0.10
	Item54	1.23	0.01	0.01	-0.01	0.00	0.02	0.02	0.05	0.06
	Item55	1.37	-0.01	-0.01	-0.01	-0.01	0.01	0.00	0.03	0.03
	Item56	-1.90	-0.01	-0.01	-0.22	-0.22	-0.04	-0.05	-0.08	-0.08
	Item57	-1.04	-0.01	-0.01	-0.06	-0.06	-0.03	-0.03	-0.04	-0.04
	Item58	0.97	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.04
	Item59	0.56	0.00	-0.01	0.01	0.00	0.00	0.00	0.00	-0.01
	Item60	2.19	0.02	0.02	0.06	0.06	0.03	0.04	0.11	0.11
MAD	NA	NA	0.01	0.01	0.08	0.08	0.03	0.03	0.06	0.07
Item discrimination	Item1	1.54	0.00	-0.01	-0.25	-0.26	-0.03	-0.04	-0.09	-0.10
	Item2	1.00	0.00	0.00	-0.07	-0.06	-0.01	-0.01	-0.01	-0.01
	Item3	1.63	0.00	-0.01	-0.38	-0.38	-0.06	-0.06	-0.08	-0.09
	Item4	0.58	0.00	0.00	-0.03	-0.02	-0.01	0.00	0.01	0.01
	Item5	2.49	-0.01	-0.01	-0.35	-0.36	-0.10	-0.10	-0.15	-0.16
	Item6	1.32	0.00	0.00	-0.14	-0.14	-0.01	-0.02	-0.05	-0.06
	Item7	0.56	0.00	0.00	-0.02	-0.01	0.00	0.00	0.00	0.01

Table A**50**

Bias of estimated item parameters for manipulation of the 25% of items and 10% of sample (60 items)

	True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
					N =	4619	4620	4549	4550
Item8	0.96	-0.01	-0.01	-0.10	-0.10	-0.03	-0.02	-0.02	-0.02
Item9	2.22	-0.01	-0.02	-0.23	-0.24	-0.05	-0.06	-0.11	-0.12
Item10	2.07	-0.02	-0.03	-0.17	-0.18	-0.06	-0.07	-0.11	-0.12
Item11	0.94	-0.01	-0.01	-0.05	-0.05	-0.02	-0.01	-0.01	-0.01
Item12	1.42	-0.01	-0.02	-0.08	-0.08	-0.03	-0.03	-0.05	-0.06
Item13	2.01	-0.01	-0.02	-0.13	-0.14	-0.05	-0.06	-0.09	-0.10
Item14	1.70	-0.01	-0.02	-0.49	-0.49	-0.09	-0.10	-0.13	-0.14
Item15	1.12	-0.01	-0.01	-0.06	-0.06	-0.02	-0.03	-0.04	-0.04
Item16	0.94	0.00	0.00	-0.08	-0.08	-0.02	-0.01	-0.01	-0.01
Item17	2.02	-0.02	-0.04	-0.35	-0.36	-0.07	-0.09	-0.14	-0.16
Item18	2.11	-0.02	-0.03	-0.17	-0.17	-0.08	-0.09	-0.13	-0.13
Item19	1.78	0.01	0.00	-0.38	-0.39	-0.04	-0.06	-0.12	-0.14
Item20	1.67	-0.01	-0.01	-0.26	-0.26	-0.04	-0.04	-0.05	-0.05
Item21	2.30	-0.01	-0.02	-0.69	-0.69	-0.11	-0.13	-0.16	-0.17
Item22	1.43	0.00	-0.02	-0.47	-0.47	-0.09	-0.10	-0.14	-0.15
Item23	1.04	-0.02	-0.02	-0.06	-0.06	-0.02	-0.02	-0.04	-0.04
Item24	1.53	0.00	-0.01	-0.25	-0.26	-0.03	-0.04	-0.09	-0.10
Item25	1.17	-0.01	-0.01	-0.06	-0.06	-0.02	-0.03	-0.04	-0.04
Item26	0.94	0.00	-0.01	-0.06	-0.06	-0.01	-0.01	-0.03	-0.03
Item27	1.93	-0.02	-0.02	-0.14	-0.15	-0.06	-0.07	-0.10	-0.10
Item28	1.31	0.01	0.01	-0.18	-0.19	-0.01	-0.02	-0.06	-0.07
Item29	1.86	-0.01	-0.02	-0.14	-0.14	-0.04	-0.05	-0.08	-0.09
Item30	1.08	0.00	0.00	-0.10	-0.10	-0.02	-0.02	-0.01	-0.01
Item31	1.82	0.00	-0.01	-0.23	-0.24	-0.03	-0.05	-0.09	-0.10
Item32	1.70	-0.01	-0.02	-0.23	-0.23	-0.05	-0.05	-0.06	-0.06
Item33	0.73	0.00	0.00	-0.03	-0.03	-0.01	0.00	0.00	0.00
Item34	1.58	0.00	-0.01	-0.18	-0.19	-0.03	-0.04	-0.08	-0.08
Item35	1.63	-0.01	-0.02	-0.09	-0.10	-0.04	-0.04	-0.07	-0.07
Item36	0.81	-0.01	-0.01	-0.04	-0.04	-0.02	-0.01	-0.01	-0.01
Item37	0.65	-0.01	-0.01	-0.04	-0.04	-0.01	-0.01	0.00	0.00
Item38	0.94	-0.01	0.00	-0.13	-0.13	-0.02	-0.02	-0.01	-0.01
Item39	1.58	-0.01	-0.02	-0.43	-0.43	-0.08	-0.09	-0.11	-0.12
Item40	1.84	-0.01	-0.03	-0.50	-0.52	-0.08	-0.11	-0.18	-0.21
Item41	1.76	-0.01	-0.03	-0.42	-0.43	-0.06	-0.08	-0.15	-0.17
Item42	2.18	-0.01	-0.02	-0.15	-0.16	-0.06	-0.07	-0.11	-0.11
Item43	2.18	-0.02	-0.02	-0.15	-0.16	-0.06	-0.07	-0.12	-0.12
Item44	1.12	0.00	0.00	-0.05	-0.05	-0.01	-0.01	-0.02	-0.02
Item45	1.12	0.01	0.01	-0.25	-0.25	-0.03	-0.03	-0.04	-0.04
Item46	1.88	-0.01	-0.01	-0.12	-0.13	-0.06	-0.06	-0.09	-0.09
Item47	1.83	0.00	-0.01	-0.18	-0.19	-0.02	-0.04	-0.08	-0.08
Item48	2.41	-0.01	-0.03	-0.86	-0.87	-0.17	-0.19	-0.23	-0.25
Item49	0.95	0.00	0.00	-0.04	-0.04	-0.01	-0.01	-0.02	-0.02
Item50	2.49	0.00	-0.01	-0.72	-0.72	-0.10	-0.12	-0.15	-0.16
Item51	1.58	-0.01	-0.02	-0.10	-0.10	-0.04	-0.05	-0.07	-0.07
Item52	1.49	0.00	0.00	-0.19	-0.20	-0.02	-0.03	-0.07	-0.08

Table A**50**
Bias of estimated item parameters for manipulation of the 25% of items and 10% of sample (60 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4619	4620	4549	4550
	Item53	1.89	0.00	-0.01	-0.30	-0.31	-0.04	-0.06	-0.11	-0.13
	Item54	2.29	-0.01	-0.02	-0.24	-0.25	-0.06	-0.07	-0.11	-0.12
	Item55	0.80	0.00	0.00	-0.03	-0.02	0.00	0.00	-0.01	0.00
	Item56	1.49	-0.01	-0.01	-0.32	-0.33	-0.05	-0.05	-0.07	-0.07
	Item57	1.82	-0.01	-0.01	-0.31	-0.31	-0.05	-0.06	-0.07	-0.08
	Item58	1.23	0.00	-0.01	-0.06	-0.06	-0.02	-0.02	-0.03	-0.04
	Item59	0.55	0.00	0.00	-0.02	-0.02	0.00	0.00	0.01	0.01
	Item60	1.27	-0.01	-0.01	-0.13	-0.14	-0.02	-0.03	-0.06	-0.06
MAD	NA	NA	0.01	0.01	0.21	0.21	0.04	0.05	0.07	0.08

