University of Alberta

Polytomous item response theory parameter recovery: An investigation of non-normal distributions and small sample size

by

Louise Marie Bahry

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Education
in
Measurement, Evaluation and Cognition

Department of Educational Psychology

Spring 2012
Edmonton, Alberta

Abstract

Item Response Theory (IRT) has been extensively used in educational research with large sample sizes and normally distributed traits. However, there are cases in which distributions are not normal, and research has shown that the estimation of parameters becomes problematic with non-normal data. This study investigates the effects of skewness on parameter estimation using the Graded Response Model (GRM) and MULTILOG. Three distribution types (extreme and moderate skewness and a baseline condition (i.e. normal) and seven sample sizes (from n = 100 to n = 3,000) were investigated using simulations. In keeping with previous findings, the extremely skewed distribution condition resulted in the poorest estimates regardless of sample size. In general, the accuracy of parameter estimation increased as sample size increased. For the normally distributed conditions, results suggest a minimum sample size of 750 for accurate estimation. Implications of these findings are discussed.

Acknowledgement

First, I would like to thank my family and friends who have supported me in this endeavour and without whom I could not have completed this very daunting task. In particular, thank you to my parents, Doug and Laureen who have always been my cheerleaders and Mrs. Anne Bahry, my Grandmother whose strength has always inspired me. My fiancé, Garnet, thank you for putting up with me throughout this process and Kristy and Ryan Carlson for taking in a wayward graduate student in the midst of her thesis, you knew not what you were in for.

I would like to extend my deepest gratitude to my advisor, Dr. Todd Rogers for his invaluable feedback throughout and Dr. Mark Gierl for introducing me to item response theory and helping me to find my way in measurement. Thank you also to my other committee members, Dr. Ying Cui and Dr. Connie Varnhagen, for your patience and support. To my friends and colleagues, Dr. Andrea Gotzmann, Ulemu Luhanga, and Dorothy Pinto thank you for your ears and minds.

Table of Contents

List of Tables

List of Figures

Chapter One

Item Response Theory (IRT) is an approach, or family of statistical models, used to analyze assessment item data. These models relate examinee ability (θ) and item parameters using logistic functions. Several IRT models have been developed to estimate examinee ability (or latent trait) and the item parameters for items that are scored either dichotomously (i.e. only two response categories) or polytomously (i.e. more than two response categories; Hambleton, Swaminathan, & Rogers, 1991).

Traditionally, IRT has been used for educational applications such as Computerized Adaptive Testing (CAT), test score equating, item analysis, and item banking. However, due to the advantages of IRT other disciplines have recently developed an interest in using IRT for scoring, validation, and other psychometric analyses (Reise & Henson, 2003).

Samejima (1969) extended the two-parameter logistic dichotomous item response theory (IRT) model to deal with ordered, categorical responses. She developed the graded responses (GRM) model to allow IRT to be used with data derived from polytomously-scored items included in an achievement test and which are scored using a scoring rubric or an analytic scoring scale. Additionally, the GRM was developed for use with assessments including likert-type response items such as those from attitude scales, psychological inventories or clinical assessments, where the different points along the response scale receive different scores.

There are several applied examples in the social sciences in which the

GRM has been used to fit item data to a model, estimate parameters, or to generally validate assessments. The assessments used vary across educational assessments and personality inventories to health questionnaires in which both dichotomously- and polytomously- scored items or only polytomously-scored items are used. Given the focus of the present study, the review of empirical studies is limited to studies outside of education with only polytomously-scored items. The sample sizes employed in these studies vary from 126 (Schrum & Salekin, 2006) to 13,059 (Chernyshenko, Stark, Chan, Drasgow & Williams, 2001) and the number of items vary from 6 (Gumpel, 1999) to 198 (Walton, 2008).

Schrum and Salekin (2006) used MULTILOG to calibrate a 20 item assessment with a 3-point graded scale and a sample size of 123. Gumpel (1999) calibrated a six item assessment with a 4-point graded scale and sample size of 139; but the program used was not identified. de Ayala (2009) recommended a minimum sample size of 500 for calibration using polytomous models (assuming normally distributed θ and IRT assumptions are met) and suggested that there may be a "point of diminishing returns" (p.223) after which increasing the sample size will not increase the accuracy of estimation. In a simulation study conducted by Reise and Yu (1990), it was suggested that a sample size above 500 is sufficient for calibration of a 25-item assessment under the GRM. Reise and Yu also found that smaller sample sizes affected the estimation of item parameters but did not affect estimation of the θ parameter.

Research Purpose and Questions

The purpose of the present study was twofold. The first purpose was to identify the effect of sample size and non-normal ability (θ) distributions on the accuracy and precision of the estimation of the item parameters at the test level using the GRM and the MULTILOG program. The second purpose was to identify the effect of sample size and non-normal ability (θ) distributions on the accuracy and precision of the estimation of the item parameters at the item level using the GRM and the MULTILOG program.

In order to address these purposes, a simulation study was conducted in which real data studies for distribution type and sample size were referenced to carry out the simulation. Two factors were varied in the study: underlying θ distribution type and sample size. The following four research questions will be addressed using simulated data:

1) Does the shape of the underlying θ distribution have an effect on test-level statistical outcomes for item and person parameter recovery under the GRM using MULTILOG?
2) Does the shape of the underlying θ distribution have an effect on item-level statistical outcomes for item and person parameter recovery under the GRM using MULTILOG?
3) Does sample size have an effect on test-level statistical outcomes for item and person parameter recovery under the GRM using MULTILOG?
4) Does sample size have an effect on item-level statistical outcomes for item and person parameter recovery under the GRM using MULTILOG?

Evaluation criteria included two outcome measures at both of the levels of analysis. RMSEs and test-level BIAS statistics calculated across items were to assess effects on total test scores and item-level BIAS and standard error of item BIAS were calculated to assess item-level effects.

Delimitations

While there are several IRT programs that can be used to complete a calibration of polytomously scored items and to estimate the latent trait parameter, only MULTILOG with the GRM was used in the study. Comparison of different computer programs and calibration and estimation procedures was not a purpose of the present study. In addition, only a 5-point score scale and 20-item assessment was simulated. This decision was made given the common use of a 5-point response scale and the average number of items included in the studies in the personality and health areas.

Organization of Thesis

The introduction of the research on applied and simulation-based studies using polytomous item response theory (PIRT) models and the presentation of the research questions was presented in Chapter One. Chapter Two contains the literature review and the logic in support of the present research. Chapter Three describes the methods that were used in this study including a description of the GRM, calibration procedures, simulation conditions, and evaluation procedures used to assess the results. Results are presented in the next two chapters. Test-level results are presented and discussed first in Chapter 4, followed by item-level results in Chapter 5. Lastly, Chapter Six contains a summary of the research findings, a discussion of the limitations of the current study, conclusions, implications for practice and future research directions.

Chapter Two: Review of the Literature

This chapter provides a review of the literature on parameter estimation and recovery using the Graded Response Model (GRM, Samejima, 1969). In the literature, there are parameter recovery studies that have incorporated three different item formats including dichotomous items only (Bahry & Gotzmann, 2011; Drasgow, 1989; Wang & Chen, 2005), mixed-item formats including both dichotomous and polytomously-scored items (Toland, 2008), and polytomous items only (Dodd, 1984; Si, 2002; Sinar & Zickar, 2002; Kang, Cohen & Sung, 2009). While the dichotomous-only and mixed-item assessment formats have been studied in great detail, polytomous-only is the focus of this study since this item format has not been evaluated to the same extent. Thus, the review of the literature is focussed on studies using assessments with only polytomously-scored items.

First, a brief introduction to Item Response Theory (IRT) is provided including a description of the GRM and the estimation process used in the MULTILOG software (Thissen, Chen, & Bock, 2003) used in this study. This is followed by a review of the application of IRT item parameter estimation and parameter recovery research using the GRM with assessments with only polytomously-scored items. The chapter concludes with a statement of the purpose of the present research.

Overview of IRT

Item Response Theory (IRT) is an approach, or family of statistical models, used to analyze assessment item data. These models relate examinee ability (θ) and item parameters using logistic functions. Several IRT models have been developed to estimate ability or person parameters that are scored either dichotomously (i.e. only two response categories) or polytomously (i.e. more than two response categories; Hambleton et al., 1991). Traditionally, IRT has been used for educational applications for Computerized Adaptive Testing (CAT), test score equating, item analysis, and test banking. However, due to the advantages of IRT other disciplines have recently developed an interest in using IRT for scoring, validation, and other psychometric analyses (Reise & Henson, 2003).

IRT ability or person parameters (θ) are not item or test dependent and item and test characteristics are not dependent on the ability or person parameters. This is called the property of invariance (Hambleton et al., 1991; Lord, 1980) and means that the test and item parameters remain the same regardless of the sample of respondents, and the ability or person parameters do not vary depending on the test items administered or time of test provided the items are relevant to and representative of the same domain of interest.

At the foundation of IRT is the item response function (IRF), which gives the probability of observing a particular response to a particular item given the examinee's latent trait value (i.e., ability, personality trait, etc.) and the parameters of the item. The item characteristic function (ICF) defines the expected item score given an examinee's ability, and the item characteristic curve (ICC) is a graphical representation of the ICF. When considering polytomous item response models, there is a curve for each scoring category; in this case, the curves are called operating characteristic curves (OCC's).

When the test items are all scored dichotomously, there are three basic models for analyzing the data: the one-, two-, and three-parameter logistic models. The one-parameter (1PL) model is the most basic and involves, as the name states, only one item parameter: the *b*-parameter is included in every IRT model and is considered the difficulty parameter (Yen & Fitzpatrick, 2006). The *b*-parameter is at the point on the θ scale where the probability of a correct response is equal to 0.50 and typically varies from -2.00 to 2.00 (Hambleton et al., 1991; Yen & Fitzpatick, 2006) increasing as items become more difficult. Figure 1 is a visual representation of the effect of changes in parameter *b*.

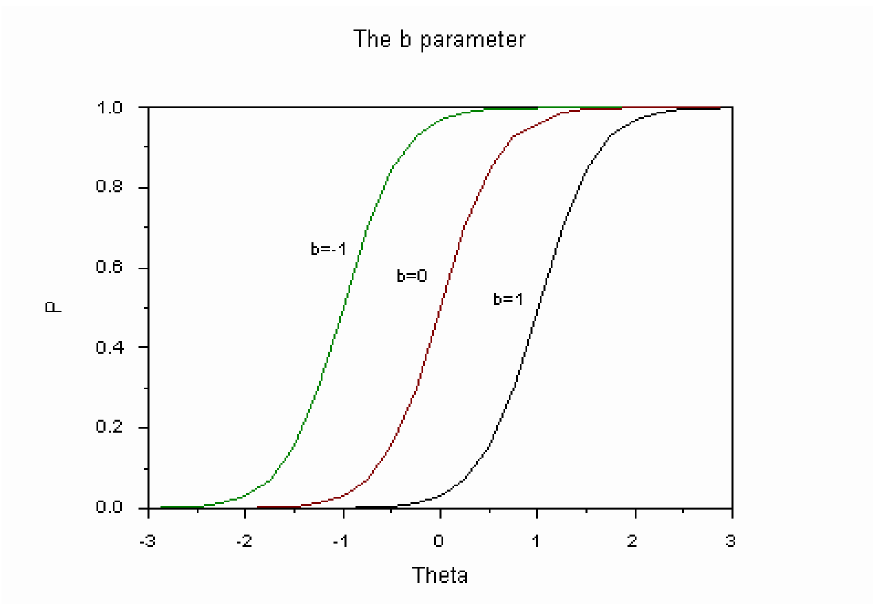

*Figure 1.*ICCs showing the effect of increasing parameter *b*

The two-parameter model (2PL) includes a second item parameter, the discrimination parameter, *a*. *a* is the slope of the ICC at the point of inflection and the higher the value of *a*, the more sharp the discrimination (Yen & Fitzpatrick, 2006). The *a*-parameter is included when it is assumed that items on an assessment vary in their discriminating power. *a*-parameters typically range from

0 to 2.00 with values ranging from 0.40 to 2.50 considered good (Hambleton et al., 1991). The *b*-parameter is at the point on the *θ* scale where the probability of a correct response is equal to 0.50. Figure 2 is a visual representation of changes in *a*. Here, we see that as *a* increases, the range of *θ* decreases for that item. That is, the information provided by an item with a large value of *a*, will be greater.

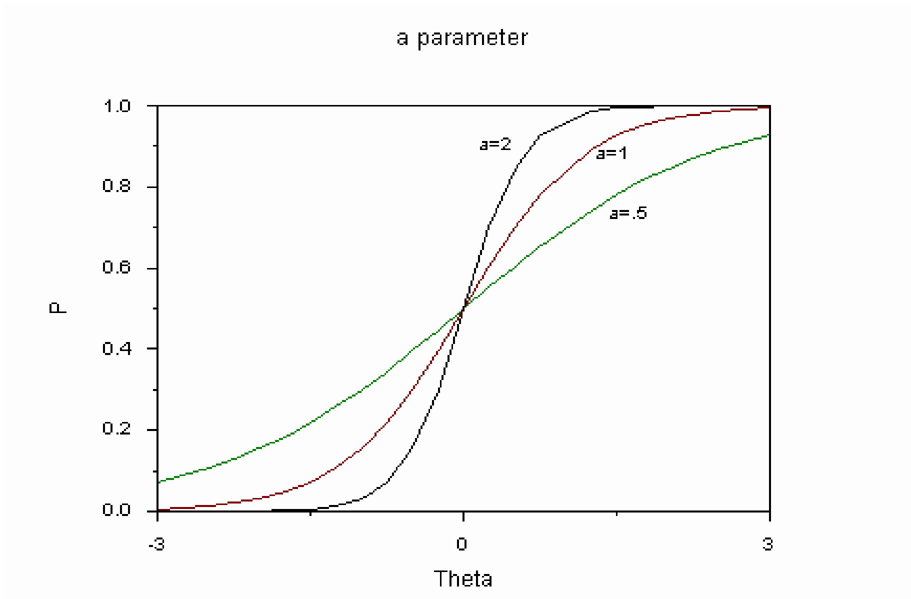

*Figure 2.*ICCs showing the effect of increasing parameter *a*

Finally, the three-parameter model (3PL) includes the *c*-parameter, called the guessing or pseudo-chance parameter. This parameter was introduced to account for the possibility that even students with low ability have some chance of answering even difficult questions correctly. This parameter is not always necessary, and if set to zero, equates the 3PL with the 2PL (Yen & Fitzpatrick, 2006). In the case of the 3PL model, the value of the *b*-parameter is dependent on the value of the lower asymptote (*c*-parameter). In this case, the *b*-parameter is at the point on the *θ* scale where the probability of a correct response is equal to

$\frac{c+100}{2}$. Figure 3 is a representation of changes to parameter c and the resulting changes to the probability of an examinee's response to an item.