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**51**
Bias of estimated item parameters for manipulation of the 25% of items and 20% of sample (60 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4460	4462	4403	4403
Item difficulty	Item1	0.63	0.01	0.01	0.01	0.01	0.03	0.03	0.04	0.04
	Item2	-0.89	-0.01	-0.01	-0.05	-0.05	-0.05	-0.05	-0.06	-0.06
	Item3	1.85	0.02	0.02	0.19	0.20	0.09	0.10	0.12	0.12
	Item4	0.19	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01
	Item5	-1.48	-0.01	0.00	-0.10	-0.08	-0.06	-0.04	-0.07	-0.06
	Item6	2.06	0.02	0.03	0.38	0.39	0.14	0.15	0.18	0.19
	Item7	1.02	0.01	0.00	0.04	0.03	0.03	0.02	0.04	0.03
	Item8	-1.39	-0.01	-0.01	-0.16	-0.16	-0.07	-0.08	-0.09	-0.10
	Item9	1.21	0.02	0.00	0.06	0.04	0.05	0.03	0.04	0.02
	Item10	-0.38	0.00	0.00	-0.01	-0.01	-0.02	-0.02	-0.03	-0.03
	Item11	-1.40	-0.02	-0.02	-0.19	-0.20	-0.09	-0.09	-0.11	-0.11
	Item12	0.68	0.00	0.00	0.01	0.01	0.02	0.02	0.04	0.03
	Item13	-0.13	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.01
	Item14	1.65	0.01	0.01	0.09	0.09	0.06	0.06	0.08	0.08
	Item15	-1.56	-0.01	-0.01	-0.19	-0.19	-0.08	-0.08	-0.10	-0.11
	Item16	1.49	0.01	0.01	0.11	0.11	0.06	0.07	0.08	0.09
	Item17	-2.23	0.00	0.01	-0.22	-0.20	-0.07	-0.06	-0.08	-0.07
	Item18	-0.11	0.00	0.00	0.00	0.00	-0.02	-0.02	-0.01	-0.01
	Item19	0.89	0.01	0.01	0.04	0.03	0.04	0.04	0.05	0.05
	Item20	-0.57	0.00	0.00	-0.01	-0.01	-0.03	-0.03	-0.04	-0.04
	Item21	2.41	0.01	0.01	0.18	0.17	0.07	0.07	0.09	0.09
	Item22	-2.35	0.00	0.01	-0.23	-0.20	-0.06	-0.05	-0.08	-0.06
	Item23	1.67	0.03	0.02	0.10	0.09	0.08	0.08	0.10	0.09
	Item24	0.74	0.01	0.01	0.02	0.03	0.04	0.04	0.05	0.05
	Item25	-1.92	-0.02	-0.02	-0.28	-0.28	-0.10	-0.10	-0.13	-0.13
	Item26	-1.69	0.00	0.00	-0.12	-0.11	-0.05	-0.05	-0.07	-0.06
	Item27	2.44	0.03	0.05	0.79	0.79	0.25	0.27	0.34	0.35
	Item28	2.21	0.02	0.02	0.19	0.18	0.08	0.09	0.11	0.11

Table A**51**
Bias of estimated item parameters for manipulation of the 25% of items and 20% of sample (60 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4460	4462	4403	4403
	Item29	-2.21	-0.02	0.00	-0.19	-0.16	-0.07	-0.04	-0.08	-0.06
	Item30	-1.38	-0.01	-0.01	-0.16	-0.16	-0.07	-0.08	-0.09	-0.10
	Item31	0.28	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
	Item32	-1.38	-0.01	-0.01	-0.11	-0.11	-0.06	-0.06	-0.08	-0.08
	Item33	-0.88	0.00	0.00	-0.04	-0.04	-0.03	-0.03	-0.04	-0.04
	Item34	-2.53	-0.04	-0.05	-1.00	-0.99	-0.34	-0.35	-0.40	-0.41
	Item35	-1.48	-0.01	0.00	-0.09	-0.08	-0.05	-0.04	-0.06	-0.05
	Item36	1.00	0.01	0.01	0.02	0.03	0.04	0.04	0.06	0.06
	Item37	-1.24	-0.01	-0.01	-0.10	-0.10	-0.07	-0.07	-0.08	-0.08
	Item38	-1.42	-0.02	-0.02	-0.11	-0.10	-0.07	-0.06	-0.08	-0.08
	Item39	0.91	0.01	0.00	0.02	0.02	0.03	0.03	0.04	0.04
	Item40	2.55	0.03	0.05	0.88	0.89	0.27	0.29	0.37	0.39
	Item41	1.62	0.02	0.02	0.13	0.13	0.07	0.08	0.10	0.10
	Item42	1.82	0.01	0.01	0.10	0.09	0.06	0.06	0.08	0.08
	Item43	0.70	0.01	0.01	0.02	0.02	0.03	0.03	0.04	0.04
	Item44	1.03	0.01	0.01	0.03	0.03	0.04	0.04	0.06	0.06
	Item45	1.21	0.02	0.02	0.05	0.05	0.06	0.06	0.07	0.07
	Item46	2.54	0.03	0.05	1.11	1.12	0.34	0.37	0.46	0.49
	Item47	-0.64	-0.01	-0.01	-0.04	-0.03	-0.03	-0.03	-0.04	-0.03
	Item48	-1.73	-0.01	-0.01	-0.18	-0.17	-0.08	-0.08	-0.10	-0.10
	Item49	1.74	0.01	0.02	0.15	0.16	0.08	0.08	0.10	0.11
	Item50	-1.33	-0.01	-0.01	-0.10	-0.10	-0.06	-0.06	-0.08	-0.08
	Item51	2.09	0.01	0.02	0.21	0.21	0.08	0.09	0.12	0.12
	Item52	1.41	0.01	0.01	0.06	0.06	0.05	0.05	0.07	0.07
	Item53	1.75	0.00	0.01	0.12	0.12	0.06	0.07	0.09	0.09
	Item54	-1.01	-0.01	-0.01	-0.07	-0.07	-0.05	-0.05	-0.06	-0.07
	Item55	1.14	0.01	0.01	0.04	0.04	0.04	0.05	0.06	0.06
	Item56	0.39	0.00	0.00	0.01	0.01	0.02	0.02	0.02	0.02
	Item57	1.41	0.02	0.02	0.09	0.09	0.06	0.07	0.08	0.09
	Item58	-0.04	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	0.00
	Item59	-2.60	-0.03	-0.05	-1.79	-1.76	-0.59	-0.61	-0.69	-0.71
	Item60	-0.28	0.00	0.00	0.00	0.00	-0.02	-0.02	-0.02	-0.02
MAD	NA	NA	0.01	0.01	0.18	0.18	0.08	0.08	0.10	0.10
Item discrimination	Item1	1.53	-0.01	-0.01	-0.21	-0.21	-0.13	-0.13	-0.14	-0.14
	Item2	2.30	-0.01	-0.02	-0.62	-0.62	-0.25	-0.26	-0.28	-0.28
	Item3	1.66	-0.01	-0.02	-0.46	-0.46	-0.16	-0.17	-0.18	-0.19
	Item4	2.14	-0.01	-0.02	-0.36	-0.37	-0.24	-0.24	-0.26	-0.26
	Item5	0.65	0.00	0.00	-0.09	-0.08	-0.04	-0.03	-0.03	-0.02
	Item6	2.16	-0.02	-0.05	-0.91	-0.92	-0.36	-0.38	-0.42	-0.44
	Item7	0.61	-0.01	0.00	-0.06	-0.06	-0.03	-0.02	-0.02	-0.02
	Item8	1.92	0.00	-0.01	-0.57	-0.57	-0.22	-0.22	-0.23	-0.24
	Item9	0.52	0.00	0.00	-0.05	-0.04	-0.02	-0.02	-0.01	-0.01
	Item10	0.98	-0.01	-0.01	-0.12	-0.12	-0.06	-0.06	-0.06	-0.06
	Item11	2.38	-0.02	-0.04	-0.88	-0.88	-0.35	-0.36	-0.38	-0.39
	Item12	1.39	-0.01	-0.01	-0.17	-0.18	-0.10	-0.10	-0.10	-0.10

Table A**51**
Bias of estimated item parameters for manipulation of the 25% of items and 20% of sample (60 items)

	True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
					N =	4460	4462	4403	4403
Item13	1.25	0.00	0.00	-0.14	-0.14	-0.08	-0.08	-0.08	-0.08
Item14	1.27	0.00	-0.01	-0.22	-0.23	-0.08	-0.09	-0.09	-0.09
Item15	1.64	0.00	-0.01	-0.46	-0.46	-0.16	-0.17	-0.18	-0.18
Item16	2.13	-0.01	-0.02	-0.60	-0.61	-0.22	-0.23	-0.25	-0.26
Item17	0.72	0.00	0.01	-0.12	-0.12	-0.03	-0.03	-0.03	-0.02
Item18	2.49	0.00	-0.01	-0.53	-0.54	-0.33	-0.34	-0.37	-0.37
Item19	0.92	-0.01	-0.01	-0.10	-0.10	-0.06	-0.05	-0.05	-0.05
Item20	1.78	-0.01	-0.02	-0.32	-0.32	-0.16	-0.16	-0.18	-0.18
Item21	0.89	0.00	0.00	-0.15	-0.15	-0.04	-0.04	-0.04	-0.04
Item22	0.70	0.00	0.01	-0.12	-0.11	-0.03	-0.03	-0.03	-0.02
Item23	0.86	-0.01	-0.01	-0.12	-0.12	-0.05	-0.05	-0.05	-0.05
Item24	1.40	-0.01	-0.01	-0.18	-0.18	-0.10	-0.10	-0.11	-0.11
Item25	1.28	0.00	-0.01	-0.34	-0.34	-0.12	-0.12	-0.12	-0.12
Item26	0.75	0.00	0.00	-0.11	-0.10	-0.03	-0.03	-0.03	-0.02
Item27	2.08	-0.02	-0.05	-1.07	-1.07	-0.47	-0.49	-0.56	-0.58
Item28	1.06	-0.01	-0.01	-0.21	-0.21	-0.07	-0.07	-0.08	-0.08
Item29	0.58	0.00	0.00	-0.09	-0.08	-0.03	-0.02	-0.02	-0.02
Item30	1.97	-0.01	-0.01	-0.60	-0.60	-0.23	-0.24	-0.25	-0.25
Item31	0.88	-0.01	0.00	-0.09	-0.09	-0.05	-0.04	-0.04	-0.04
Item32	1.20	-0.01	-0.01	-0.23	-0.23	-0.08	-0.08	-0.09	-0.09
Item33	0.88	0.00	0.00	-0.11	-0.11	-0.05	-0.05	-0.05	-0.04
Item34	1.66	-0.02	-0.03	-0.82	-0.82	-0.40	-0.41	-0.43	-0.44
Item35	0.71	-0.01	0.00	-0.10	-0.09	-0.04	-0.03	-0.03	-0.03
Item36	1.93	-0.01	-0.02	-0.36	-0.37	-0.17	-0.18	-0.18	-0.19
Item37	1.56	-0.01	-0.01	-0.36	-0.36	-0.14	-0.15	-0.15	-0.16
Item38	0.81	-0.01	-0.01	-0.12	-0.12	-0.05	-0.05	-0.05	-0.04
Item39	0.88	0.00	0.00	-0.09	-0.09	-0.04	-0.04	-0.04	-0.04
Item40	1.97	-0.01	-0.04	-1.02	-1.02	-0.44	-0.46	-0.53	-0.55
Item41	1.76	-0.02	-0.03	-0.45	-0.45	-0.17	-0.18	-0.19	-0.20
Item42	0.95	-0.01	-0.01	-0.14	-0.14	-0.05	-0.05	-0.05	-0.05
Item43	0.94	-0.01	-0.01	-0.10	-0.10	-0.06	-0.06	-0.06	-0.05
Item44	1.40	-0.01	-0.01	-0.20	-0.20	-0.10	-0.10	-0.11	-0.11
Item45	1.28	-0.02	-0.02	-0.19	-0.19	-0.09	-0.10	-0.10	-0.10
Item46	2.35	-0.01	-0.07	-1.41	-1.41	-0.71	-0.73	-0.82	-0.85
Item47	0.66	-0.01	0.00	-0.07	-0.07	-0.04	-0.03	-0.03	-0.02
Item48	1.09	0.00	0.00	-0.22	-0.22	-0.08	-0.08	-0.08	-0.08
Item49	1.73	-0.01	-0.02	-0.46	-0.46	-0.16	-0.17	-0.18	-0.19
Item50	1.24	0.00	0.00	-0.23	-0.23	-0.08	-0.09	-0.09	-0.09
Item51	1.32	0.00	-0.01	-0.32	-0.32	-0.10	-0.10	-0.12	-0.12
Item52	1.27	-0.01	-0.01	-0.20	-0.20	-0.08	-0.08	-0.08	-0.09
Item53	1.42	0.00	0.00	-0.30	-0.30	-0.10	-0.10	-0.11	-0.11
Item54	2.14	0.00	-0.01	-0.57	-0.57	-0.22	-0.23	-0.24	-0.25
Item55	2.01	-0.01	-0.02	-0.42	-0.43	-0.18	-0.19	-0.19	-0.20
Item56	1.67	-0.01	-0.01	-0.23	-0.23	-0.14	-0.15	-0.16	-0.16
Item57	2.04	-0.02	-0.03	-0.52	-0.53	-0.20	-0.22	-0.22	-0.23