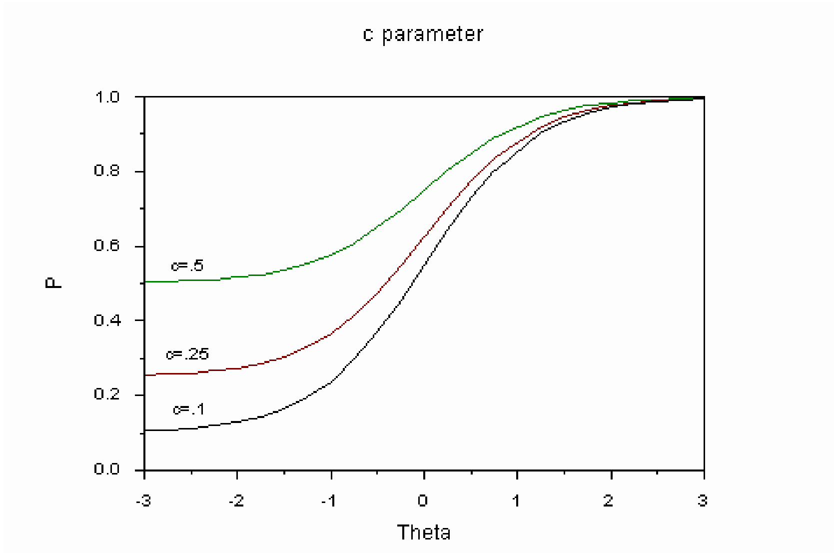

*Figure 3.*ICCs showing the effect of increasing parameter *c*

Although there are clear benefits to the invariance property, there are two integral assumptions of IRT. First, there is an assumption regarding the dimensionality of the underlying ability or trait. While there are multi-dimensional IRT models (MIRT), the model used in this study requires that a single trait or ability accounts for an individual's θ score. When this assumption of the data holds, the examinees can be placed along a single, meaningful scale (Hambleton et al., 1991).

Second, there is the assumption of local independence. When the items on an assessment are locally independent, a response to any item is independent of a response to any other item on the same assessment for a given individual. This assumption allows us to determine the probability of an individual response pattern occurring given the individual's ability or trait level (Hambleton et al., 1991; Lord, 1980). It is the case that if the first assumption of unidimensionality is met, then the assumption of local independence will also be met.

In addition to these assumptions, an assessment of model-data fit is also important in IRT. A poorly specified model creates problems with estimating both item parameters and θ scores. Consider the following: an analyst mistakenly specifies a model which only specifies *a*- and *b*-parameters when in fact the data fit a model consisting of all three item parameters. Because the *c*-parameter has not been specified, the θ values may be over-estimated as the individual’s ability to correctly guess the answer has not been taken into consideration. Guessing is not considered to be included in ability and, as such, it should not be allowed to unduly influence scores.

Graded Response Model

Samejima (1969) extended the 2PL dichotomous IRT model to deal with ordered, categorical responses. She developed the graded responses (GRM) model to allow IRT to be used with data derived from polytomously-scored items included in an achievement test and which are scored using a scoring rubric or analytic scoring scale and with likert-type response data used in attitude scales, psychological inventories or clinical assessments, where the different points along the response scale receive different scores. In essence, the GRM is an application of the 2PL to an ordered series of dichotomous responses and specifies the probability of responding in *k* or higher response categories as opposed to lower than *k* response categories (e.g., for a three point scale, 0 vs. 1 and 2 and 0 and 1 vs. 2; de Ayala, 2009). The probability (P) of obtaining a score (x_j) or higher is defined as:

$$P^*_{xj}\ (\theta) = \frac{e^{\alpha j(\theta - \delta xj)}}{1+e^{\alpha j(\theta-\delta xj)}}\,, \tag{2.1}$$

where θ is the latent trait,

α_j is the discrimination parameter for item j,

δ_{xj} is the category boundary location for category score x_j, and

$x_j = \{0,1\ldots m_j\}$ where m_j is the largest category score for item j. The value of m_j need not be the same for all items.

The GRM is considered as a difference model because the probability of obtaining a specific category score x_j on item j involves a two-step process. Equation 2.1 provides the probability of attaining a category score or higher and must be solved for each score category (i.e. $x_j = 0,1,\ldots m$). This provides the operating characteristic functions for the k thresholds. Next, the following equation is used:

$$p_k = P^*_k - P^*_{k+1} \tag{2.2}$$

where P^*_k is P^*_{xj} from equation 2.1. And p_k gives the probability of responding in a particular category given θ by subtracting adjacent $P_k^*(\theta)$ values. Because by definition responding above the highest response category is $p_k = 0.00$, the probability of responding within the highest category is equal to the highest operating characteristic function calculated using Equation 2.1.

Parameter Estimation Using MULTILOG[1]

In MULTILOG, item parameter estimation can be done in one of three ways depending on whether θ is assumed to be a fixed or random variable. If θ is assumed to be fixed and linearly related to the observable variable, parameters can be estimated using nonlinear regression (Roche, Wainer & Thissen, 1975). If θ is assumed to be fixed but unknown, simultaneous estimation of the fixed values of θ and item parameters is used (Bock, 1976) by dividing the examinees into homogenous groups.

Finally, when θ is assumed to be a random unobserved variable Bock and Aitken (1981) proposed using marginal maximum likelihood estimation (MMLE) which integrates the unknown ability parameter out over the parameter distributions and uses the marginal distributions to estimate item parameters. Their reformulation of the algorithm initially proposed by Dempster, Laird and Rubin (1977) allows for an unknown ability distribution to be estimated along with the item parameters.

Trait Score Estimation Using MULTILOG

Trait score (θ) estimation in MULTILOG can be done in one of two ways: maximum likelihood (MLE) or expected a posteriori (EAP). The MLE of θ is the value at which an examinee has the highest likelihood of responding given the observed response pattern and item properties. However, in order for MLE to be computed, an examinee must have both correct and incorrect responses on an assessment. That is, given a dichotomous assessment, the response patterns

[1] PARSCALE (Muraki & Bock, 1997) was considered for calibration. However, when attempted with a skewed distribution condition, the program stopped running and produced an error file due to a lack of data in all possible categories.

[0,0,0,0,0] and [1,1,1,1,1] will produce an estimation error when using MLE. In contrast, the EAP procedure uses the mean of the posterior distribution rather than the mode as in the MLE (Bock & Mislevy, 1982). In this case, all response patterns can be used.

Empirical Studies Using the GRM

There are several applied examples in the social sciences in which the GRM has been used to fit item data to a model, estimate parameters, or to generally validate assessments. The assessments used vary across educational assessments and personality inventories to health questionnaires in which both dichotomously- and polytomously- scored items or only polytomously-scored items are used. As mentioned early, given the focus of the present study, the review of empirical studies is limited to studies outside of education with only polytomously-scored items. The sample sizes employed in these studies vary from 126 (Schrum & Salekin, 2006) to 13,059 (Chernyshenko, Stark, Chan, Drasgow & Williams, 2001) and the number of items vary from 6 (Gumpel, 1999) to 198 (Walton, Roberts, Krueger, Blonigen & Hicks, 2008).

One assessment that has been analysed more than once using Polytomous Item Response Theory (PIRT) is the 20-item Psychopathy Checklist (PCL), both the Revised (PCL-R; Hare, 1991) and Youth Version (PCL-YV; Forth, Kosson & Hare, 2003) forms. PCL items are scored on a 3 point scale wherein 0 translates to a complete absence of the behaviour, 1 translates to an occasional presence of the behaviour and 3 translates to the continuous presence of the behaviour.

Cooke, Michie and Kosson (2001) evaluated the structural, item, and test generalizability of the PCL-R using IRT methods. Two samples, one with 359 participants and another with 356 participants, were calibrated using the GRM and the computer program MULTILOG. Cooke et al. used IRT methods to investigate Differential Item Functioning (DIF) of the PCL-R for Caucasians and African Americans. DIF, in the context of the PCL, is expected to occur when individuals with the same level of psychopathy from different groups have differing probabilities of obtaining the same score on a particular item. Two PCL-R factor models, one using 13 items and another using all 20, were calibrated for both samples in MULTILOG using the GRM. Five items showed significant differences across the two samples and the magnitude was small.

Bolt, Hare, Vitale & Newman (2004), also investigated DIF on the PCL-R using three methods across four samples: male criminal offenders (n = 3,847), female criminal offenders (n = 1,219), male psychiatric forensic patients (n = 1,246) and male criminal offenders scored only from file review (n = 2,626). Each sample was calibrated using MULTILOG with the GRM and both item and θ parameters were estimated. A large number of items displayed DIF in the study but as with the results of Cooke et al. (2001) the magnitude was small.

Finally, Schrum and Salekin (2006) analysed the assessment data from a sample of 123 responses to the PCL-YV from adolescent females from a detention centre. They also used the GRM and MULTILOG program to calibrate item and person parameters and to investigate item discrimination. Results showed that items discriminated the sample of juveniles differently from other samples.

Health research has also seen an increase in the use of IRT for test and item development. Cook et al. (2007) calibrated the data from 1,714 patient responses on a two scales from a health-related quality of life (HRQOL) measure: the general distress pool (15 items) and the physical function pool (23 items). Three different PIRT models were compared for fit with the data: the partial credit model (PCM; Masters, 1982), the generalized partial credit model (GPCM; Muraki, 1992), and the GRM, and two software programs were used: WINSTEPS (Linacre, 2002) and PARSCALE 3 (Muraki & Bock, 1997).

In addition to item and DIF analyses using IRT, item parameters estimated were used to simulate a computerized adaptive testing (CAT) environment with the items from the HRQOL instrument. Results indicated that in the health sciences, multidimensional IRT models may be of more use.

Hays, Liu, Spritzer, and Cella (2007) also calibrated sample data from 15 items assessing physical functioning from the HRQOL measure (n = 3,223) in order to inform the creation of an item bank. MULTILOG software was used in the calibration of data with the GRM and results indicated good fit with the model. However, the *b*-parameters for the majority of the 15 items were very low on the θ scale and recommendations include the creation of items more evenly placed along the scale.

Simulation Studies Using the GRM

There are only a small number of simulated data parameter recovery studies using PIRT models. A seminal article by Reise and Yu (1990) posits that the minimum number of participants be 500 in order to estimate the parameters using the GRM when using an instrument with 5 response categories. The authors used the MULTILOG program to estimate parameters across 36 conditions: sample size (n = 250, 500, 1,000, and 2,000), true θ distribution (normal, uniform, and positively skewed), and true *a*-parameters (poor, moderate, and average discrimination). Outcome measures for the study included root mean square errors (RMSE), correlations between the true and estimated parameters, and mean comparisons between true and estimated parameters.

Reise and Yu's results indicated that the accuracy of the recovery of *a*-parameters increased across θ distributions from uniform to normal to positively skewed. Five hundred examinees were necessary to bring the RMSE below 0.10, and 1,000 examinees were need to obtain correlations between the 'true' and estimated *a*-parameter values above 0.90. The results for the *b*-parameters were similar to those for the *a*-parameters, with RMSEs decreasing with increasing sample size and correlations between 'true' and estimated *b*-parameters increasing with increasing sample size. Recovery of the θ parameters was generally poorer than the *a*- and *b*-parameters and was less affected by changes in sample size.

Sinar and Zickar (2002) used simulation methods to investigate the influence of the inclusion of deviant items that did not assess the construct of interest. A total of 45 conditions were calibrated: scale intercorrelations (-0.60, -0.30, 0.00, 0.30, 0.60), *a*-parameters for the focal scale (low, average, and high discrimination), and *a*-parameters for the scale with deviant items (low, average, and high discrimination). They use the GRM and the MULTILOG program to obtain parameter estimates.

Eight ANOVAs were run to investigate the influence of deviant items on traditional psychometric measures (classical test theory) and IRT with the dependent variables as the change in discrimination. Results indicated that construct irrelevant items were not significantly problematic for IRT analysis results due to the property of invariance and that when the item pool was well-defined an IRT model may be preferable to a classical model.

Purpose of the Study

Though developed and utilized heavily in the field of Education, IRT has been increasingly used in the social sciences and medicine for scale analysis and validation. When looking at large-scale assessment data in Education, large sized samples often with scores that are approximately normally distributed is the norm. However, as evidenced above, the recommended samples sizes were not met for many of the studies in which PIRT was used in the social and heath sciences areas. In addition, non-normal distributions are often seen in the social or health sciences due to the nature of the domain that is assessed.

Schrum and Salekin (2006) used MULTILOG to calibrate a 20-item assessment with a 3-point graded scale with a sample size of 123. Gumpel (1999) calibrated a six-item assessment with a 4-point graded scale and sample size of 139; but the program used was not identified. de Ayala (2009) recommended a minimum sample size of 500 for calibration using polytomous models (assuming normally distributed θ and IRT assumptions are met) and suggested that there may be a “point of diminishing returns” (p.223) after which increasing the sample size will not increase the accuracy of estimation. In a simulation study conducted by

Reise and Yu (1990), it was suggested that a sample size above 500 is sufficient for calibration of a 25-item assessment under the GRM. Reise and Yu also found that smaller sample sizes affected the estimation of item parameters but did not affect estimation of the θ parameter.