Table A**51**
Bias of estimated item parameters for manipulation of the 25% of items and 20% of sample (60 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4460	4462	4403	4403
	Item58	1.61	-0.02	-0.02	-0.23	-0.23	-0.15	-0.15	-0.16	-0.16
	Item59	2.28	0.01	-0.04	-1.52	-1.52	-0.92	-0.94	-0.97	-0.98
	Item60	1.30	-0.01	-0.01	-0.17	-0.17	-0.09	-0.09	-0.10	-0.10
MAD	NA	NA	0.01	0.01	0.35	0.35	0.16	0.16	0.17	0.17

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**52**
Bias of estimated item parameters for manipulation of the 25% of items and 30% of sample (60 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4415	4418	4352	4352
Item difficulty	Item1	-0.39	0.00	0.00	-0.02	-0.02	-0.06	-0.06	-0.07	-0.07
	Item2	-1.14	0.00	0.00	-0.10	-0.08	-0.06	-0.04	-0.09	-0.08
	Item3	2.31	0.02	0.04	0.77	0.77	0.34	0.36	0.40	0.41
	Item4	2.04	0.01	0.02	0.38	0.38	0.19	0.20	0.22	0.23
	Item5	1.48	0.02	0.02	0.14	0.14	0.11	0.12	0.12	0.13
	Item6	-0.55	0.00	0.00	-0.03	-0.02	-0.01	0.00	-0.04	-0.03
	Item7	-2.15	-0.02	0.00	-0.33	-0.29	-0.15	-0.12	-0.22	-0.19
	Item8	-1.51	-0.01	-0.01	-0.31	-0.31	-0.19	-0.19	-0.24	-0.25
	Item9	-0.37	0.00	0.00	-0.01	-0.01	-0.05	-0.05	-0.06	-0.06
	Item10	2.62	0.01	0.01	0.30	0.28	0.16	0.16	0.16	0.16
	Item11	2.55	0.04	0.05	0.36	0.35	0.21	0.21	0.21	0.21
	Item12	-1.54	-0.02	-0.02	-0.22	-0.21	-0.14	-0.14	-0.20	-0.19
	Item13	-1.50	-0.01	0.00	-0.16	-0.14	-0.09	-0.08	-0.14	-0.13
	Item14	-0.20	-0.01	-0.01	0.00	0.00	-0.04	-0.04	-0.04	-0.05
	Item15	-1.72	-0.01	0.00	-0.23	-0.21	-0.13	-0.11	-0.19	-0.17
	Item16	0.36	0.01	0.01	0.03	0.02	0.02	0.02	0.02	0.02
	Item17	0.89	0.01	0.01	0.03	0.03	0.05	0.05	0.06	0.06
	Item18	0.17	0.01	0.01	0.02	0.02	0.02	0.02	0.01	0.01
	Item19	-0.41	0.00	0.00	-0.02	-0.02	-0.04	-0.04	-0.06	-0.06
	Item20	2.42	0.03	0.01	0.19	0.16	0.16	0.13	0.13	0.11
	Item21	-2.06	-0.01	-0.02	-0.88	-0.87	-0.45	-0.46	-0.56	-0.56
	Item22	1.55	0.01	0.00	0.08	0.08	0.10	0.09	0.10	0.09
	Item23	-1.81	-0.01	-0.01	-0.49	-0.49	-0.26	-0.27	-0.34	-0.34
	Item24	1.31	0.01	0.02	0.08	0.08	0.09	0.09	0.10	0.10
	Item25	0.99	0.01	0.01	0.04	0.04	0.06	0.06	0.07	0.07
	Item26	-1.34	-0.01	-0.01	-0.21	-0.21	-0.14	-0.14	-0.19	-0.19
	Item27	-0.52	0.00	0.00	-0.04	-0.04	-0.04	-0.04	-0.06	-0.06
	Item28	2.69	0.02	0.03	0.73	0.72	0.30	0.31	0.34	0.35
	Item29	1.59	0.02	0.02	0.16	0.16	0.12	0.13	0.14	0.14
	Item30	1.08	0.02	0.02	0.05	0.05	0.07	0.07	0.08	0.08
	Item31	0.24	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.01
	Item32	-0.38	0.00	0.00	-0.01	-0.01	-0.04	-0.04	-0.05	-0.05
	Item33	-1.38	-0.02	-0.01	-0.16	-0.13	-0.09	-0.07	-0.13	-0.11

Table A**52**
Bias of estimated item parameters for manipulation of the 25% of items and 30% of sample (60 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4415	4418	4352	4352
	Item34	-1.85	-0.01	-0.02	-0.58	-0.57	-0.31	-0.32	-0.39	-0.40
	Item35	0.47	0.00	0.00	0.03	0.02	0.03	0.03	0.02	0.02
	Item36	-0.64	-0.01	-0.01	-0.05	-0.05	-0.06	-0.06	-0.08	-0.08
	Item37	1.75	0.02	0.02	0.12	0.11	0.12	0.11	0.12	0.11
	Item38	2.37	0.04	0.05	1.11	1.10	0.48	0.50	0.56	0.58
	Item39	1.11	0.01	0.01	0.05	0.05	0.07	0.07	0.08	0.08
	Item40	-0.90	-0.01	-0.01	-0.09	-0.09	-0.09	-0.09	-0.12	-0.12
	Item41	0.94	0.00	0.00	0.04	0.04	0.05	0.05	0.06	0.06
	Item42	2.41	0.01	0.01	0.21	0.20	0.13	0.12	0.13	0.12
	Item43	-0.34	0.00	0.00	-0.02	-0.02	-0.05	-0.05	-0.06	-0.06
	Item44	-0.24	-0.01	-0.01	-0.01	-0.01	-0.04	-0.04	-0.05	-0.05
	Item45	-0.91	0.00	0.00	-0.08	-0.08	-0.07	-0.07	-0.11	-0.10
	Item46	-0.95	-0.01	-0.01	-0.10	-0.10	-0.10	-0.10	-0.13	-0.13
	Item47	0.99	0.01	0.01	0.03	0.03	0.06	0.06	0.07	0.07
	Item48	0.08	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01
	Item49	2.35	0.03	0.00	0.14	0.09	0.12	0.07	0.07	0.02
	Item50	-1.34	-0.01	0.00	-0.14	-0.13	-0.09	-0.08	-0.13	-0.12
	Item51	0.56	0.00	0.00	0.03	0.03	0.04	0.03	0.03	0.03
	Item52	1.70	0.01	0.01	0.22	0.22	0.14	0.14	0.15	0.16
	Item53	1.54	0.01	0.00	0.08	0.06	0.08	0.06	0.07	0.05
	Item54	0.37	0.00	0.00	0.02	0.02	0.01	0.01	0.02	0.02
	Item55	-1.33	-0.01	-0.01	-0.24	-0.24	-0.16	-0.16	-0.21	-0.21
	Item56	1.06	0.02	0.02	0.06	0.05	0.07	0.07	0.07	0.07
	Item57	0.80	0.00	0.00	0.03	0.03	0.05	0.05	0.06	0.06
	Item58	1.46	0.02	0.03	0.12	0.13	0.11	0.12	0.12	0.13
	Item59	2.04	0.02	0.03	0.42	0.42	0.22	0.23	0.24	0.25
	Item60	1.74	0.02	0.02	0.17	0.17	0.13	0.13	0.14	0.14
MAD	NA	NA	0.01	0.01	0.18	0.17	0.11	0.11	0.14	0.13
Item discrimination	Item1	2.34	-0.02	-0.02	-0.76	-0.76	-0.52	-0.53	-0.57	-0.58
	Item2	0.58	0.00	0.00	-0.10	-0.09	-0.06	-0.05	-0.05	-0.05
	Item3	2.15	-0.01	-0.04	-1.20	-1.20	-0.70	-0.71	-0.76	-0.77
	Item4	1.90	0.00	-0.02	-0.82	-0.82	-0.43	-0.44	-0.46	-0.47
	Item5	2.28	-0.03	-0.05	-0.86	-0.86	-0.50	-0.51	-0.52	-0.53
	Item6	0.52	0.00	0.00	-0.08	-0.07	-0.05	-0.04	-0.04	-0.03
	Item7	0.65	0.00	0.00	-0.16	-0.15	-0.08	-0.07	-0.08	-0.08
	Item8	1.66	0.00	-0.01	-0.64	-0.64	-0.38	-0.38	-0.40	-0.40
	Item9	2.41	-0.01	-0.02	-0.80	-0.80	-0.55	-0.56	-0.60	-0.61
	Item10	0.89	0.00	0.00	-0.22	-0.22	-0.10	-0.10	-0.10	-0.10
	Item11	0.98	-0.02	-0.02	-0.28	-0.27	-0.14	-0.14	-0.14	-0.14
	Item12	0.88	-0.01	-0.01	-0.22	-0.21	-0.13	-0.12	-0.13	-0.13
	Item13	0.65	0.00	0.00	-0.13	-0.12	-0.07	-0.07	-0.07	-0.07
	Item14	2.07	-0.03	-0.03	-0.58	-0.58	-0.42	-0.43	-0.46	-0.46
	Item15	0.69	0.00	0.00	-0.15	-0.14	-0.08	-0.08	-0.08	-0.08
	Item16	0.85	-0.01	-0.01	-0.13	-0.13	-0.09	-0.09	-0.09	-0.09
	Item17	1.36	0.00	-0.01	-0.25	-0.25	-0.18	-0.18	-0.18	-0.18