Thus, the purpose of the present study was twofold. First, to identify the effect of sample size and non-normal ability (θ) distributions on the accuracy and precision of the estimation of the item parameters at the test level using the GRM and the MULTILOG program at the test level. The second purpose was to conduct the analysis and provide outcome data at the item level to obtain information at the individual item level. Thus, the following four research questions will be addressed using simulated data:

1) Does the shape of the underlying θ distribution have an effect on test-level statistical outcomes for item and person parameter recovery under the GRM using MULTILOG?
2) Does the shape of the underlying θ distribution have an effect on item-level statistical outcomes for item and person parameter recovery under the GRM using MULTILOG?
3) Does sample size have an effect on test-level statistical outcomes for item and person parameter recovery under the GRM using MULTILOG?
4) Does sample size have an effect on item-level statistical outcomes for item and person parameter recovery under the GRM using MULTILOG?

The test simulated and the rationale for each factor it’s levels is presented in Chapter Three.

Chapter Three: Method

The simulation methods used in this research study are presented in this chapter. First, the independent variables investigated are described and the rationale for the levels chosen for these variables is presented. Second, descriptions of the processes carried out to simulate and calibrate the data for the study are described. Finally, the outcome measures used to evaluate the accuracy and precision of the estimates produced using the GRM and the MULTILOG program are described.

Independent Variables

Two independent factors were considered: type of underlying latent trait distribution (θ) and sample size.

Underlying latent trait distribution (θ). The type of underlying latent trait distribution (θ) was varied in this study because it has been shown that in some cases, the shape of the distribution of θ can affect parameter estimation (Reise & Yu, 1990; Toland, 2008). In order to accurately represent the type of data that one would collect with a clinical assessment, the level of negative skewness was varied for the underlying θ distribution. As the program WinGen3 (Han, 2007) was used, it was not possible to have complete control over the exact value of the skewness statistic. However, three levels of skewness were considered: extreme negative, moderate negative, and no skewness (i.e., normal).

Sample size. Sample size was chosen as a factor because, as shown in the previous chapter, research has shown that sample size does have an effect on the accuracy and precision of item parameter estimation (de Ayala, 2009; Drasgow, 1989; Seong, 1990; Reise &Yu, 1990). Seven sample sizes were investigated (n = 100, 250, 500, 750, 1,000, 1,500, 3,000). These sample sizes represent those found in applied literature and those generally found in clinical assessment situations where PIRT has been used. Of particular note are the two smallest sample sizes, which have been used in applied research studies and do not meet the recommendations provided by de Ayala (2009).

The three distribution shapes were crossed with the seven sample sizes to yield a 3 x 7 research design.

Data Generation and Calibration

The first step in the simulation was to generate item parameters for the 20 item assessment using WinGen3. The 20 *a*-parameters were simulated using a uniform distribution with a range of 0.400 to 2.500. The values of *a*-parameters typically range from 0 to 2.00 with values ranging from 0.40 to 2.50 considered good (Hambleton et al., 1991). The *b1*-, *b2*-, *b3*-, and *b4*-location parameters were simulated using a normal distribution (M=0.000, SD=1.000) and they ranged from -2.00 to 2.00 since this is the typical range for *b*-parameters (Hambleton et al., 1991; Yen & Fitzpatick, 2006). The item parameters used for the simulation are reported in Table 1 and are similar to those found in applied literature (Chernyshenko et al., 2001; Cooke et al., 2001; Schrum & Salekin, 2006). Next, three population distributions were sampled using WinGen3 to create θ distributions for each sample size. Two degrees of negative skewness developed using the 2-parameter beta distribution in an attempt to model the different distributions of clinical scores on a diagnostic instrument. Parameters of the population beta distributions were varied to keep the value of skewness at approximately -0.500 for the moderately-skewed conditions and -1.000 for the

Table 1

Item Parameters for 20-Item Assessment

	Parameters				
	a-	***b1-***	***b2-***	***b3-***	***b4-***
Item 1	0.735	-0.482	-0.073	0.121	2.030
Item 2	0.596	-0.750	1.112	1.643	2.343
Item 3	2.400	-1.090	-0.054	0.288	1.916
Item 4	0.637	-2.116	-0.420	0.481	0.987
Item 5	1.594	-0.779	-0.314	0.874	1.602
Item 6	1.804	-2.090	-1.360	-0.461	1.631
Item 7	0.629	-1.206	0.779	0.900	1.469
Item 8	1.252	-0.542	0.024	0.549	1.019
Item 9	1.372	-1.447	-0.786	-0.443	0.847
Item 10	1.522	-1.646	-1.585	-1.231	0.515
Item 11	2.376	-0.398	0.845	1.694	1.973
Item 12	1.204	-1.911	0.161	1.373	1.410
Item 13	2.466	-1.044	0.070	0.121	0.523
Item 14	0.833	-1.058	0.273	0.518	0.585
Item 15	1.793	0.223	0.426	0.730	0.998
Item 16	0.413	-2.044	-1.132	-0.292	0.866
Item 17	1.511	-0.706	-0.049	0.942	1.308
Item 18	1.857	-1.384	-0.505	0.474	1.399
Item 19	1.877	-0.987	-0.004	0.796	1.633
Item 20	2.440	-0.419	0.854	1.190	1.723
Mean	**1.466**	**-1.094**	**-0.082**	**0.513**	**1.339**

extremely skewed distributions. The normal distribution (approximately M = 0.000, SD = 1.000) was used as a baseline. Tables 2 and 3 contain the descriptive statistics for the normal distribution for each sample size (Table 2) and both skewed distribution conditions for each sample size (Table 3). In order to obtain stable results, 1,000 replications of each condition were conducted.

Table 2

Descriptive Statistics for all Normal Distribution Conditions

Distribution	Sample Size	Mean (M)	Standard Deviation	Skewness	Kurtosis
Normal	100	0.088	1.074	0.040	0.020
Normal	250	0.055	1.046	-0.130	0.350
Normal	500	0.060	0.951	0.010	-0.080
Normal	750	0.016	1.004	0.060	0.110
Normal	1000	-0.019	1.015	0.010	-0.200
Normal	1500	0.039	1.022	0.070	-0.030
Normal	3000	-0.012	0.988	-0.030	-0.060

Appendix ‘A’ contains the MULTILOG syntax used for the estimation of item parameters. The “RANDOM” command was used for marginal maximum likelihood (MMLE) parameter estimation, with “INDIVIDUAL” indicating the input format is individual item response vectors. Convergence was set to 0.001 with 500 calibration cycles in order to allow the software time to come to convergence. As MULTILOG does not produce an error message in the output parameter file, all output was utilized in calculating the outcome measures. Since MMLE uses the empirical θ distribution rather than making theoretical assumptions and inconsistencies due to problematic local maxima when estimating item parameters are eliminated (Bock & Aitkin, 1981).

Further, since the item parameters are estimated separately from ability, calibration using MMLE is more efficient than Joint Maximum Likelihood Estimation (JMLE) which estimates item and person parameters simultaneously (de Ayala, 2009). In addition, whereas MMLE has been shown to improve accuracy of estimation for shorter instruments, JMLE has been shown to produce biased estimates for instruments 15 items or shorter (Lord, 1983, 1986).

Table 3

Descriptive Statistics for all Skewed Conditions

		Beta Parameters					
Distribution	**Sample Size**	**α**	**β**	**Mean (M)**	**Standard Deviation**	**Skewness**	**Kurtosis**
Moderate Negative	100	3	2	0.640	1.143	-0.550	-0.280
Moderate Negative	250	4	2	0.993	1.119	-0.480	-0.360
Moderate Negative	500	5	2	1.334	0.875	-0.500	-0.210
Moderate Negative	750	4	2	0.989	1.029	-0.480	-0.220
Moderate Negative	1000	5	2	1.303	0.945	-0.560	-0.150
Moderate Negative	1500	5	2	1.234	0.974	-0.530	-0.170
Moderate Negative	3000	5	2	1.270	0.932	-0.510	-0.320
Extreme Negative	100	6	2	1.591	0.859	-1.050	0.960
Extreme Negative	250	8	2	1.819	0.769	-1.080	0.770
Extreme Negative	500	10	2	1.965	0.647	-1.090	1.570
Extreme Negative	750	8	2	1.794	0.775	-0.970	0.740
Extreme Negative	1000	10	2	2.013	0.611	-0.970	0.880
Extreme Negative	1500	10	2	1.990	0.641	-1.000	0.890
Extreme Negative	3000	10	2	2.001	0.619	-1.000	1.040

Appendix ‘B’ contains the MULTILOG syntax used to calibrate the person parameters *(θ).* The “SCORE” command computes θ scores, using

Maximum A Posteriori (MAP) estimation as default. MAP is a Bayesian approach to parameter estimation that uses an iterative method and a continuous prior distribution. Bayesian estimation procedures can be used for any response pattern, including those with ‘perfect’ (all correct or incorrect) response patterns, unlike maximum likelihood procedures which demand both correct and incorrect responses in an individual’s response set. The possibility of perfect response patterns when dealing with extremely skewed distributions is large. Thus using a Bayesian approach was necessary for this study.

Data Analysis

Once all the MULTILOG runs were completed, the item and person parameters were read back into SAS (Version 9.2) and four outcome measures were presented: Root Mean Square Errors (RMSEs) across the 20 items by replication, Test-Level BIAS averaged across all 20 items and replications, Item-Level BIAS for each item across replications, and frequencies of non-convergence for each condition. The syntax used to combine the results into SAS and calculate the outcome measures is presented in Appendix ‘C’. In order to gain a true sense of the outcomes from an applied PIRT calibration wherein there are no ‘true’ parameters to use in a calibration procedure, estimated parameters were not scaled to the ‘true’ parameter scale for the purposes of this study.

RMSEs were calculated in three stages as follows:

Step 1: The $MSE_{r.}$ was calculated across the 20 assessment items for each replication:

$$MSE_{r.} = \frac{\sum_{i=1}^{20}\left(\xi_i - \hat{\xi}_{ir}\right)^2}{19}, \tag{3.1}$$

where $\hat{\xi}_{ir}$ = the estimated parameter for item *i* on replication *r*, and

ξ_i = the 'true' parameter for item *i*.

Step 2: The mean $MSE_{r.}$ was calculated across the 1000 replications:

$$MSE_{..} = \frac{\sum_{r=1}^{1000} MSE_{r.}}{999} \tag{3.2}$$

Step 3: The square root of the $MSE_{..}$ is equal to the RMSE across the 20 items for the 1000 replications.

The RMSE is the most commonly used and recommended statistic for parameter recovery studies such as this (Sass, Schmitt, & Walker, 2008; Seong, 1990; Stone, 1992; Tate, 1995). And the RMSE is also highly interpretable (Harwell, Stone, Hsu, & Kirisci, 1996) as it is calculated in parameter units. Thus, an RMSE = 1 translates to an absolute difference of one parameter unit between the estimated and 'true' parameters.

The second outcome measure to be used in this study is the average estimate of bias (BIAS). Test-Level BIAS is defined by:

$$BIAS = \frac{1}{1000}\sum_{r=1}^{1000}\left[\frac{1}{20}\sum_{i=1}^{20}\left(\xi_i - \hat{\xi}_i\right)\right] \tag{3.3}$$

where ξ_i = the 'true' parameter value for item *i*, and

$\hat{\xi}_i$= the estimated parameter for item *i*.

Test-level BIAS provides information regarding the direction and magnitude of bias for an estimated parameter relative to the corresponding 'true' parameter.

Item-Level BIAS was also calculated for each item across the 1,000 replications to aid in interpretation. These statistics were calculated as follows:

$$BIAS_i = \left[\frac{\sum_{r=1}^{1000}\left(\xi_i - \hat{\xi}_i\right)^2}{999}\right] \tag{3.3}$$

where ξ_i = the 'true' parameter value for item *i*, and

$\hat{\xi}_i$= the estimated parameter for item *i*.

Item-Level BIAS provides item-level information regarding the magnitude of the BIAS in item parameter estimates. The standard error (S.E.) of Item-Level BIAS was also calculated and provides information regarding the precision of those estimates.

Non-Convergence Frequencies

Before moving to the presentation of results in the next two chapters, it is first necessary to address the issue of non-convergence during the calibration phase. Problems were encountered, especially when the distribution was extremely negatively skewed and for the smaller sample sizes.

Table 4 shows the percentage of replications that did not converge when calibrating the data using MULTILOG. As shown, non-convergence was an issue with small sample sizes regardless of the distribution type. The default criterion for convergence for MULTILOG is set at 0.001. As shown in Table 4, in 51.4% of the replications failed to converge for the extreme negative (EN) distribution with n = 100 and this decreased to 11.4% with n = 3,000. The non-convergence for the moderate negative skewed (MN) distributions and normal distributions were more comparable with non-convergence for 26.8% and 26.9%, with n =

100 and 0.00% with n = 3,000.

Table 4

Percentages of Replications Without Convergence with Criterion Set at 0.01 and 0.001

Distribution Type	Sample Size	Criterion = 0.001	Criterion = 0.01
Normal	100	26.90%	9.20%
	250	21.50%	16.50%
	500	2.50%	1.90%
	750	0.60%	0.30%
	1000	0.10%	0.10%
	1500	0.00%	0.00%
	3000	0.00%	0.00%
Moderate Negative	100	26.80%	6.20%
	250	40.00%	12.60%
	500	16.30%	8.00%
	750	3.00%	1.00%
	1000	0.90%	0.10%
	1500	0.30%	0.00%
	3000	0.00%	0.00%
Extreme Negative	100	51.40%	9.80%
	250	38.70%	34.00%
	500	20.70%	9.80%
	750	14.50%	2.90%
	1000	20.10%	3.90%
	1500	16.10%	1.90%
	3000	11.40%	0.30%

Using a less conservative criterion of 0.01, the non-convergence rates decreased substantially. For example, for the EN distribution, the non-convergence rate was 9.80% with n = 100 and 0.30% with n = 3,000. Likewise, the non-convergence rates were reduced for the MN and normal distributions. For example, the non-convergence rates for the MN distribution was 6.2% with n = 100 and 0.0% with n = 1,500 and n = 3,000.

However, a feature of MULTILOG is to provide an estimate of the parameter of interest after the last completed cycle in the iterative procedure used (MML in this case). All 1,000 parameter estimates for each condition were included the calculation of the outcome measures. In such cases where the convergence criterion was not met, it is not known whether the estimates provided at the end of the 500 calibration cycles were over, under, or accurate estimates. As a result, the mean outcome measures provided may be too large, too small or correct. Non-convergence was taken into consideration when interpretations of the outcome measures were made.