Table A**52**
Bias of estimated item parameters for manipulation of the 25% of items and 30% of sample (60 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4415	4418	4352	4352
	Item18	0.77	-0.01	-0.01	-0.12	-0.12	-0.08	-0.08	-0.08	-0.08
	Item19	1.33	0.00	0.00	-0.28	-0.28	-0.19	-0.19	-0.20	-0.20
	Item20	0.67	0.00	0.00	-0.12	-0.12	-0.06	-0.06	-0.05	-0.05
	Item21	1.95	0.00	-0.01	-1.09	-1.09	-0.75	-0.76	-0.78	-0.79
	Item22	0.83	0.00	0.00	-0.14	-0.14	-0.08	-0.08	-0.08	-0.08
	Item23	1.61	0.00	-0.01	-0.70	-0.70	-0.43	-0.43	-0.45	-0.46
	Item24	1.46	-0.02	-0.02	-0.34	-0.34	-0.22	-0.22	-0.22	-0.22
	Item25	1.78	-0.01	-0.02	-0.41	-0.42	-0.28	-0.29	-0.28	-0.29
	Item26	1.46	0.00	0.00	-0.47	-0.47	-0.27	-0.27	-0.29	-0.29
	Item27	0.79	-0.01	-0.01	-0.14	-0.14	-0.09	-0.09	-0.09	-0.09
	Item28	1.40	0.00	-0.02	-0.62	-0.62	-0.31	-0.32	-0.34	-0.34
	Item29	1.99	-0.02	-0.03	-0.70	-0.70	-0.39	-0.40	-0.41	-0.42
	Item30	1.85	-0.02	-0.03	-0.46	-0.46	-0.31	-0.31	-0.31	-0.32
	Item31	0.98	-0.01	-0.01	-0.16	-0.16	-0.12	-0.12	-0.12	-0.12
	Item32	1.34	-0.01	-0.01	-0.28	-0.28	-0.19	-0.19	-0.21	-0.21
	Item33	0.57	-0.01	0.00	-0.11	-0.10	-0.06	-0.05	-0.06	-0.05
	Item34	1.77	-0.01	-0.02	-0.84	-0.84	-0.54	-0.55	-0.57	-0.57
	Item35	0.82	0.00	0.00	-0.12	-0.12	-0.08	-0.08	-0.08	-0.08
	Item36	1.22	0.00	0.00	-0.26	-0.26	-0.16	-0.16	-0.17	-0.17
	Item37	0.74	-0.01	-0.01	-0.12	-0.12	-0.07	-0.07	-0.07	-0.06
	Item38	2.50	-0.04	-0.10	-1.62	-1.62	-1.04	-1.06	-1.14	-1.15
	Item39	1.92	-0.01	-0.01	-0.51	-0.51	-0.33	-0.34	-0.33	-0.34
	Item40	1.28	-0.01	-0.01	-0.32	-0.32	-0.19	-0.19	-0.20	-0.20
	Item41	0.97	0.00	0.00	-0.16	-0.15	-0.11	-0.11	-0.10	-0.10
	Item42	0.87	0.00	0.00	-0.19	-0.19	-0.09	-0.09	-0.09	-0.09
	Item43	1.68	-0.01	-0.01	-0.41	-0.41	-0.28	-0.29	-0.31	-0.31
	Item44	1.90	-0.01	-0.02	-0.50	-0.50	-0.36	-0.36	-0.39	-0.39
	Item45	0.87	0.00	0.00	-0.17	-0.17	-0.11	-0.10	-0.11	-0.11
	Item46	1.28	0.00	0.00	-0.33	-0.33	-0.20	-0.20	-0.21	-0.21
	Item47	2.17	-0.01	-0.02	-0.59	-0.59	-0.40	-0.40	-0.40	-0.41
	Item48	0.84	-0.01	-0.01	-0.13	-0.13	-0.09	-0.09	-0.09	-0.09
	Item49	0.52	0.00	0.01	-0.07	-0.06	-0.03	-0.02	-0.02	-0.01
	Item50	0.76	-0.01	0.00	-0.16	-0.15	-0.09	-0.08	-0.09	-0.08
	Item51	0.73	0.00	0.00	-0.11	-0.10	-0.07	-0.07	-0.07	-0.06
	Item52	2.34	0.00	-0.02	-1.00	-1.01	-0.56	-0.57	-0.59	-0.60
	Item53	0.62	0.00	0.00	-0.09	-0.08	-0.05	-0.04	-0.04	-0.04
	Item54	1.53	0.00	-0.01	-0.29	-0.29	-0.22	-0.22	-0.23	-0.23
	Item55	1.91	0.00	0.00	-0.76	-0.76	-0.45	-0.46	-0.48	-0.48
	Item56	0.79	0.00	0.00	-0.12	-0.12	-0.08	-0.08	-0.07	-0.07
	Item57	1.18	0.00	-0.01	-0.20	-0.20	-0.14	-0.14	-0.15	-0.15
	Item58	1.93	-0.02	-0.03	-0.62	-0.62	-0.36	-0.37	-0.37	-0.38
	Item59	1.99	-0.01	-0.03	-0.91	-0.91	-0.50	-0.51	-0.54	-0.55
	Item60	1.50	0.00	-0.01	-0.43	-0.44	-0.23	-0.24	-0.24	-0.24
MAD	NA	NA	0.01	0.01	0.41	0.41	0.25	0.25	0.26	0.27

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting

Table A**52**
Bias of estimated item parameters for manipulation of the 25% of items and 30% of sample (60 items)

	True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
					N =	4415	4418	4352	4352

data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**53**
Bias of estimated item parameters for manipulation of the 50% of items and 10% of sample (60 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4550	4550	4509	4509
Item difficulty	Item1	-1.26	-0.01	-0.01	-0.02	-0.02	0.00	0.00	-0.04	-0.04
	Item2	0.42	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
	Item3	2.29	0.00	0.00	-0.27	-0.27	-0.06	-0.06	-0.04	-0.04
	Item4	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Item5	-0.36	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00
	Item6	-0.36	-0.01	-0.01	-0.02	-0.02	-0.01	-0.01	-0.02	-0.02
	Item7	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Item8	0.68	0.01	0.01	-0.03	-0.03	0.01	0.01	0.01	0.01
	Item9	-0.09	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00
	Item10	0.45	0.00	0.00	0.03	0.03	0.01	0.01	0.02	0.02
	Item11	-2.07	-0.02	-0.03	0.06	0.06	0.03	0.03	-0.03	-0.03
	Item12	-0.18	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	Item13	1.81	0.01	0.01	-0.15	-0.15	-0.02	-0.02	-0.01	-0.01
	Item14	-1.75	-0.01	-0.01	0.09	0.09	0.03	0.03	0.00	0.00
	Item15	-0.38	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02
	Item16	-2.20	-0.01	0.00	0.13	0.13	0.05	0.05	-0.01	0.00
	Item17	0.75	0.01	0.01	-0.02	-0.02	0.00	0.00	0.01	0.00
	Item18	2.48	0.02	0.03	-0.28	-0.27	-0.06	-0.05	-0.01	0.00
	Item19	1.46	0.01	0.01	-0.04	-0.04	0.00	0.00	0.02	0.02
	Item20	2.15	0.02	0.03	-0.17	-0.16	-0.04	-0.02	0.01	0.02
	Item21	0.15	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
	Item22	-2.02	-0.02	-0.01	0.17	0.18	0.04	0.06	0.02	0.03
	Item23	2.19	0.03	0.04	-0.16	-0.14	-0.03	-0.01	0.02	0.03
	Item24	2.06	0.02	0.03	-0.14	-0.13	-0.03	-0.02	0.01	0.02
	Item25	-1.90	-0.01	-0.01	0.03	0.03	0.03	0.03	-0.02	-0.03
	Item26	-1.39	-0.01	-0.02	-0.03	-0.03	0.00	0.00	-0.04	-0.04
	Item27	0.45	0.00	0.00	-0.01	-0.02	0.00	0.00	0.00	0.00
	Item28	0.14	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00
	Item29	2.45	0.03	0.04	-0.25	-0.24	-0.05	-0.03	0.00	0.02
	Item30	2.34	0.02	0.02	-0.29	-0.30	-0.05	-0.06	-0.05	-0.05
	Item31	-0.42	0.00	0.00	0.03	0.03	0.01	0.01	0.00	0.00
	Item32	1.15	0.00	0.01	0.00	0.00	0.00	0.01	0.02	0.02
	Item33	-2.49	-0.02	-0.03	-0.08	-0.09	0.04	0.03	-0.06	-0.07
	Item34	1.51	0.01	0.02	-0.05	-0.05	0.00	0.00	0.02	0.02
	Item35	-2.54	-0.02	-0.04	-0.06	-0.07	0.04	0.03	-0.06	-0.08
	Item36	-0.33	0.00	0.00	0.00	0.00	0.00	0.00	-0.02	-0.02
	Item37	-2.36	-0.02	-0.02	0.06	0.05	0.04	0.04	-0.04	-0.04
	Item38	-0.34	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01
	Item39	-0.87	-0.01	-0.01	-0.02	-0.02	-0.01	-0.01	-0.03	-0.03