Chapter Four: Results and Discussion - Test Level Analysis

The results of the simulations are presented in this and the next chapter. The current chapter presents results at the test level, whereas chapter 5 contains results at the item level. Results are presented in both chapters for each of the parameters separately. The RMSE and test level bias measures were used at the test level, and the item level BIAS and standard error of the BIAS were used at the item level. Each chapter concludes with general comments across conditions.

a-parameters

RMSE. With the notable exception of n = 500, the RMSEs for the *a*-parameters decreased in general for each distribution as the sample size increased (Figure 1.). As shown, the values of RMSE for all three distributions were large for n = 100 but dropped significantly for n = 250. As suggested above, RMSEs unexpectedly increased for n = 500, particularly for the MN distribution and the normal distribution conditions. There is no clear reason for this latter result. Beginning with n = 750, the RMSEs for the MN distribution and the normal distribution conditions were essentially the same and all less than 0.20. In contrast, the RMSEs for the EN distribution conditions were larger, varying from 0.32 to 0.52 across the four larger sample sizes.

Test-Level BIAS. As with the RMSE, the test-level BIAS results show the accuracy of the recovered parameters increased across all distribution conditions as sample size increased with a spike at n = 500 (Figure 2.). Note that as the

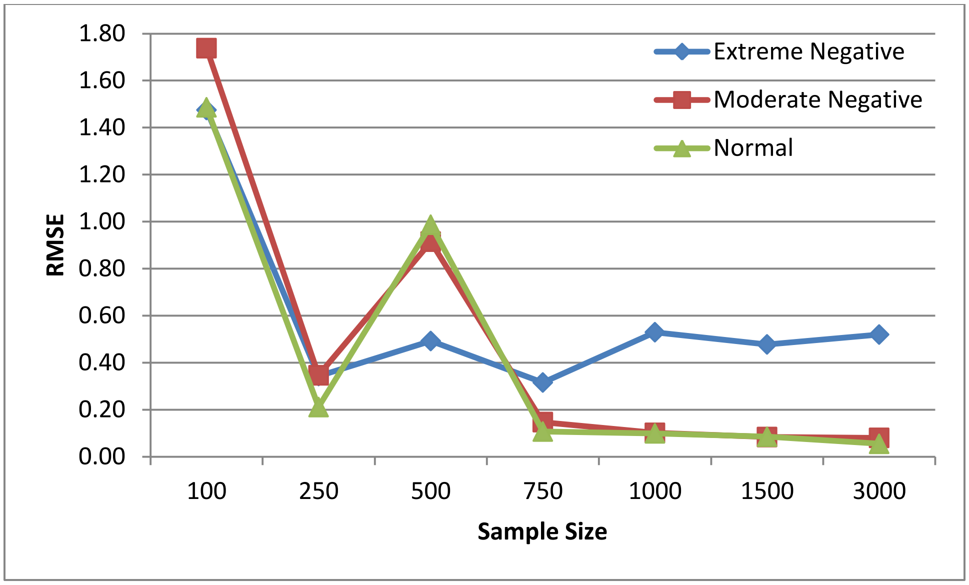

*Figure 4.*RMSEs of 'true' and estimated *a*-parameters by condition.

subtraction for both BIAS measures was computed 'true' minus estimated, a negative BIAS indicates an overestimate and a positive BIAS indicates an underestimate. With one exception (EN, n = 250), the *a*-parameter was overestimated for the three smaller sample size conditions.

Continuing with the four larger sample size, the test level bias for the *a*-parameter was slightly underestimated for the MN distribution and the normal distribution conditions for n = 750, and essentially zero for the remaining three sample sizes. In contrast, the test level bias for the EN distribution conditions was 0.25 for n = 750, after which it increased to close to 0.50 for the three larger sample sizes.

b1-parameters

RMSE. As shown in Figure 3, and unlike the case for the test-level *a*-parameter, the RMSEs for the *b1*-location parameter differed across the three

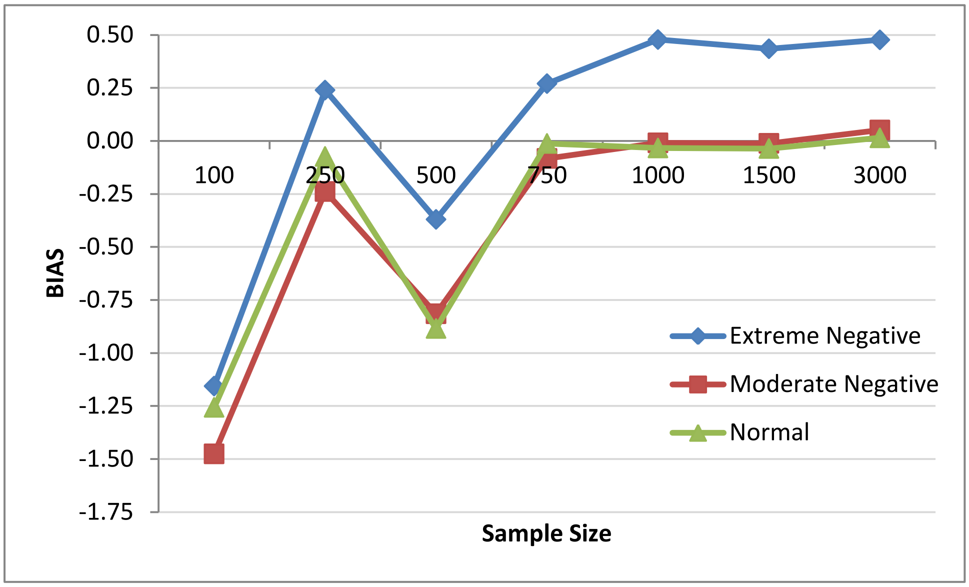

Figure 5.BIAS of 'true' and estimated *a*-parameters by condition.

distributions. RMSEs for the EN distribution conditions are all larger than the RMSEs for the MN distribution conditions which, with the exception of n =100, are all larger than the RMSEs for the normal distribution. The same spike in error occurs for the n=500 sample size with both the EN and MN distributions as with the *a*-parameter.

As expected, the RMSEs consistently decreased from 0.52 to close to zero for the normal, or baseline distribution conditions. In contrast, RMSEs increased for the EN and MN distribution conditions as the sample size increased from 100 to 500, then decreased for n = 750 in essentially parallel ways. The RMSEs then increased for both EN and MN distribution conditions, but more so for the EN distribution, at n = 1,000. After this point, values for the two distributions diverged from each other, with the RMSE remaining close to 1.40 for the MN distribution conditions, while the RMSEs for the EN distribution conditions varying between 3.62 and 4.04.

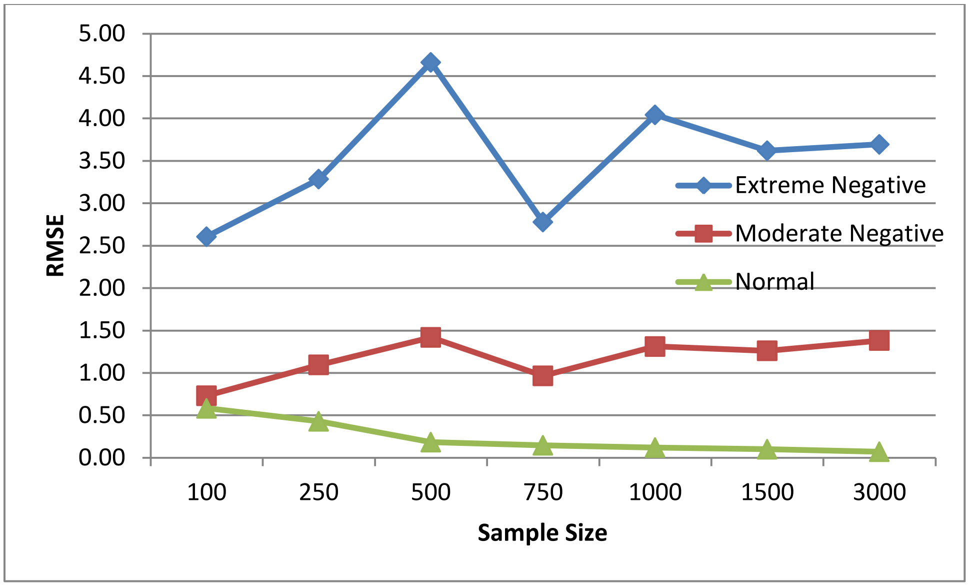

*Figure 6.*RMSEs of ‘true’ and estimated *b1*-parameters by condition.

Test-Level BIAS. As with the RMSE, the test-level BIAS for the *b1*-location parameter differed across the three distributions. And as with the RMSE, the BIAS was greatest for the EN distribution, followed in by the MN distribution and the normal distribution conditions. While the test-level BIAS was essentially zero across the seven sample sizes for the normal distribution, it increased for the EN and MN distribution conditions as the sample size increased from 100 to 500, then decreased for n = 750 in the same ways. BIAS then increased for both EN and MN distribution conditions, but more so for the EN distribution at n = 1,000. After this point, values for the two distributions diverged from one other, with the BIAS varying between 1.23 and 1.38 for the MN distribution and between 3.22 and 3.51 for the EN distribution conditions.

Test-Level BIAS. As with the RMSE, the test-level BIAS for the *b1*-location parameter differed across the three distributions. And as with the RMSE, the BIAS was greatest for the EN distribution, followed in by the MN distribution

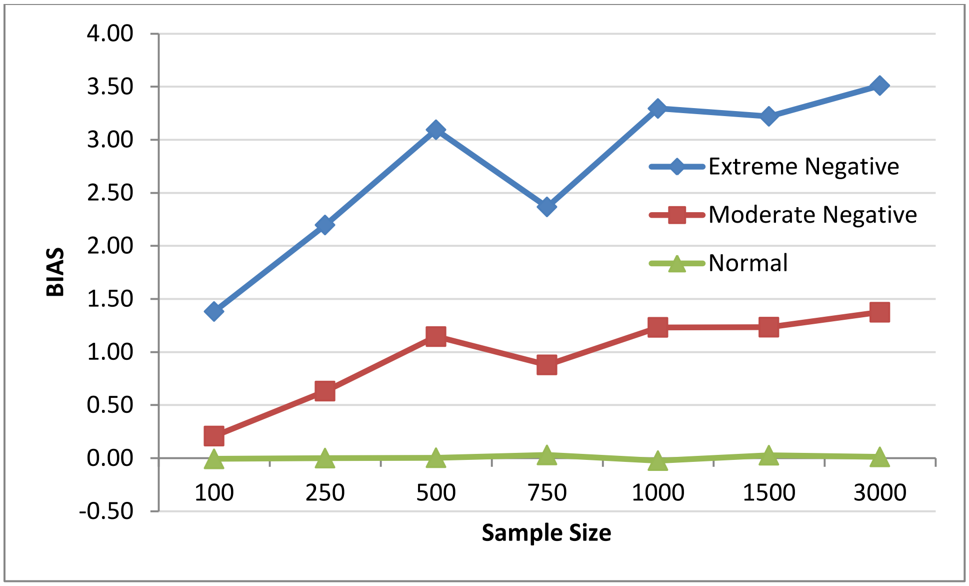

*Figure 7.*BIAS of 'true' and estimated *b1*-parameters by condition.

and the normal distribution conditions. While the test-level BIAS was essentially zero across the seven sample sizes for the normal distribution, it increased for the EN and MN distribution conditions as the sample size increased from 100 to 500, then decreased for n = 750 in the same ways. BIAS then increased for both EN and MN distribution conditions, but more so for the EN distribution at n = 1,000. After this point, values for the two distributions diverged from one other, with the BIAS varying between 1.23 and 1.38 for the MN distribution and between 3.22 and 3.51 for the EN distribution conditions.

b2-parameters

RMSE. As shown in Figure 5, the RMSEs for the *b2*-location parameter differed across the three distributions in the same way as for the *b1*-parameters. RMSEs for the EN distribution conditions are all larger than RMSEs for the MN distribution conditions which are all larger than the RMSEs for the normal distribution. As expected, the RMSEs consistently decreased from 0.79 to close to zero for the normal distribution conditions. In contrast, RMSEs increased for the EN and MN distribution conditions as the sample size increased from 100 to 500, then decreased for n = 750 in essentially parallel ways. The RMSEs then increased for both EN and MN distribution conditions in similar ways with the RMSE staying close to 1.30 for the MN distribution conditions, and between 3.20 and 3.48 for the EN distribution conditions.

Test-Level BIAS. As with the RMSE, the test-level BIAS for the *b2*-location parameter differed across the three distributions (Figure 6). And as with the RMSE, the BIAS was greatest for the EN distribution, followed in by the MN distribution and the normal distribution conditions. *b2*-parameters were overestimated for the n = 100 and n = 250 sample sizes for the normal distribution

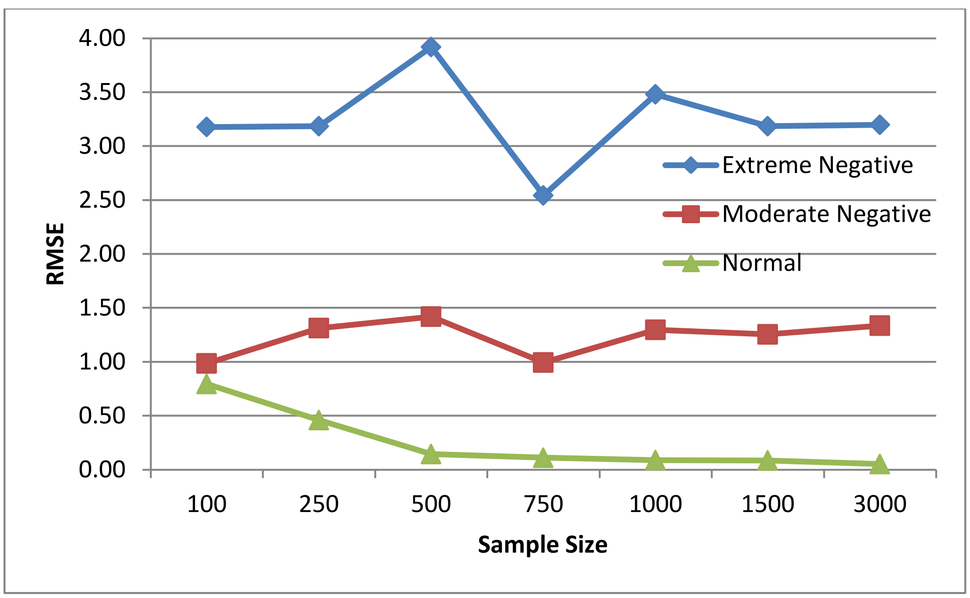

*Figure 8.*RMSEs of ‘true’ and estimated *b2*-parameters by condition.

overestimated for the n = 100 and n = 250 sample sizes for the normal distribution
conditions, and from n = 500 as sample size increased BIAS was essentially zero.