Table A**53**
Bias of estimated item parameters for manipulation of the 50% of items and 10% of sample (60 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4550	4550	4509	4509
	Item40	1.96	0.02	0.02	-0.18	-0.18	-0.02	-0.02	-0.02	-0.02
	Item41	1.67	0.01	0.01	-0.15	-0.16	-0.02	-0.03	-0.02	-0.02
	Item42	0.47	0.00	0.00	0.03	0.03	0.01	0.01	0.02	0.02
	Item43	-1.47	-0.02	-0.02	-0.03	-0.04	0.00	0.00	-0.04	-0.04
	Item44	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
	Item45	0.07	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00
	Item46	-2.35	-0.01	-0.02	0.04	0.04	0.05	0.04	-0.03	-0.04
	Item47	-2.62	-0.05	-0.02	0.26	0.28	0.06	0.08	0.02	0.04
	Item48	0.36	0.01	0.01	0.04	0.04	0.02	0.02	0.02	0.02
	Item49	2.15	0.01	0.02	-0.20	-0.19	-0.04	-0.03	0.00	0.00
	Item50	1.28	0.00	0.00	-0.09	-0.09	-0.02	-0.02	-0.01	-0.01
	Item51	-1.03	0.00	0.00	-0.03	-0.03	-0.01	-0.01	-0.03	-0.03
	Item52	-0.32	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01
	Item53	-2.13	-0.01	0.00	0.20	0.21	0.06	0.08	0.04	0.05
	Item54	-1.36	-0.01	-0.01	-0.01	-0.01	0.00	0.00	-0.03	-0.04
	Item55	2.13	0.04	0.01	-0.32	-0.33	-0.04	-0.07	-0.08	-0.10
	Item56	1.36	0.00	0.01	-0.03	-0.03	0.00	0.00	0.02	0.02
	Item57	2.61	0.02	0.04	-0.28	-0.26	-0.08	-0.05	-0.01	0.01
	Item58	1.75	0.01	0.02	-0.09	-0.09	-0.02	-0.01	0.02	0.02
	Item59	-1.31	-0.02	-0.01	0.12	0.12	0.03	0.04	0.02	0.03
	Item60	2.58	0.04	0.06	-0.20	-0.18	-0.04	-0.01	0.03	0.06
MAD	NA	NA	0.01	0.01	0.08	0.08	0.02	0.02	0.02	0.02
Item discrimination	Item1	1.74	-0.02	-0.02	-0.24	-0.24	0.00	-0.01	-0.04	-0.04
	Item2	1.26	-0.01	-0.01	-0.10	-0.10	-0.01	-0.01	-0.02	-0.02
	Item3	0.86	0.00	0.00	0.06	0.06	0.03	0.03	0.03	0.03
	Item4	1.37	-0.01	-0.01	-0.14	-0.14	-0.02	-0.02	-0.03	-0.03
	Item5	0.80	-0.01	0.00	-0.01	-0.01	0.00	0.01	0.01	0.01
	Item6	1.71	-0.02	-0.02	-0.26	-0.26	-0.04	-0.04	-0.07	-0.07
	Item7	0.99	-0.01	-0.01	-0.05	-0.05	0.00	0.00	0.00	0.00
	Item8	0.76	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
	Item9	1.33	-0.01	-0.01	-0.13	-0.13	-0.01	-0.02	-0.02	-0.03
	Item10	2.48	0.00	-0.01	-0.55	-0.55	-0.08	-0.09	-0.14	-0.15
	Item11	1.16	-0.01	-0.01	-0.07	-0.07	0.03	0.02	0.00	0.00
	Item12	2.49	0.00	-0.01	-0.60	-0.60	-0.09	-0.10	-0.15	-0.16
	Item13	0.89	0.00	0.00	0.04	0.04	0.02	0.02	0.03	0.03
	Item14	0.84	0.00	0.00	-0.01	-0.01	0.02	0.02	0.01	0.01
	Item15	1.67	0.00	-0.01	-0.23	-0.23	-0.02	-0.02	-0.05	-0.05
	Item16	0.94	0.00	0.00	-0.02	-0.01	0.03	0.03	0.01	0.01
	Item17	0.94	-0.01	-0.01	-0.02	-0.02	0.00	0.00	0.00	0.00
	Item18	1.38	0.00	-0.01	0.03	0.02	0.05	0.04	0.02	0.01
	Item19	2.18	0.00	-0.02	-0.21	-0.22	0.02	0.01	-0.02	-0.04
	Item20	2.05	-0.03	-0.05	-0.18	-0.20	0.04	0.02	-0.03	-0.05
	Item21	1.54	-0.01	-0.01	-0.19	-0.19	-0.02	-0.03	-0.05	-0.05
	Item22	0.62	0.00	0.00	0.02	0.03	0.01	0.02	0.02	0.02
	Item23	2.29	-0.02	-0.06	-0.31	-0.33	0.05	0.01	-0.04	-0.07

Table A**53**

Bias of estimated item parameters for manipulation of the 50% of items and 10% of sample (60 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4550	4550	4509	4509
	Item24	2.45	-0.01	-0.05	-0.35	-0.38	0.06	0.02	-0.03	-0.07
	Item25	1.49	0.00	0.00	-0.15	-0.16	0.04	0.04	0.00	-0.01
	Item26	2.31	-0.01	-0.02	-0.45	-0.45	0.02	0.01	-0.04	-0.06
	Item27	0.82	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.02
	Item28	0.59	-0.01	0.00	0.02	0.02	0.01	0.01	0.02	0.02
	Item29	1.56	-0.02	-0.03	-0.03	-0.05	0.04	0.03	0.00	-0.02
	Item30	0.75	0.00	0.00	0.06	0.07	0.02	0.03	0.03	0.03
	Item31	0.71	-0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.01
	Item32	1.96	-0.01	-0.01	-0.21	-0.21	0.00	0.00	-0.04	-0.04
	Item33	2.21	0.00	-0.04	-0.72	-0.73	-0.02	-0.06	-0.15	-0.18
	Item34	1.63	-0.01	-0.02	-0.09	-0.09	0.02	0.01	-0.01	-0.02
	Item35	2.09	0.00	-0.04	-0.63	-0.64	0.00	-0.04	-0.12	-0.16
	Item36	1.31	-0.01	-0.01	-0.13	-0.13	-0.01	-0.01	-0.03	-0.03
	Item37	1.45	-0.01	-0.01	-0.18	-0.18	0.03	0.02	-0.02	-0.03
	Item38	1.55	-0.02	-0.02	-0.20	-0.20	-0.03	-0.03	-0.05	-0.05
	Item39	1.48	-0.01	-0.02	-0.17	-0.17	-0.01	-0.02	-0.04	-0.04
	Item40	0.84	0.00	0.00	0.04	0.04	0.02	0.02	0.02	0.02
	Item41	0.74	0.00	0.00	0.04	0.04	0.02	0.02	0.02	0.03
	Item42	2.47	-0.02	-0.03	-0.56	-0.56	-0.10	-0.10	-0.16	-0.17
	Item43	2.43	-0.03	-0.05	-0.51	-0.52	0.02	0.00	-0.06	-0.07
	Item44	1.05	0.00	0.00	-0.05	-0.05	0.00	0.00	0.00	0.00
	Item45	1.91	-0.02	-0.02	-0.33	-0.33	-0.06	-0.06	-0.09	-0.10
	Item46	1.60	0.00	-0.01	-0.24	-0.24	0.04	0.03	-0.02	-0.04
	Item47	0.61	-0.01	0.00	0.03	0.03	0.01	0.02	0.01	0.02
	Item48	2.49	-0.02	-0.02	-0.58	-0.58	-0.10	-0.11	-0.16	-0.17
	Item49	1.45	0.00	-0.01	0.00	-0.01	0.05	0.04	0.01	0.00
	Item50	0.82	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.02
	Item51	2.27	0.00	-0.01	-0.44	-0.44	-0.01	-0.01	-0.06	-0.07
	Item52	1.19	-0.01	-0.01	-0.10	-0.10	-0.01	-0.01	-0.02	-0.02
	Item53	0.64	0.00	0.00	0.03	0.03	0.02	0.03	0.02	0.03
	Item54	1.68	-0.01	-0.01	-0.21	-0.21	0.01	0.01	-0.02	-0.03
	Item55	0.50	-0.01	0.00	0.06	0.07	0.02	0.02	0.03	0.04
	Item56	1.75	0.01	0.00	-0.12	-0.13	0.03	0.02	0.00	0.00
	Item57	1.95	0.00	-0.03	-0.21	-0.23	0.08	0.04	-0.02	-0.05
	Item58	2.14	0.00	-0.02	-0.17	-0.18	0.05	0.03	0.00	-0.02
	Item59	0.56	-0.01	0.00	0.03	0.03	0.01	0.02	0.02	0.02
	Item60	2.33	-0.03	-0.09	-0.52	-0.55	0.01	-0.05	-0.13	-0.18
MAD	NA	NA	0.01	0.02	0.19	0.19	0.03	0.03	0.04	0.05