For the EN and MN distribution conditions BIAS increased as the sample size increased from 100 to 500, then decreased for n = 750 in the same way. BIAS then increased for both EN and MN distribution conditions, more so for the EN distribution at n = 1,000. After this point, values for the two distributions diverged from one other, with the BIAS varying between 1.24 and 1.33 for the MN distribution and between 3.05 and 3.26 for the EN distribution conditions.

b3-parameters

RMSE. As shown in Figure 7, the RMSEs for the *b3*-location parameter differed across the three distributions differently than both the *b1*- and *b2*-parameters. However, as with the other *b*-parameters, RMSEs for the EN

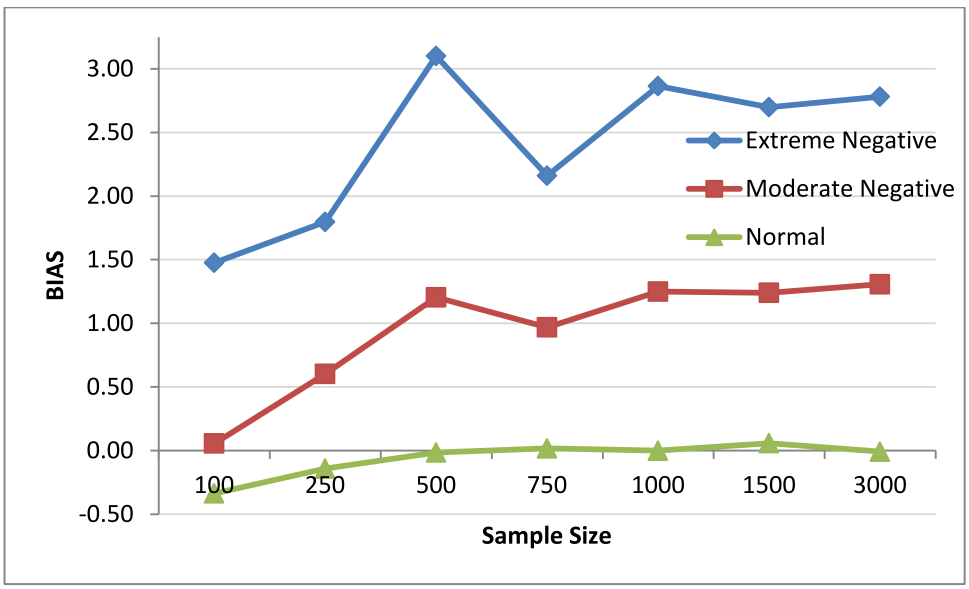

*Figure 9.*BIAS of 'true' and estimated *b2*-parameters by condition.

distribution conditions are all larger than RMSEs for the MN distribution conditions which are all larger than the RMSEs for the normal distribution. As in all cases, the RMSEs consistently decreased from 2.55 to close to zero for the normal distribution conditions.

In contrast, RMSEs for the MN distribution conditions also steadily

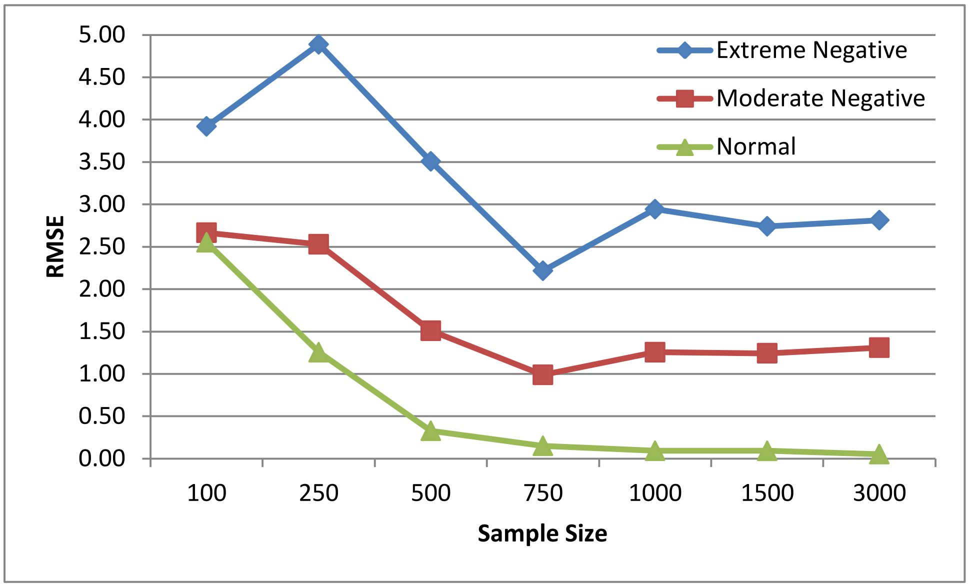

*Figure 10.*RMSEs of 'true' and estimated *b3*-parameters by condition.

decreased from 2.67 at n = 100 to 0.99 at n = 750 and then increased slightly to 1.31 at n = 3,000. In contrast, for the EN distribution conditions the RMSE increased from 3.92 at n = 100 to 4.89 at n = 250 and then decreased again to 2.22 at n = 750. At that point, the RMSE varied between 2.78 and 2.94 for the remaining EN distribution conditions.

Test-Level BIAS. As with the RMSE, the test-level BIAS for the *b3*-location parameter differed across the three distributions (Figure 8). And as with the RMSE, the BIAS was greatest for the EN distribution, followed in by the MN distribution and the normal distribution conditions. *b3*-parameters were overestimated for the n = 100 and n = 250 sample sizes for the normal distribution conditions, and from n = 500 as sample size increased BIAS was essentially zero. For the EN and MN distribution conditions BIAS increased as the sample size increased from 100 to 500, then decreased for n = 750 in the same way. BIAS then increased for both EN and MN distribution conditions, more so for the

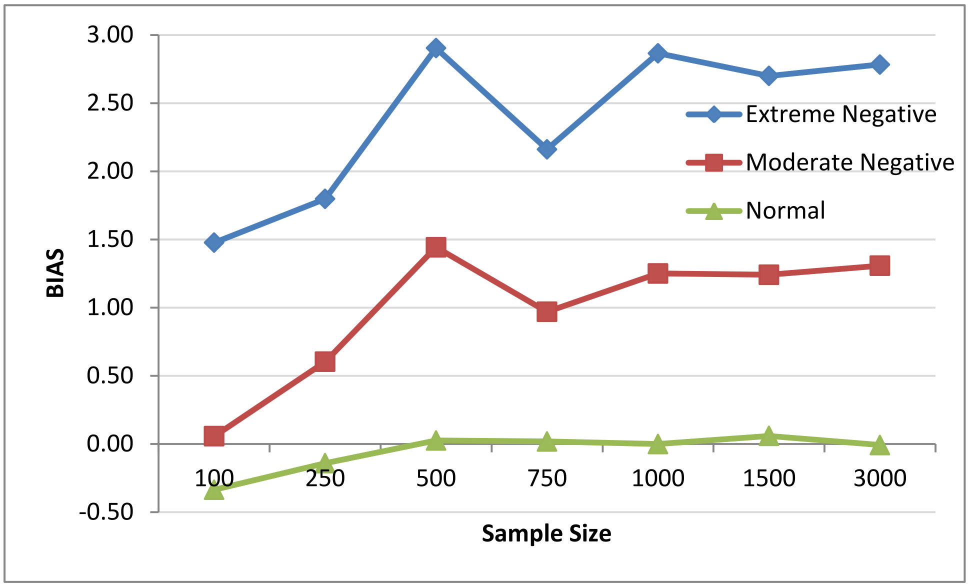

*Figure 11.*BIAS of 'true' and estimated *b3*-parameters by condition.

EN distribution at n = 1,000. After this point, values for the two distributions levelled out, with the BIAS varying between 1.24 and 1.31 for the MN distribution and between 2.70 and 2.86 for the EN distribution conditions.

b4-parameters

RMSE. As shown in Figure 7, the RMSEs for the *b4*-location parameter differed across the three distributions in a similar pattern to the *b3*-parameters. As with the other *b*-parameters, RMSEs for the EN distribution conditions are all larger than RMSEs for the MN distribution conditions which are all larger than the RMSEs for the normal distribution. And as in all cases, the RMSEs consistently decreased from 2.28 to close to zero for the normal distribution conditions.

In contrast, RMSEs for the MN distribution conditions steadily decreased from 2.38 at n = 100 to 1.02 at n = 750 and then increased slightly to 1.27 at n = 3,000. Similarly, for the EN distribution conditions the RMSE increased from

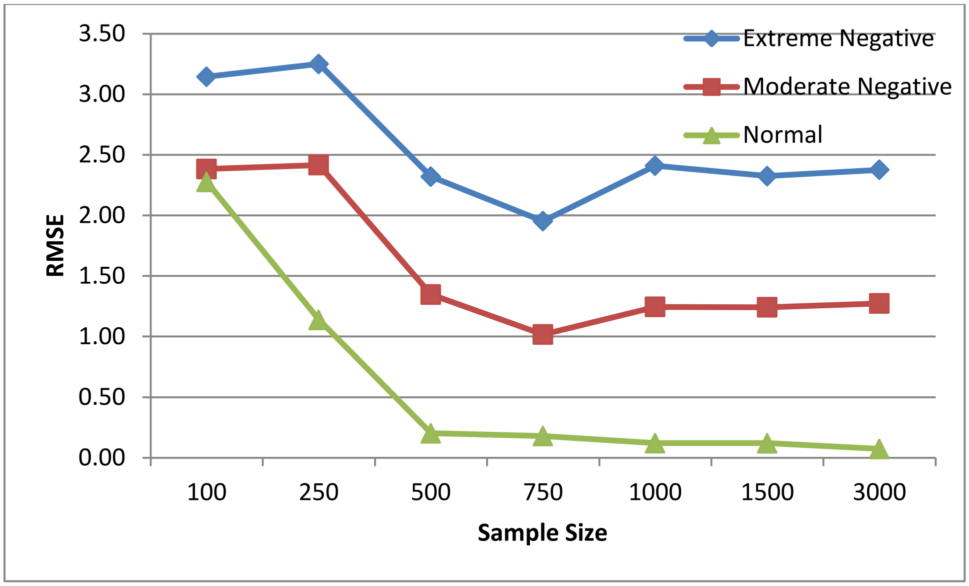

*Figure 12.*RMSEs of 'true' and estimated *b4*-parameters by condition.

3.14 at n = 100 to 3.25 at n = 250 and then decreased again to 1.95 at n = 750. At that point, the RMSE varied between 2.32 and 2.41 for the remaining EN distribution conditions.

Test-Level BIAS. As with the RMSE, the test-level BIAS for the *b4*-location parameter differed across the three distributions (Figure 8) in similar ways to the *b3*-parameter. And as with the RMSE, the BIAS was greatest for the EN distribution, followed in by the MN distribution and the normal distribution conditions. *b4*-parameters were overestimated at n = 100 for the normal distribution conditions, and from n = 250 as sample size increased BIAS was essentially zero.

For the EN and MN distribution conditions BIAS increased as the sample size increased from 100 to 500, then decreased for n = 750 in the same way. BIAS then increased for both EN and MN distribution conditions, more so for the EN distribution at n = 1,000. After this point, values for the two distributions

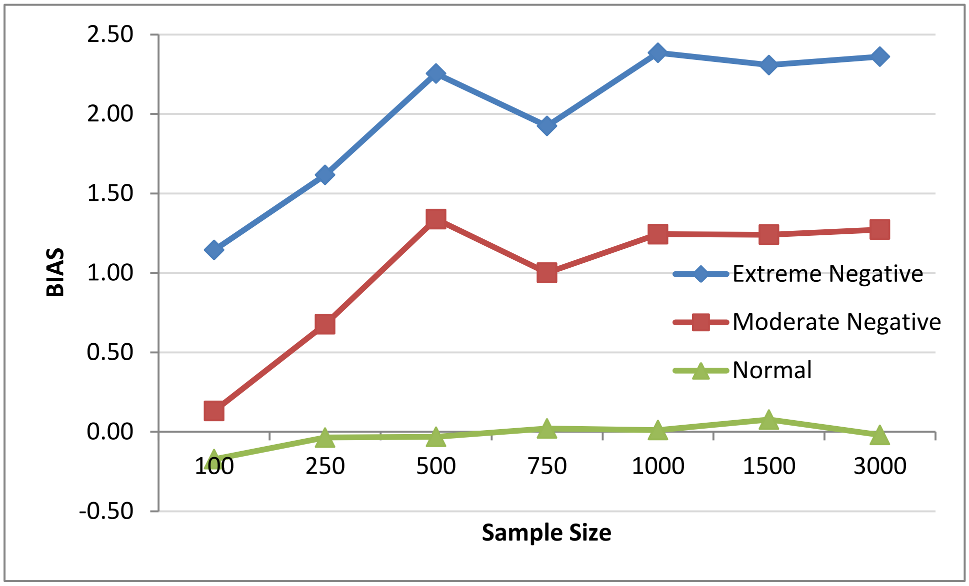

*Figure 13.*BIAS of 'true' and estimated *b4*-parameters by condition.

levelled out, with the BIAS varying between 1.24 and 1.27 for the MN distribution and between 2.31 and 2.38 for the EN distribution conditions.

Theta (θ)

RMSE. Compared to the item parameters at the test level and as shown in Figure 11, the values of the RMSEs for θ are much less variable and except for n = 1,000 and, particularly, n = 3,000 essentially equal for the normal, MN, and EN distribution conditions. The values ranged from 0.33 parameter units for sample size 100 to 0.42 for n = 1,000. However, while the RMSE stayed the same for n= 3,000, the RMSE increased to 0.80 units for the EN distribution and, unexpectedly, to 1.22 units for the MN distribution.

BIAS. In contrast to the RMSEs for θ, the BIAS in the θ estimates were more variable across the seven sample sizes and three distributions Figure 12). Whereas BIAS for the normal distribution was, with the exception of n = 750,

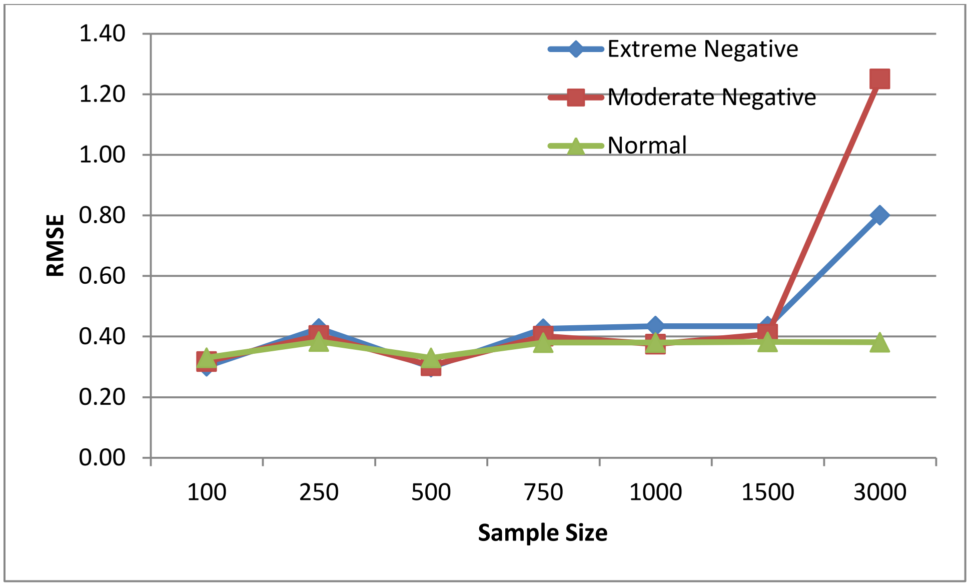

*Figure 14.*RMSEs of 'true' and estimated θ by condition.

essentially unchanged (0.20 units), the patterns of BIAS for the MN and EN distributions varied across the sample sizes. For example, the BIAS for the MN distribution was greater for n = 100 and n = 500 (0.15 vs. 0.05 and 0.10 vs. 0.0, respectively). And the BIAS was essentially equal for n = 250 (0.22), 750 (0.22) and 1,500 (0.22). At n = 1,000, for the EN distribution condition BIAS was greater than for the MN distribution (0.26 vs. 0.23) but at n = 3,000 BIAS for the MN distribution condition was much greater than for the EN condition (0.27 vs. 0.96). With one exception (bias =0 for EN, n = 500) θ was consistently underestimated.

Summary

In general, and as expected, the normal distribution conditions produced better test-level results than either skewed distribution across the seven sample sizes (see Appendix D). Additionally, aside from θ estimates, the EN distribution conditions produced the poorest results overall. Recovery of the *a*-parameters

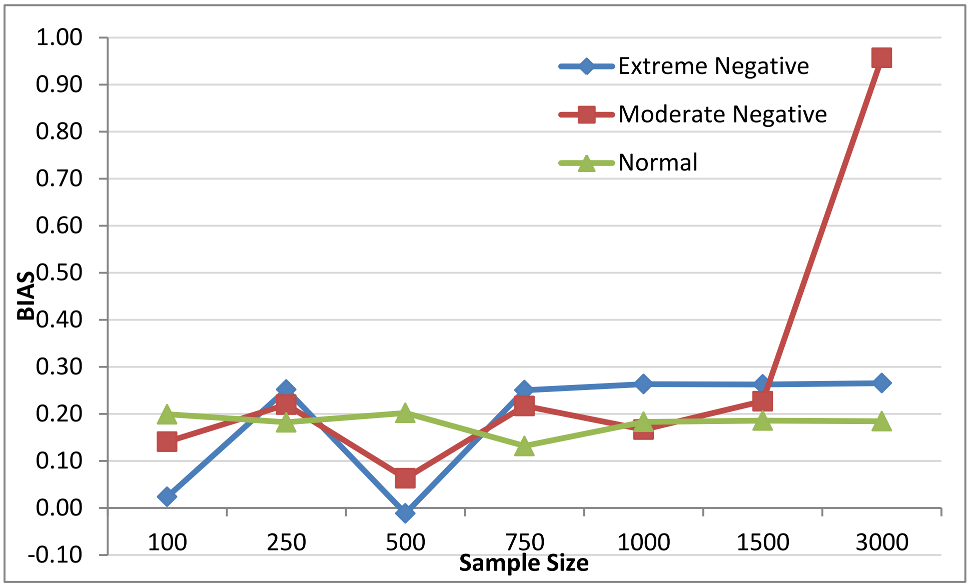

*Figure 15.*BIAS of 'true' and estimated θ by condition.

showed the most consistent improvement as the sample size increased across all distribution conditions, and *b3*- and *b4*-location parameters were more accurately recovered than *b1*- and *b2*- location parameters. The RMSEs for the normal and MN distribution conditions were comparable across sample sizes and all three distribution conditions were comparable when *n* was small. The BIAS results revealed that for the EN and MN distribution conditions, locations parameters were in general overestimated. In the case of the *a*-parameters, they tended to be overestimated at small sample sizes and underestimated with larger sample sizes. As with the RMSEs, the test-level BIAS results for normal and MN were comparable across sample size and for n = 100 all three distribution conditions were comparable.

Chapter 5: Results and Discussion - Item Level Analysis

As indicated in the previous chapter, the results at the item level are presented in this chapter. As in Chapter 4, the results are presented for each of the parameters separately. However, the results in this chapter include the item level BIAS and standard error of the BIAS. The chapter also concludes with general comments across conditions.

a-parameters

Tables 5, 6 and 7 contain the item BIAS and standard error of the item BIAS for the *a*-parameters across all sample sizes for the normal distribution, MN, and EN conditions. As with the test level results, unexpected results were obtained for n = 500 for all three distributions. While the largest BIAS for all items and the three distributions was for n = 100, the BIAS was smallest for the EN distribution and more similar for the MN and normal distributions.

There was less BIAS at n = 250 than at n = 100, and greater BIAS was observed for the EN distribution followed in turn by the MN distribution and the normal distribution conditions. And while generally the size of BIAS decreased as the sample size increased for the MN and normal distributions, BIAS increased as sample size increased for the EN distribution conditions. Further, the decrease noted for the MN and normal distribution conditions was greater for the normal distribution than for the MN distribution. BIAS was less than or equal to 0.05 with three exceptions for the normal distribution, $n \geq 750$, nine exceptions for the MN distribution, $n \geq 1,000$, and for no items for the EN distribution.

Further inspection of the full set of BIAS values reveals that the amount of BIAS was also dependent upon item: larger ‘true’ *a*-parameters tended to result in greater bias. However, while the standard error of the BIAS generally decreased as the sample size increased across the 20 items for all three distribution conditions, the standard errors tended to be close in value or larger than their corresponding bias except for the n = 100 for the MN and normal distribution conditions. Consequently, when the value of the BIAS was divided by it’s standard error for $n \geq 250$, the results suggest that the BIAS values were not significantly different from zero for these two distributions. In contrast the standard errors for the BIAS across EN distribution conditions tended to be less that their corresponding BIAS, resulting in the ratio of the BIAS to it’s standard error being large, suggesting that the BIAS was significantly different from zero for these conditions.

Table 5

Item-Level Bias for a-parameters Under the Normal Conditions

	n=100	**n=250**	**n=500**	**n=750**	**n=1000**	**n=1500**	**n=3000**
Item 1 Bias	-0.623	-0.039	-0.440	-0.005	-0.016	-0.019	0.005
(S.E.)	(0.281)	(0.135)	(0.109)	(0.082)	(0.070)	(0.058)	(0.040)
Item 2 Bias	-0.501	-0.038	-0.357	-0.005	-0.009	-0.016	0.008
(S.E.)	(0.242)	(0.136)	(0.100)	(0.078)	(0.063)	(0.055)	(0.038)
Item 3 Bias	-2.115	-0.124	-1.471	-0.030	-0.059	-0.050	0.020
(S.E.)	(0.898)	(0.252)	(0.290)	(0.143)	(0.126)	(0.097)	(0.070)
Item 4 Bias	-0.527	-0.037	-0.383	-0.004	-0.013	-0.018	0.008
(S.E.)	(0.244)	(0.131)	(0.097)	(0.074)	(0.066)	(0.053)	(0.037)
Item 5 Bias	-1.372	-0.073	-0.960	-0.020	-0.034	-0.037	0.017
(S.E.)	(0.472)	(0.183)	(0.174)	(0.101)	(0.089)	(0.072)	(0.049)
Item 6 Bias	-1.616	-0.113	-1.091	-0.021	-0.040	-0.048	0.019
(S.E.)	(0.592)	(0.211)	(0.213)	(0.116)	(0.096)	(0.079)	(0.054)
Item 7 Bias	-0.491	-0.032	-0.375	-0.007	-0.012	-0.014	0.008
(S.E.)	(0.258)	(0.133)	(0.100)	(0.075)	(0.062)	(0.052)	(0.039)
Item 8 Bias	-1.074	-0.071	-0.746	-0.008	-0.026	-0.029	0.013
(S.E.)	(0.375)	(0.169)	(0.143)	(0.096)	(0.082)	(0.068)	(0.046)
Item 9 Bias	-1.164	-0.084	-0.818	-0.002	-0.027	-0.029	0.010
(S.E.)	(0.391)	(0.175)	(0.158)	(0.093)	(0.083)	(0.068)	(0.048)
Item 10 Bias	-1.243	-0.063	-0.919	-0.011	-0.034	-0.038	0.015
(S.E.)	(0.569)	(0.207)	(0.194)	(0.103)	(0.095)	(0.076)	(0.052)
Item 11 Bias	-2.226	-0.113	-1.427	-0.021	-0.060	-0.063	0.026
(S.E.)	(0.945)	(0.267)	(0.288)	(0.142)	(0.126)	(0.100)	(0.074)
Item 12 Bias	-0.943	-0.018	-0.720	-0.006	-0.028	-0.032	0.012
(S.E.)	(0.447)	(0.186)	(0.152)	(0.091)	(0.080)	(0.064)	(0.043)
Item 13 Bias	-2.001	-0.127	-1.518	-0.010	-0.067	-0.050	0.018
(S.E.)	(0.945)	(0.315)	(0.304)	(0.157)	(0.125)	(0.108)	(0.074)
Item 14 Bias	-0.612	-0.037	-0.496	-0.004	-0.019	-0.020	0.007
(S.E.)	(0.357)	(0.160)	(0.114)	(0.084)	(0.068)	(0.060)	(0.042)
Item 15 Bias	-1.551	-0.086	-1.083	-0.014	-0.036	-0.049	0.019
(S.E.)	(0.655)	(0.234)	(0.227)	(0.126)	(0.119)	(0.097)	(0.063)
Item 16 Bias	-0.356	-0.023	-0.243	-0.006	-0.009	-0.012	0.004
(S.E.)	(0.218)	(0.132)	(0.090)	(0.073)	(0.063)	(0.052)	(0.037)
Item 17 Bias	-1.276	-0.066	-0.909	-0.012	-0.032	-0.041	0.016
(S.E.)	(0.433)	(0.181)	(0.164)	(0.098)	(0.083)	(0.073)	(0.049)
Item 18 Bias	-1.587	-0.112	-1.125	-0.014	-0.044	-0.046	0.019
(S.E.)	(0.522)	(0.193)	(0.190)	(0.111)	(0.096)	(0.077)	(0.052)
Item 19 Bias	-1.620	-0.090	-1.128	-0.017	-0.045	-0.045	0.017
(S.E.)	(0.545)	(0.193)	(0.190)	(0.109)	(0.096)	(0.078)	(0.056)
Item 20 Bias	-2.229	-0.119	-1.479	-0.024	-0.057	-0.065	0.025
(S.E.)	(0.930)	(0.266)	(0.294)	(0.146)	(0.131)	(0.105)	(0.076)