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**54**
Bias of estimated item parameters for manipulation of the 50% of items and 20% of sample (60 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4276	4277	4222	4223
Item difficulty	Item1	0.29	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.02
	Item2	-1.98	-0.03	-0.04	0.37	0.36	0.19	0.17	0.09	0.08
	Item3	0.00	0.00	0.00	-0.01	-0.01	0.00	0.00	-0.01	-0.01
	Item4	1.93	0.01	0.01	-0.12	-0.12	-0.09	-0.09	-0.07	-0.07
	Item5	-0.37	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00
	Item6	0.32	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00
	Item7	1.49	0.01	0.01	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01
	Item8	-2.21	-0.02	-0.04	0.50	0.49	0.27	0.25	0.15	0.13
	Item9	1.08	0.01	0.01	0.05	0.05	0.02	0.02	0.02	0.02
	Item10	-2.62	-0.03	-0.06	0.79	0.77	0.44	0.40	0.26	0.23
	Item11	2.55	0.02	0.03	-0.25	-0.25	-0.20	-0.20	-0.15	-0.14
	Item12	0.10	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
	Item13	-1.92	-0.02	-0.02	0.37	0.37	0.17	0.17	0.10	0.10
	Item14	-0.26	0.00	0.00	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02
	Item15	1.85	0.02	0.02	-0.01	-0.01	-0.06	-0.06	-0.04	-0.03
	Item16	-0.72	-0.01	-0.01	-0.03	-0.03	-0.02	-0.02	-0.03	-0.03
	Item17	2.21	0.03	0.02	-0.28	-0.28	-0.14	-0.14	-0.12	-0.13
	Item18	-0.14	0.00	0.00	-0.04	-0.03	-0.02	-0.02	-0.02	-0.02
	Item19	0.42	0.00	0.00	-0.06	-0.07	-0.03	-0.03	-0.04	-0.04
	Item20	-1.60	-0.02	-0.03	0.15	0.14	0.07	0.06	0.01	0.00
	Item21	2.19	0.02	0.01	-0.39	-0.40	-0.19	-0.20	-0.18	-0.19
	Item22	-1.73	-0.02	-0.02	0.22	0.22	0.11	0.10	0.04	0.03
	Item23	2.19	0.02	0.03	-0.14	-0.15	-0.12	-0.12	-0.08	-0.08
	Item24	-2.53	-0.03	-0.05	0.72	0.70	0.40	0.37	0.24	0.21
	Item25	1.44	0.02	0.01	-0.05	-0.05	-0.02	-0.03	-0.02	-0.02
	Item26	-0.77	-0.01	-0.01	-0.07	-0.07	-0.04	-0.04	-0.05	-0.05
	Item27	0.52	0.00	0.00	-0.08	-0.08	-0.03	-0.03	-0.04	-0.04
	Item28	-0.55	-0.01	-0.01	-0.06	-0.06	-0.04	-0.04	-0.04	-0.04
	Item29	0.46	0.01	0.01	-0.06	-0.06	-0.02	-0.02	-0.03	-0.03
	Item30	2.46	0.02	0.04	-0.05	-0.05	-0.13	-0.12	-0.07	-0.05
	Item31	-2.31	-0.03	-0.04	0.56	0.55	0.28	0.26	0.16	0.14
	Item32	1.23	0.01	0.01	0.04	0.04	0.01	0.01	0.01	0.01
	Item33	-1.56	-0.02	-0.02	0.15	0.14	0.06	0.05	0.01	0.00
	Item34	-2.52	-0.02	-0.05	0.72	0.70	0.39	0.36	0.23	0.20
	Item35	-2.64	-0.02	-0.04	0.80	0.79	0.44	0.42	0.26	0.24
	Item36	-2.32	-0.03	-0.03	0.55	0.54	0.26	0.25	0.15	0.14
	Item37	-0.01	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00
	Item38	-2.14	-0.02	-0.02	0.45	0.44	0.21	0.20	0.12	0.11
	Item39	0.91	0.01	0.01	0.02	0.02	0.01	0.01	0.01	0.01
	Item40	2.54	0.04	0.03	-0.37	-0.38	-0.18	-0.20	-0.15	-0.16
	Item41	-1.14	-0.02	-0.02	0.09	0.09	0.03	0.03	0.01	0.01
	Item42	2.48	0.01	0.03	-0.09	-0.09	-0.15	-0.14	-0.09	-0.07
	Item43	-1.36	-0.01	-0.01	0.15	0.15	0.06	0.06	0.02	0.02
	Item44	-2.15	-0.03	-0.04	0.46	0.45	0.25	0.23	0.13	0.12
	Item45	0.60	0.01	0.01	0.08	0.08	0.04	0.04	0.04	0.04

Table A**54**
Bias of estimated item parameters for manipulation of the 50% of items and 20% of sample (60 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4276	4277	4222	4223
	Item46	0.18	0.01	0.01	-0.01	-0.01	0.00	0.00	0.00	0.00
	Item47	2.29	0.02	0.03	-0.07	-0.07	-0.14	-0.13	-0.09	-0.08
	Item48	1.00	0.01	0.01	0.08	0.08	0.03	0.03	0.03	0.03
	Item49	1.67	0.03	0.02	-0.26	-0.27	-0.11	-0.12	-0.11	-0.12
	Item50	0.71	0.00	0.00	0.06	0.06	0.03	0.03	0.03	0.03
	Item51	-2.33	-0.03	-0.03	0.57	0.57	0.26	0.26	0.17	0.17
	Item52	0.49	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.01
	Item53	-0.51	-0.01	-0.01	-0.06	-0.06	-0.04	-0.04	-0.04	-0.04
	Item54	-2.33	-0.03	-0.04	0.59	0.57	0.32	0.29	0.18	0.16
	Item55	0.82	0.00	0.00	0.06	0.06	0.03	0.02	0.03	0.02
	Item56	1.18	0.01	0.01	0.08	0.08	0.02	0.02	0.02	0.02
	Item57	0.39	0.00	0.00	0.03	0.03	0.02	0.02	0.02	0.02
	Item58	-1.45	-0.01	-0.02	0.09	0.09	0.04	0.03	-0.01	-0.01
	Item59	-1.91	-0.02	-0.03	0.32	0.31	0.15	0.14	0.07	0.06
	Item60	1.55	0.01	0.01	-0.02	-0.03	-0.02	-0.03	-0.02	-0.02
MAD	NA	NA	0.01	0.02	0.20	0.19	0.11	0.10	0.07	0.07
Item discrimination	Item1	1.79	-0.01	-0.01	-0.55	-0.55	-0.19	-0.19	-0.17	-0.17
	Item2	2.19	-0.01	-0.04	-0.01	-0.04	0.30	0.25	0.13	0.09
	Item3	0.99	-0.01	-0.01	-0.11	-0.10	-0.01	-0.01	0.00	0.00
	Item4	1.19	0.00	-0.01	-0.14	-0.14	0.04	0.04	0.02	0.02
	Item5	0.89	0.00	0.00	-0.04	-0.04	0.01	0.01	0.01	0.02
	Item6	1.02	0.00	0.00	-0.12	-0.12	-0.02	-0.02	-0.01	-0.01
	Item7	1.29	-0.01	-0.01	-0.22	-0.21	0.00	0.00	-0.01	-0.01
	Item8	2.40	0.00	-0.05	-0.01	-0.06	0.39	0.31	0.15	0.10
	Item9	1.63	-0.02	-0.02	-0.41	-0.41	-0.07	-0.07	-0.07	-0.07
	Item10	2.02	0.01	-0.04	0.24	0.19	0.41	0.33	0.18	0.12
	Item11	1.39	-0.01	-0.01	-0.23	-0.22	0.08	0.08	0.04	0.03
	Item12	2.25	-0.02	-0.03	-0.88	-0.88	-0.35	-0.35	-0.32	-0.32
	Item13	0.93	-0.01	-0.01	0.11	0.10	0.09	0.08	0.06	0.06
	Item14	1.53	-0.02	-0.02	-0.37	-0.37	-0.13	-0.13	-0.12	-0.12
	Item15	1.93	-0.02	-0.03	-0.53	-0.53	0.02	0.01	-0.02	-0.03
	Item16	1.46	-0.01	-0.01	-0.26	-0.26	-0.06	-0.06	-0.06	-0.06
	Item17	0.84	-0.01	-0.01	0.00	0.00	0.04	0.05	0.04	0.04
	Item18	2.10	-0.01	-0.02	-0.75	-0.75	-0.28	-0.28	-0.26	-0.26
	Item19	0.63	-0.01	0.00	0.03	0.03	0.03	0.03	0.03	0.04
	Item20	2.49	-0.03	-0.05	-0.38	-0.40	0.14	0.10	0.00	-0.02
	Item21	0.63	-0.01	0.00	0.05	0.06	0.05	0.05	0.05	0.05
	Item22	2.06	-0.02	-0.03	-0.12	-0.14	0.17	0.14	0.05	0.03
	Item23	1.36	-0.02	-0.02	-0.21	-0.21	0.05	0.04	0.02	0.01
	Item24	2.22	0.00	-0.05	0.12	0.06	0.40	0.32	0.16	0.09
	Item25	1.07	-0.01	-0.01	-0.13	-0.13	0.00	0.00	0.00	0.00
	Item26	2.21	-0.02	-0.02	-0.67	-0.67	-0.20	-0.21	-0.21	-0.21
	Item27	0.62	-0.01	0.00	0.03	0.04	0.03	0.04	0.04	0.04
	Item28	1.95	-0.02	-0.03	-0.58	-0.58	-0.18	-0.19	-0.18	-0.18
	Item29	0.64	0.00	0.00	0.03	0.03	0.03	0.03	0.03	0.04

Table A**54**
Bias of estimated item parameters for manipulation of the 50% of items and 20% of sample (60 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4276	4277	4222	4223
	Item30	2.41	0.00	-0.05	-1.04	-1.04	-0.16	-0.20	-0.28	-0.31
	Item31	1.43	-0.01	-0.02	0.22	0.20	0.22	0.20	0.11	0.10
	Item32	1.62	-0.01	-0.02	-0.39	-0.39	-0.04	-0.05	-0.05	-0.05
	Item33	1.60	-0.02	-0.02	-0.10	-0.11	0.06	0.05	0.01	0.00
	Item34	2.06	0.00	-0.04	0.21	0.16	0.39	0.32	0.17	0.12
	Item35	1.88	0.01	-0.02	0.29	0.24	0.40	0.33	0.18	0.13
	Item36	1.06	-0.01	-0.02	0.18	0.17	0.13	0.12	0.08	0.07
	Item37	1.56	-0.01	-0.01	-0.40	-0.40	-0.13	-0.13	-0.11	-0.11
	Item38	1.21	-0.01	-0.01	0.15	0.14	0.14	0.13	0.08	0.07
	Item39	1.28	-0.01	-0.01	-0.24	-0.24	-0.04	-0.04	-0.04	-0.04
	Item40	0.85	-0.01	-0.01	0.00	0.01	0.05	0.06	0.04	0.04
	Item41	0.98	-0.02	-0.02	-0.02	-0.02	0.02	0.02	0.02	0.02
	Item42	2.26	0.00	-0.04	-0.90	-0.90	-0.09	-0.13	-0.20	-0.24
	Item43	1.02	-0.01	-0.01	0.01	0.00	0.04	0.04	0.03	0.03
	Item44	2.38	-0.01	-0.05	-0.01	-0.06	0.38	0.31	0.16	0.11
	Item45	2.46	-0.02	-0.02	-1.01	-1.01	-0.33	-0.33	-0.30	-0.31
	Item46	0.98	-0.01	-0.01	-0.11	-0.10	-0.02	-0.02	-0.01	-0.01
	Item47	2.14	-0.01	-0.03	-0.75	-0.75	-0.01	-0.03	-0.09	-0.12
	Item48	2.17	-0.02	-0.03	-0.74	-0.74	-0.15	-0.15	-0.15	-0.15
	Item49	0.61	-0.01	0.00	0.04	0.05	0.04	0.05	0.04	0.05
	Item50	1.88	-0.01	-0.01	-0.58	-0.58	-0.15	-0.15	-0.14	-0.14
	Item51	0.83	-0.01	-0.01	0.16	0.16	0.10	0.10	0.07	0.07
	Item52	1.15	-0.01	-0.01	-0.19	-0.19	-0.04	-0.04	-0.03	-0.03
	Item53	1.85	-0.02	-0.03	-0.52	-0.52	-0.18	-0.18	-0.17	-0.17
	Item54	2.42	0.00	-0.05	0.03	-0.02	0.40	0.31	0.15	0.08
	Item55	1.72	0.00	0.00	-0.48	-0.48	-0.11	-0.11	-0.10	-0.10
	Item56	2.13	-0.01	-0.01	-0.68	-0.68	-0.08	-0.09	-0.09	-0.09
	Item57	1.71	-0.01	-0.01	-0.50	-0.50	-0.15	-0.15	-0.13	-0.13
	Item58	2.19	-0.01	-0.02	-0.34	-0.35	0.07	0.04	-0.02	-0.04
	Item59	1.84	-0.02	-0.03	0.01	-0.01	0.19	0.16	0.08	0.06
	Item60	1.25	-0.01	-0.01	-0.20	-0.20	0.00	0.00	0.00	-0.01
MAD	NA	NA	0.01	0.02	0.30	0.30	0.14	0.12	0.09	0.09

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.

Table A**55**
Bias of estimated item parameters for manipulation of the 50% of items and 30% of sample (60 items)

		True	MLfit	BMfit	MLmisfit	BMmisfit	LzML	LzBM	HTML	HTBM
						N =	4226	4225	4105	4097
Item difficulty	Item1	-1.78	-0.02	-0.03	0.31	0.31	0.23	0.23	0.13	0.13
	Item2	-1.14	-0.02	-0.01	0.13	0.13	0.09	0.08	0.04	0.04
	Item3	-1.52	-0.01	-0.01	0.16	0.15	0.12	0.12	0.04	0.04
	Item4	-2.08	-0.01	-0.02	0.49	0.48	0.41	0.40	0.27	0.26
	Item5	0.63	0.01	0.01	0.11	0.11	0.09	0.09	0.09	0.09
	Item6	2.50	0.02	0.04	-0.48	-0.48	-0.48	-0.48	-0.41	-0.39
	Item7	1.77	0.02	0.02	-0.14	-0.13	-0.17	-0.17	-0.15	-0.14
	Item8	1.24	0.01	0.01	-0.14	-0.14	-0.10	-0.10	-0.09	-0.09
	Item9	-1.35	-0.01	0.00	0.35	0.36	0.25	0.26	0.21	0.21
	Item10	0.69	0.01	0.01	0.11	0.11	0.09	0.09	0.08	0.08
	Item11	2.53	0.02	0.04	-0.56	-0.56	-0.53	-0.53	-0.47	-0.46
	Item12	-1.00	-0.01	-0.01	0.00	0.00	-0.01	-0.01	-0.04	-0.04
	Item13	2.62	0.03	0.00	-0.89	-0.90	-0.70	-0.71	-0.67	-0.67
	Item14	1.05	0.01	0.01	0.08	0.08	0.05	0.05	0.05	0.05
	Item15	-1.77	-0.02	-0.02	0.28	0.28	0.24	0.23	0.13	0.12
	Item16	1.56	0.01	0.01	-0.09	-0.09	-0.10	-0.10	-0.09	-0.08
	Item17	-1.81	-0.01	-0.01	0.38	0.38	0.28	0.28	0.19	0.18
	Item18	-0.32	0.00	0.00	0.03	0.03	0.02	0.02	0.01	0.01
	Item19	-1.88	0.00	-0.01	0.36	0.35	0.29	0.29	0.18	0.17
	Item20	1.30	0.01	0.01	0.03	0.03	-0.01	-0.01	0.00	0.00
	Item21	-0.71	-0.01	-0.01	0.00	0.00	0.00	0.00	-0.02	-0.02
	Item22	-2.64	-0.01	-0.03	0.91	0.90	0.76	0.74	0.56	0.55
	Item23	1.09	0.00	0.00	-0.21	-0.21	-0.15	-0.15	-0.15	-0.15
	Item24	1.40	0.00	-0.01	-0.36	-0.36	-0.26	-0.26	-0.26	-0.26
	Item25	-1.18	-0.01	-0.01	0.08	0.08	0.05	0.05	0.01	0.01
	Item26	-0.70	-0.01	0.00	0.12	0.13	0.08	0.09	0.06	0.06
	Item27	0.78	0.01	0.01	0.10	0.10	0.08	0.08	0.07	0.07
	Item28	-1.58	-0.02	-0.02	0.17	0.17	0.14	0.14	0.06	0.05
	Item29	-0.21	0.00	0.01	0.06	0.06	0.05	0.05	0.03	0.03
	Item30	1.50	0.01	0.01	-0.10	-0.10	-0.10	-0.10	-0.09	-0.09
	Item31	-1.31	-0.01	-0.01	0.04	0.04	0.03	0.03	-0.03	-0.03
	Item32	-2.21	0.00	-0.01	0.57	0.57	0.49	0.48	0.33	0.32
	Item33	-1.62	-0.01	-0.01	0.19	0.19	0.16	0.16	0.07	0.07
	Item34	0.03	0.00	0.00	-0.01	-0.01	0.00	0.00	-0.01	-0.01
	Item35	2.37	0.03	0.03	-0.51	-0.51	-0.43	-0.43	-0.38	-0.38
	Item36	-1.44	-0.01	-0.01	0.15	0.15	0.10	0.10	0.04	0.04
	Item37	1.84	0.02	0.03	-0.19	-0.19	-0.20	-0.20	-0.17	-0.17
	Item38	-0.82	0.00	0.00	-0.09	-0.09	-0.08	-0.08	-0.10	-0.10
	Item39	-1.68	-0.01	-0.02	0.23	0.22	0.20	0.19	0.10	0.09
	Item40	1.04	0.01	0.01	0.08	0.08	0.05	0.05	0.05	0.05
	Item41	1.27	0.01	0.01	0.01	0.01	-0.01	-0.01	-0.01	0.00
	Item42	-2.45	-0.03	-0.01	0.83	0.83	0.62	0.62	0.50	0.50
	Item43	-2.48	-0.02	-0.02	0.78	0.77	0.61	0.61	0.45	0.44
	Item44	1.47	0.01	0.02	-0.02	-0.02	-0.07	-0.07	-0.05	-0.05
	Item45	-1.43	-0.01	0.00	0.49	0.49	0.37	0.37	0.32	0.32