Table 6

Item Bias for a-parameters Across Moderate Negative Conditions

	n=100	n=250	n=500	n=750	n=1000	n=1500	n=3000
Item 1 Bias	-0.698	-0.099	-0.374	-0.027	0.014	0.008	0.044
(S.E.)	(0.280)	(0.140)	(0.104)	(0.080)	(0.067)	(0.057)	(0.040)
Item 2 Bias	-0.566	-0.089	-0.295	-0.016	0.014	0.013	0.035
(S.E.)	(0.235)	(0.131)	(0.098)	(0.075)	(0.065)	(0.053)	(0.038)
Item 3 Bias	-2.669	-0.399	-1.281	-0.126	0.003	0.000	0.103
(S.E.)	(1.054)	(0.295)	(0.311)	(0.147)	(0.132)	(0.101)	(0.071)
Item 4 Bias	-0.627	-0.095	-0.341	-0.033	0.009	0.006	0.033
(S.E.)	(0.260)	(0.141)	(0.110)	(0.078)	(0.069)	(0.056)	(0.040)
Item 5 Bias	-1.588	-0.271	-0.878	-0.084	-0.003	-0.003	0.063
(S.E.)	(0.482)	(0.196)	(0.180)	(0.106)	(0.090)	(0.074)	(0.054)
Item 6 Bias	-1.878	-0.297	-0.985	-0.091	0.005	-0.005	0.080
(S.E.)	(0.716)	(0.223)	(0.252)	(0.123)	(0.103)	(0.092)	(0.062)
Item 7 Bias	-0.564	-0.086	-0.325	-0.028	0.008	0.011	0.034
(S.E.)	(0.273)	(0.141)	(0.104)	(0.081)	(0.065)	(0.055)	(0.039)
Item 8 Bias	-1.244	-0.210	-0.699	-0.071	-0.004	-0.008	0.046
(S.E.)	(0.407)	(0.181)	(0.166)	(0.092)	(0.087)	(0.069)	(0.049)
Item 9 Bias	-1.433	-0.230	-0.780	-0.086	-0.019	-0.020	0.040
(S.E.)	(0.482)	(0.184)	(0.191)	(0.104)	(0.093)	(0.074)	(0.053)
Item 10 Bias	-1.531	-0.225	-0.876	-0.097	-0.022	-0.035	0.033
(S.E.)	(0.669)	(0.244)	(0.247)	(0.124)	(0.107)	(0.088)	(0.061)
Item 11 Bias	-2.387	-0.379	-1.288	-0.104	0.003	0.002	0.104
(S.E.)	(0.880)	(0.278)	(0.264)	(0.139)	(0.125)	(0.096)	(0.073)
Item 12 Bias	-1.078	-0.138	-0.661	-0.063	-0.001	0.003	0.052
(S.E.)	(0.472)	(0.223)	(0.154)	(0.093)	(0.082)	(0.065)	(0.047)
Item 13 Bias	-2.560	-0.463	-1.463	-0.224	-0.107	-0.095	0.009
(S.E.)	(1.284)	(0.413)	(0.433)	(0.193)	(0.160)	(0.135)	(0.092)
Item 14 Bias	-0.702	-0.093	-0.463	-0.042	0.001	-0.003	0.035
(S.E.)	(0.396)	(0.203)	(0.137)	(0.087)	(0.074)	(0.064)	(0.045)
Item 15 Bias	-1.828	-0.319	-1.035	-0.120	-0.029	-0.033	0.039
(S.E.)	(0.696)	(0.249)	(0.233)	(0.132)	(0.114)	(0.091)	(0.065)
Item 16 Bias	-0.396	-0.056	-0.217	-0.015	0.005	0.003	0.024
(S.E.)	(0.217)	(0.132)	(0.098)	(0.076)	(0.068)	(0.051)	(0.038)
Item 17 Bias	-1.518	-0.250	-0.855	-0.081	-0.017	-0.013	0.053
(S.E.)	(0.491)	(0.195)	(0.186)	(0.107)	(0.090)	(0.075)	(0.053)
Item 18 Bias	-1.875	-0.328	-1.069	-0.111	-0.023	-0.027	0.055
(S.E.)	(0.610)	(0.219)	(0.222)	(0.117)	(0.105)	(0.082)	(0.062)
Item 19 Bias	-1.942	-0.317	-1.045	-0.105	-0.009	-0.014	0.068
(S.E.)	(0.603)	(0.217)	(0.218)	(0.116)	(0.102)	(0.083)	(0.057)
Item 20 Bias	-2.428	-0.431	-1.371	-0.128	-0.015	-0.022	0.077
(S.E.)	(0.836)	(0.305)	(0.287)	(0.150)	(0.127)	(0.102)	(0.074)

Table 7

Item Bias for a-parameters Across Extreme Negative Conditions

	n=100	n=250	n=500	n=750	n=1000	n=1500	n=3000
Item 1 Bias	-0.446	0.153	-0.107	0.166	0.271	0.250	0.270
(S.E.)	(0.310)	(0.147)	(0.111)	(0.080)	(0.075)	(0.058)	(0.043)
Item 2 Bias	-0.371	0.131	-0.082	0.140	0.223	0.204	0.222
(S.E.)	(0.263)	(0.133)	(0.098)	(0.075)	(0.069)	(0.056)	(0.039)
Item 3 Bias	-1.821	0.421	-0.490	0.484	0.823	0.751	0.821
(S.E.)	(1.039)	(0.257)	(0.297)	(0.147)	(0.127)	(0.108)	(0.071)
Item 4 Bias	-0.443	0.127	-0.116	0.135	0.225	0.211	0.230
(S.E.)	(0.291)	(0.139)	(0.121)	(0.084)	(0.076)	(0.060)	(0.044)
Item 5 Bias	-1.270	0.260	-0.378	0.306	0.535	0.490	0.532
(S.E.)	(0.590)	(0.195)	(0.180)	(0.108)	(0.098)	(0.079)	(0.052)
Item 6 Bias	-1.426	0.320	-0.388	0.358	0.615	0.560	0.615
(S.E.)	(0.813)	(0.218)	(0.232)	(0.125)	(0.113)	(0.092)	(0.060)
Item 7 Bias	-0.394	0.125	-0.110	0.136	0.222	0.205	0.220
(S.E.)	(0.304)	(0.146)	(0.111)	(0.084)	(0.072)	(0.058)	(0.047)
Item 8 Bias	-1.053	0.196	-0.329	0.232	0.414	0.378	0.411
(S.E.)	(0.516)	(0.181)	(0.179)	(0.103)	(0.096)	(0.076)	(0.053)
Item 9 Bias	-1.156	0.201	-0.393	0.237	0.439	0.401	0.436
(S.E.)	(0.638)	(0.203)	(0.223)	(0.113)	(0.111)	(0.084)	(0.058)
Item 10 Bias	-1.466	0.222	-0.489	0.247	0.466	0.424	0.462
(S.E.)	(0.996)	(0.249)	(0.318)	(0.151)	(0.153)	(0.104)	(0.080)
Item 11 Bias	-1.844	0.418	-0.508	0.466	0.801	0.730	0.799
(S.E.)	(0.873)	(0.253)	(0.250)	(0.133)	(0.122)	(0.099)	(0.066)
Item 12 Bias	-0.808	0.256	-0.270	0.228	0.400	0.370	0.410
(S.E.)	(0.646)	(0.204)	(0.177)	(0.128)	(0.140)	(0.104)	(0.056)
Item 13 Bias	-1.878	0.300	-0.935	0.309	0.656	0.579	0.646
(S.E.)	(1.938)	(0.547)	(0.832)	(0.233)	(0.242)	(0.180)	(0.113)
Item 14 Bias	-0.439	0.177	-0.189	0.164	0.286	0.258	0.286
(S.E.)	(0.555)	(0.201)	(0.152)	(0.094)	(0.094)	(0.070)	(0.048)
Item 15 Bias	-1.629	0.226	-0.549	0.292	0.543	0.491	0.548
(S.E.)	(0.949)	(0.252)	(0.260)	(0.137)	(0.129)	(0.097)	(0.070)
Item 16 Bias	-0.271	0.080	-0.070	0.091	0.152	0.140	0.148
(S.E.)	(0.249)	(0.131)	(0.106)	(0.082)	(0.070)	(0.055)	(0.039)
Item 17 Bias	-1.241	0.232	-0.389	0.282	0.492	0.452	0.491
(S.E.)	(0.620)	(0.198)	(0.187)	(0.111)	(0.101)	(0.082)	(0.054)
Item 18 Bias	-1.589	0.273	-0.498	0.332	0.592	0.529	0.587
(S.E.)	(0.773)	(0.217)	(0.228)	(0.123)	(0.115)	(0.090)	(0.061)
Item 19 Bias	-1.481	0.307	-0.475	0.352	0.622	0.558	0.615
(S.E.)	(0.680)	(0.199)	(0.213)	(0.118)	(0.111)	(0.086)	(0.060)
Item 20 Bias	-2.099	0.364	-0.639	0.442	0.775	0.703	0.778
(S.E.)	(1.018)	(0.263)	(0.282)	(0.144)	(0.130)	(0.106)	(0.066)

b1-parameters

Tables 8, 9 and 10 contain the item BIAS and standard error of the item BIAS for the *b1*-parameters across all sample sizes for the normal distribution, MN, and EN conditions. As with the test level results, unexpected results were obtained for n = 500 for all three distributions. The largest BIAS for all items with the normal distribution conditions was when n = 500. However, in contrast to the *a*-parameters, the MN and EN conditions were more similar to one another and for both distribution condition, BIAS increased as sample size increased.

The greatest amount of BIAS was observed for the EN distribution followed in turn by the MN distribution and the normal distribution conditions. And while generally the size of BIAS decreased as the sample size increased for the normal distributions, BIAS increased as sample size increased for the EN and MN distribution conditions. Further, the increase noted for the MN and EN distribution conditions was more extreme for the EN distribution than for the MN distribution. BIAS was less than or equal to 0.05 with two exceptions for the normal distribution, $n \geq 750$, but for no items for both the EN and MN distribution conditions.

Further inspection of the full set of BIAS values reveals that the amount of BIAS was also dependent upon item: items which had *b1*- and *b2*-parameters very close in value had much larger BIAS, particularly when sample size was small. As with the *a*-parameters, the standard error of the BIAS generally decreased as the sample size increased across the 20 items for all three distribution conditions, and in this case, the standard errors tended to be close in value or larger than their corresponding bias only for the normal distribution conditions where $n \geq 100$. For the MN and EN distribution conditions most standard errors were smaller than the BIAS values.

Consequently, when the value of the BIAS was divided by it's standard error for $n \geq 250$, the results suggest that the BIAS values were not significantly different from zero for the normal distribution conditions. In contrast the standard errors for the BIAS across EN and MN distribution conditions tended to be less that their corresponding BIAS, resulting in the ratio of the BIAS to its standard error being large, suggesting that the BIAS was significantly different from zero for these conditions.

Table 8

Item Bias for b1-parameters Across Normal Conditions

	n=100	**n=250**	**n=500**	**n=750**	**n=1000**	**n=1500**	**n=3000**
Item 1 Bias	0.083	0.048	0.119	0.033	-0.013	0.043	0.007
(S.E.)	(0.328)	(0.200)	(0.115)	(0.123)	(0.102)	(0.080)	(0.060)
Item 2 Bias	0.117	0.056	0.137	0.036	-0.013	0.033	0.011
(S.E.)	(0.301)	(0.296)	(0.139)	(0.165)	(0.133)	(0.114)	(0.081)
Item 3 Bias	0.065	0.037	0.154	0.025	-0.030	0.032	0.009
(S.E.)	(0.159)	(0.098)	(0.077)	(0.059)	(0.050)	(0.040)	(0.028)
Item 4 Bias	0.093	0.051	0.242	0.051	-0.023	0.005	0.033
(S.E.)	(0.508)	(0.494)	(0.231)	(0.278)	(0.239)	(0.179)	(0.135)
Item 5 Bias	0.080	0.042	0.137	0.021	-0.021	0.034	0.007
(S.E.)	(0.161)	(0.116)	(0.085)	(0.067)	(0.058)	(0.047)	(0.035)
Item 6 Bias	0.047	-0.029	0.231	0.015	-0.043	0.004	0.023
(S.E.)	(0.326)	(0.195)	(0.138)	(0.121)	(0.097)	(0.080)	(0.054)
Item 7 Bias	0.085	0.058	0.171	0.033	-0.015	0.030	0.019
(S.E.)	(0.352)	(0.397)	(0.159)	(0.190)	(0.154)	(0.124)	(0.093)
Item 8 Bias	0.093	0.049	0.121	0.024	-0.012	0.036	0.006
(S.E.)	(0.171)	(0.126)	(0.084)	(0.075)	(0.063)	(0.052)	(0.038)
Item 9 Bias	0.060	0.006	0.180	0.031	-0.031	0.022	0.012
(S.E.)	(0.206)	(0.173)	(0.109)	(0.104)	(0.087)	(0.071)	(0.052)
Item 10 Bias	-1.766	-0.855	0.059	0.019	-0.037	0.015	0.019
(S.E.)	(1.982)	(2.273)	(0.794)	(0.243)	(0.088)	(0.071)	(0.054)
Item 11 Bias	0.097	0.059	0.111	0.027	-0.015	0.038	0.005
(S.E.)	(0.138)	(0.080)	(0.065)	(0.045)	(0.039)	(0.031)	(0.022)
Item 12 Bias	0.101	0.022	0.229	0.032	-0.035	0.010	0.020
(S.E.)	(0.331)	(0.259)	(0.150)	(0.141)	(0.115)	(0.093)	(0.070)
Item 13 Bias	0.074	0.036	0.154	0.029	-0.030	0.032	0.008
(S.E.)	(0.161)	(0.101)	(0.076)	(0.055)	(0.047)	(0.041)	(0.026)
Item 14 Bias	0.048	0.017	0.161	0.037	-0.028	0.027	0.013
(S.E.)	(0.292)	(0.334)	(0.126)	(0.143)	(0.112)	(0.095)	(0.069)
Item 15 Bias	0.134	0.069	0.066	0.027	-0.002	0.050	-0.004
(S.E.)	(0.164)	(0.086)	(0.069)	(0.049)	(0.044)	(0.035)	(0.025)
Item 16 Bias	0.142	0.210	0.269	0.077	-0.002	0.033	0.037
(S.E.)	(0.766)	(0.918)	(0.347)	(0.441)	(0.351)	(0.292)	(0.209)
Item 17 Bias	0.088	0.050	0.134	0.025	-0.018	0.034	0.005
(S.E.)	(0.164)	(0.119)	(0.077)	(0.066)	(0.056)	(0.047)	(0.035)
Item 18 Bias	0.064	0.008	0.172	0.025	-0.036	0.024	0.012
(S.E.)	(0.189)	(0.133)	(0.091)	(0.077)	(0.065)	(0.054)	(0.039)
Item 19 Bias	0.071	0.039	0.148	0.026	-0.026	0.030	0.009
(S.E.)	(0.159)	(0.111)	(0.080)	(0.064)	(0.052)	(0.045)	(0.032)
Item 20 Bias	0.102	0.058	0.113	0.026	-0.015	0.036	0.004
(S.E.)	(0.138)	(0.075)	(0.065)	(0.043)	(0.038)	(0.030)	(0.022)