Table A**55**
Bias of estimated item parameters for manipulation of the 50% of items and 30% of sample (60 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4226	4225	4105	4097
	Item46	1.57	0.01	0.01	-0.09	-0.09	-0.11	-0.11	-0.09	-0.09
	Item47	-2.57	-0.03	-0.03	0.89	0.89	0.68	0.68	0.54	0.54
	Item48	0.34	0.00	0.00	-0.02	-0.02	0.00	0.00	0.00	0.00
	Item49	-2.48	-0.04	-0.05	0.78	0.77	0.65	0.64	0.47	0.45
	Item50	-1.48	-0.01	0.01	0.49	0.49	0.37	0.37	0.31	0.32
	Item51	-2.60	-0.02	-0.04	0.87	0.86	0.73	0.72	0.53	0.52
	Item52	-1.47	-0.02	-0.02	0.17	0.17	0.12	0.12	0.05	0.05
	Item53	-0.63	0.00	0.00	0.11	0.11	0.08	0.08	0.06	0.06
	Item54	2.29	0.01	0.02	-0.42	-0.42	-0.41	-0.41	-0.36	-0.35
	Item55	-1.40	-0.01	-0.01	0.12	0.12	0.08	0.08	0.02	0.02
	Item56	0.36	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00
	Item57	2.28	0.02	0.02	-0.44	-0.44	-0.40	-0.40	-0.36	-0.35
	Item58	0.85	0.01	0.01	-0.22	-0.22	-0.15	-0.16	-0.16	-0.16
	Item59	0.82	0.00	0.00	0.11	0.11	0.08	0.08	0.08	0.08
	Item60	1.30	0.00	-0.01	-0.27	-0.27	-0.19	-0.20	-0.19	-0.19
MAD	NA	NA	0.01	0.01	0.27	0.27	0.22	0.22	0.17	0.17
Item discrimination	Item1	1.30	-0.02	-0.02	-0.05	-0.05	0.12	0.12	0.05	0.05
	Item2	1.00	-0.02	-0.02	-0.07	-0.07	0.02	0.01	0.00	0.00
	Item3	1.69	0.00	-0.01	-0.29	-0.30	0.03	0.02	-0.05	-0.06
	Item4	1.99	0.01	-0.01	-0.14	-0.16	0.41	0.37	0.18	0.15
	Item5	2.37	-0.02	-0.03	-1.15	-1.15	-0.68	-0.69	-0.61	-0.61
	Item6	2.39	0.00	-0.05	-0.86	-0.87	-0.12	-0.15	-0.26	-0.30
	Item7	1.96	-0.01	-0.02	-0.56	-0.56	0.06	0.05	0.02	0.00
	Item8	0.93	-0.01	-0.01	-0.07	-0.07	0.03	0.03	0.04	0.04
	Item9	0.64	0.00	0.00	0.11	0.12	0.11	0.12	0.11	0.11
	Item10	2.28	-0.01	-0.02	-1.06	-1.06	-0.60	-0.60	-0.54	-0.54
	Item11	1.62	0.00	-0.02	-0.24	-0.24	0.24	0.22	0.16	0.14
	Item12	1.50	0.00	0.00	-0.36	-0.36	-0.13	-0.14	-0.15	-0.15
	Item13	0.58	0.00	0.00	0.17	0.18	0.18	0.18	0.18	0.18
	Item14	2.18	-0.01	-0.02	-0.89	-0.89	-0.36	-0.36	-0.32	-0.32
	Item15	2.28	-0.01	-0.03	-0.44	-0.45	0.23	0.20	0.03	0.01
	Item16	1.52	-0.01	-0.02	-0.34	-0.34	-0.01	-0.01	-0.01	-0.01
	Item17	1.05	0.00	-0.01	0.05	0.05	0.14	0.14	0.09	0.09
	Item18	0.87	-0.01	-0.01	-0.06	-0.06	0.00	0.00	0.01	0.01
	Item19	1.79	0.00	-0.01	-0.16	-0.17	0.25	0.23	0.09	0.08
	Item20	2.20	-0.01	-0.02	-0.83	-0.83	-0.20	-0.20	-0.19	-0.20
	Item21	1.23	-0.01	-0.01	-0.25	-0.25	-0.11	-0.11	-0.10	-0.10
	Item22	1.50	0.01	-0.01	0.23	0.22	0.53	0.50	0.32	0.30
	Item23	0.71	0.00	0.00	0.04	0.04	0.08	0.08	0.08	0.09
	Item24	0.60	0.00	0.00	0.09	0.09	0.10	0.10	0.11	0.11
	Item25	1.25	-0.01	-0.01	-0.18	-0.18	-0.04	-0.04	-0.06	-0.06
	Item26	0.72	-0.01	0.00	0.04	0.04	0.06	0.06	0.06	0.07
	Item27	2.04	-0.01	-0.02	-0.85	-0.85	-0.44	-0.44	-0.39	-0.39
	Item28	1.95	-0.02	-0.03	-0.39	-0.40	0.07	0.05	-0.06	-0.06
	Item29	0.53	0.00	0.00	0.11	0.11	0.10	0.10	0.11	0.11

Table A**55**

Bias of estimated item parameters for manipulation of the 50% of items and 30% of sample (60 items)

		True	**MLfit**	**BMfit**	**MLmisfit**	**BMmisfit**	**LzML**	**LzBM**	**HTML**	**HTBM**
						N =	4226	4225	4105	4097
	Item30	1.32	-0.01	-0.01	-0.24	-0.24	0.00	0.00	0.01	0.01
	Item31	2.16	-0.03	-0.04	-0.66	-0.66	-0.17	-0.18	-0.26	-0.26
	Item32	2.09	0.01	-0.02	-0.16	-0.17	0.48	0.44	0.21	0.18
	Item33	2.25	-0.01	-0.02	-0.52	-0.52	0.10	0.08	-0.06	-0.08
	Item34	0.62	0.00	0.00	0.07	0.07	0.08	0.08	0.08	0.09
	Item35	1.08	-0.01	-0.02	-0.01	-0.01	0.16	0.16	0.14	0.13
	Item36	1.35	-0.01	-0.01	-0.17	-0.17	0.02	0.01	-0.03	-0.03
	Item37	1.53	-0.02	-0.02	-0.29	-0.29	0.08	0.07	0.06	0.05
	Item38	2.23	-0.02	-0.02	-0.94	-0.94	-0.53	-0.54	-0.54	-0.54
	Item39	2.35	-0.01	-0.03	-0.53	-0.54	0.18	0.14	-0.02	-0.05
	Item40	2.31	-0.01	-0.02	-0.98	-0.98	-0.39	-0.39	-0.35	-0.35
	Item41	1.81	-0.01	-0.02	-0.57	-0.57	-0.14	-0.14	-0.13	-0.13
	Item42	0.70	0.00	0.00	0.21	0.21	0.20	0.20	0.16	0.16
	Item43	1.05	0.00	-0.01	0.21	0.21	0.30	0.29	0.20	0.19
	Item44	2.38	-0.02	-0.04	-0.90	-0.90	-0.12	-0.13	-0.14	-0.15
	Item45	0.50	0.00	0.00	0.16	0.17	0.14	0.14	0.14	0.14
	Item46	1.65	-0.01	-0.01	-0.41	-0.41	-0.01	-0.01	-0.01	-0.02
	Item47	0.80	-0.01	-0.01	0.24	0.24	0.24	0.24	0.19	0.18
	Item48	1.04	0.00	0.00	-0.18	-0.17	-0.06	-0.06	-0.05	-0.04
	Item49	1.71	-0.02	-0.04	0.11	0.09	0.51	0.48	0.28	0.25
	Item50	0.53	0.00	0.00	0.17	0.17	0.14	0.14	0.14	0.14
	Item51	1.50	-0.01	-0.02	0.20	0.19	0.51	0.48	0.30	0.28
	Item52	1.27	-0.01	-0.02	-0.12	-0.13	0.04	0.03	0.00	-0.01
	Item53	0.71	0.00	0.00	0.04	0.04	0.06	0.07	0.07	0.07
	Item54	1.59	0.00	-0.01	-0.25	-0.26	0.20	0.19	0.14	0.12
	Item55	1.47	-0.01	-0.01	-0.24	-0.24	-0.02	-0.02	-0.06	-0.06
	Item56	1.10	-0.01	-0.01	-0.22	-0.22	-0.10	-0.10	-0.08	-0.08
	Item57	1.37	-0.01	-0.02	-0.14	-0.14	0.18	0.18	0.14	0.14
	Item58	0.57	-0.01	0.00	0.10	0.10	0.10	0.10	0.11	0.11
	Item59	2.39	0.00	-0.01	-1.12	-1.12	-0.59	-0.59	-0.52	-0.52
	Item60	0.70	0.00	0.00	0.05	0.05	0.08	0.09	0.09	0.09
MAD	NA	NA	0.01	0.01	0.34	0.34	0.19	0.18	0.15	0.15

Notes: True= true item parameter; ML/BMfit= ML/BM estimate of fitting data set; ML/BMmisfit= ML/BM estimate of misfitting data set; LzML/BM= ML/BM item parameter estimate based on selected fitting data set using $\boldsymbol{l_z}$; HTML/BM= ML/BM item parameter estimate based on selected fitting data set using H^T; MAD= mean absolute difference.