Table 9

Item Bias for b1-parameters Across Moderate Negative Conditions

	n=100	n=250	n=500	n=750	n=1000	n=1500	n=3000
Item 1 Bias	0.363	0.850	1.570	0.956	1.283	1.262	1.386
(S.E.)	(0.381)	(0.302)	(0.208)	(0.169)	(0.182)	(0.147)	(0.121)
Item 2 Bias	0.369	0.826	1.646	0.961	1.311	1.289	1.405
(S.E.)	(0.341)	(0.392)	(0.276)	(0.237)	(0.255)	(0.206)	(0.163)
Item 3 Bias	0.242	0.710	1.603	0.876	1.229	1.222	1.353
(S.E.)	(0.200)	(0.133)	(0.196)	(0.090)	(0.095)	(0.076)	(0.072)
Item 4 Bias	0.263	0.699	1.781	0.865	1.328	1.289	1.473
(S.E.)	(0.560)	(0.579)	(0.473)	(0.361)	(0.388)	(0.301)	(0.241)
Item 5 Bias	0.326	0.762	1.556	0.904	1.240	1.234	1.351
(S.E.)	(0.207)	(0.149)	(0.164)	(0.098)	(0.105)	(0.083)	(0.076)
Item 6 Bias	0.359	0.623	2.584	0.845	1.264	1.225	1.422
(S.E.)	(0.743)	(0.282)	(1.202)	(0.182)	(0.207)	(0.168)	(0.138)
Item 7 Bias	0.321	0.771	1.680	0.926	1.297	1.300	1.419
(S.E.)	(0.409)	(0.450)	(0.338)	(0.284)	(0.279)	(0.238)	(0.175)
Item 8 Bias	0.368	0.802	1.529	0.916	1.243	1.234	1.344
(S.E.)	(0.224)	(0.170)	(0.159)	(0.104)	(0.117)	(0.092)	(0.083)
Item 9 Bias	0.230	0.687	1.641	0.864	1.225	1.213	1.363
(S.E.)	(0.265)	(0.226)	(0.280)	(0.150)	(0.161)	(0.130)	(0.108)
Item 10 Bias	-1.916	-1.804	-1.287	0.435	0.757	1.077	1.349
(S.E.)	(2.100)	(3.146)	(2.559)	(1.667)	(1.806)	(0.941)	(0.193)
Item 11 Bias	0.392	0.819	1.506	0.923	1.235	1.234	1.330
(S.E.)	(0.191)	(0.105)	(0.116)	(0.062)	(0.064)	(0.053)	(0.052)
Item 12 Bias	0.288	0.705	1.787	0.857	1.250	1.252	1.416
(S.E.)	(0.388)	(0.389)	(0.519)	(0.197)	(0.208)	(0.175)	(0.135)
Item 13 Bias	0.245	0.692	1.576	0.858	1.196	1.193	1.319
(S.E.)	(0.202)	(0.134)	(0.180)	(0.089)	(0.095)	(0.076)	(0.073)
Item 14 Bias	0.293	0.699	1.603	0.907	1.257	1.248	1.382
(S.E.)	(0.355)	(0.412)	(0.273)	(0.195)	(0.203)	(0.171)	(0.139)
Item 15 Bias	0.455	0.904	1.448	0.955	1.240	1.232	1.310
(S.E.)	(0.254)	(0.095)	(0.094)	(0.061)	(0.062)	(0.048)	(0.046)
Item 16 Bias	0.354	0.913	1.839	0.963	1.384	1.312	1.505
(S.E.)	(0.810)	(1.051)	(0.655)	(0.588)	(0.635)	(0.455)	(0.357)
Item 17 Bias	0.332	0.779	1.539	0.910	1.230	1.225	1.344
(S.E.)	(0.216)	(0.148)	(0.159)	(0.099)	(0.104)	(0.084)	(0.072)
Item 18 Bias	0.224	0.678	1.664	0.863	1.225	1.204	1.357
(S.E.)	(0.231)	(0.180)	(0.383)	(0.117)	(0.135)	(0.106)	(0.094)
Item 19 Bias	0.279	0.728	1.577	0.888	1.232	1.221	1.350
(S.E.)	(0.208)	(0.145)	(0.178)	(0.093)	(0.103)	(0.083)	(0.072)
Item 20 Bias	0.388	0.806	1.499	0.918	1.230	1.223	1.323
(S.E.)	(0.189)	(0.099)	(0.117)	(0.063)	1.283	(0.054)	(0.051)

Table 10

Item Bias for b1-parameters Across Extreme Negative Conditions

	n=100	**n=250**	**n=500**	**n=750**	**n=1000**	**n=1500**	**n=3000**
Item 1 Bias	1.475	2.628	3.222	2.515	3.580	3.322	3.479
(S.E.)	(1.200)	(0.990)	(0.495)	(0.427)	(0.699)	(0.474)	(0.359)
Item 2 Bias	1.799	2.790	3.414	2.670	3.812	3.500	3.678
(S.E.)	(0.877)	(1.611)	(0.639)	(0.606)	(1.050)	(0.645)	(0.468)
Item 3 Bias	1.945	2.707	4.054	2.507	3.754	3.415	3.552
(S.E.)	(1.260)	(0.893)	(1.190)	(0.302)	(0.779)	(0.496)	(0.304)
Item 4 Bias	1.729	2.594	3.948	2.967	3.965	4.052	4.424
(S.E.)	(2.239)	(3.260)	(1.432)	(1.143)	(2.956)	(1.498)	(0.727)
Item 5 Bias	1.299	2.428	3.158	2.438	3.476	3.249	3.406
(S.E.)	(1.239)	(0.500)	(0.704)	(0.272)	(0.407)	(0.300)	(0.206)
Item 6 Bias	1.315	2.975	1.341	3.095	3.420	3.809	4.048
(S.E.)	(2.390)	(2.782)	(2.691)	(1.078)	(4.454)	(2.580)	(1.662)
Item 7 Bias	1.793	2.742	3.521	2.730	3.934	3.618	3.779
(S.E.)	(1.136)	(1.762)	(0.738)	(0.721)	(1.010)	(0.699)	(0.548)
Item 8 Bias	1.476	2.399	2.943	2.380	3.348	3.132	3.277
(S.E.)	(0.582)	(0.475)	(0.356)	(0.287)	(0.444)	(0.309)	(0.223)
Item 9 Bias	1.838	2.606	3.774	2.560	3.799	3.475	3.664
(S.E.)	(2.153)	(0.886)	(1.862)	(0.391)	(0.655)	(0.435)	(0.315)
Item 10 Bias	-3.334	-4.305	-2.506	-1.350	-1.613	-0.231	2.019
(S.E.)	(4.799)	(5.284)	(4.602)	(4.321)	(5.258)	(4.959)	(3.921)
Item 11 Bias	1.490	2.319	2.867	2.316	3.229	3.024	3.182
(S.E.)	(0.394)	(0.297)	(0.319)	(0.175)	(0.291)	(0.208)	(0.142)
Item 12 Bias	2.133	2.806	4.195	2.758	4.122	3.824	4.054
(S.E.)	(3.543)	(2.075)	(1.571)	(0.576)	(0.896)	(0.636)	(0.410)
Item 13 Bias	1.740	2.430	3.745	2.337	3.343	3.113	3.239
(S.E.)	(1.013)	(0.836)	(1.111)	(0.307)	(0.750)	(0.454)	(0.254)
Item 14 Bias	1.417	2.451	3.277	2.575	3.757	3.441	3.649
(S.E.)	(3.410)	(1.786)	(0.645)	(0.513)	(0.835)	(0.564)	(0.409)
Item 15 Bias	1.285	2.159	2.627	2.159	2.873	2.723	2.843
(S.E.)	(0.935)	(0.327)	(0.183)	(0.155)	(0.247)	(0.170)	(0.115)
Item 16 Bias	1.257	2.304	3.701	3.002	3.247	4.097	4.392
(S.E.)	(3.760)	(4.355)	(2.636)	(2.124)	(4.769)	(2.171)	(1.127)
Item 17 Bias	1.511	2.398	3.052	2.406	3.401	3.179	3.333
(S.E.)	(0.611)	(0.446)	(0.484)	(0.271)	(0.402)	(0.299)	(0.211)
Item 18 Bias	2.302	2.725	4.237	2.540	3.721	3.413	3.603
(S.E.)	(1.584)	(1.014)	(1.315)	(0.351)	(0.971)	(0.400)	(0.290)
Item 19 Bias	1.705	2.511	3.578	2.454	3.546	3.280	3.468
(S.E.)	(0.952)	(0.590)	(1.143)	(0.272)	(0.444)	(0.300)	(0.233)
Item 20 Bias	1.472	2.275	2.837	2.294	3.180	2.978	3.132
(S.E.)	(0.417)	(0.267)	(0.298)	(0.174)	(0.290)	(0.213)	(0.139)

b2-parameters

Tables 11, 12 and 13 contain the item BIAS and standard error of the item BIAS for the *b2*-parameters across all sample sizes for the normal distribution, MN, and EN conditions. As with the test level results, unexpected results were obtained for n = 500 for all three distributions. The largest BIAS for all items with the normal distribution conditions was with n = 500. As with the *b1*-parameters, the MN and EN conditions were more similar to one another and for both distribution condition, and BIAS increased as sample size increased.

The greatest amount of BIAS was observed for the EN distribution followed in turn by the MN distribution and the normal distribution conditions. And while generally the size of BIAS decreased as the sample size increased for the normal distributions, BIAS actually increased as sample size increased for the EN and MN distribution conditions. Further, the increase noted for the MN and EN distribution conditions was more extreme for the EN distribution than for the MN distribution but less extreme than for the *b1*-parameters. BIAS was less than or equal to 0.05 with eight exceptions for the normal distribution, $n \geq 750$, but for no items for both the EN and MN distribution conditions.

Further inspection of the full set of BIAS values reveals that the amount of BIAS was also dependent upon item: items which had *b1*- and *b2*-parameters very close in value had much larger BIAS, particularly when sample size was small. As with the *a*- and *b1*-parameters, the standard error of the BIAS generally decreased as the sample size increased across the 20 items for all three distribution conditions, and in this case, the standard errors tended to be close in value or larger than their corresponding bias only for the normal distribution conditions where $n \geq 100$. For the MN and EN distribution conditions most standard errors were smaller than the BIAS values.

Consequently, when the value of the BIAS was divided by it's standard error for $n \geq 250$, the results suggest that the BIAS values were not significantly different from zero for the normal distribution conditions. In contrast the standard errors for the BIAS across EN and MN distribution conditions tended to be less that their corresponding BIAS, resulting in the ratio of the BIAS to its standard error being large, suggesting that the BIAS was significantly different from zero for these conditions.

Table 11

Item Bias for b2-parameters Across Normal Conditions

	n=100	n=250	n=500	n=750	n=1000	n=1500	n=3000
Item 1 Bias	0.033	0.053	0.090	0.031	-0.007	0.049	0.002
(S.E.)	(1.122)	(0.179)	(0.101)	(0.107)	(0.093)	(0.073)	(0.052)
Item 2 Bias	0.157	0.069	-0.007	0.015	-0.003	0.067	-0.020
(S.E.)	(0.309)	(0.319)	(0.154)	(0.195)	(0.158)	(0.122)	(0.095)
Item 3 Bias	0.128	0.064	0.075	0.026	-0.005	0.047	0.000
(S.E.)	(0.135)	(0.069)	(0.064)	(0.041)	(0.036)	(0.029)	(0.021)
Item 4 Bias	0.108	0.060	0.121	0.036	-0.011	0.039	0.005
(S.E.)	(0.237)	(0.231)	(0.116)	(0.134)	(0.115)	(0.088)	(0.066)
Item 5 Bias	0.104	0.055	0.103	0.026	-0.014	0.042	0.002
(S.E.)	(0.150)	(0.097)	(0.072)	(0.055)	(0.048)	(0.040)	(0.030)
Item 6 Bias	0.056	0.010	0.174	0.018	-0.033	0.020	0.013
(S.E.)	(0.184)	(0.135)	(0.091)	(0.078)	(0.065)	(0.054)	(0.038)
Item 7 Bias	-0.783	-0.237	0.017	0.035	0.005	0.062	-0.017
(S.E.)	(3.336)	(4.877)	(0.129)	(0.157)	(0.132)	(0.102)	(0.077)
Item 8 Bias	0.117	0.066	0.078	0.024	-0.003	0.048	-0.001
(S.E.)	(0.153)	(0.113)	(0.076)	(0.068)	(0.056)	(0.044)	(0.033)
Item 9 Bias	0.085	0.036	0.135	0.032	-0.020	0.033	0.007
(S.E.)	(0.175)	(0.123)	(0.086)	(0.073)	(0.063)	(0.051)	(0.038)
Item 10 Bias	1.688	0.749	0.299	0.031	-0.036	0.016	0.018
(S.E.)	(1.836)	(2.008)	(0.673)	(0.217)	(0.085)	(0.068)	(0.052)
Item 11 Bias	0.153	0.094	0.023	0.029	0.008	0.067	-0.012
(S.E.)	(0.154)	(0.078)	(0.069)	(0.049)	(0.043)	(0.034)	(0.025)
Item 12 Bias	0.060	-0.042	0.066	0.018	-0.005	0.052	-0.003
(S.E.)	(0.187)	(0.275)	(0.081)	(0.097)	(0.065)	(0.046)	(0.035)
Item 13 Bias	-0.399	-0.051	0.074	0.024	-0.004	0.047	0.000
(S.E.)	(1.218)	(0.784)	(0.062)	(0.039)	(0.035)	(0.029)	(0.020)
Item 14 Bias	-0.310	-0.096	0.060	0.023	-0.007	0.051	-0.005
(S.E.)	(1.034)	(1.038)	(0.092)	(0.099)	(0.084)	(0.070)	(0.048)
Item 15 Bias	0.143	0.077	0.053	0.027	0.001	0.055	-0.005
(S.E.)	(0.163)	(0.087)	(0.072)	(0.052)	(0.045)	(0.035)	(0.027)
Item 16 Bias	0.141	0.130	0.181	0.058	-0.006	0.043	0.020
(S.E.)	(0.519)	(0.570)	(0.225)	(0.296)	(0.234)	(0.195)	(0.136)
Item 17 Bias	0.119	0.062	0.082	0.023	-0.007	0.048	-0.001
(S.E.)	(0.152)	(0.095)	(0.070)	(0.055)	(0.048)	(0.039)	(0.028)
Item 18 Bias	0.095	0.050	0.118	0.025	-0.018	0.038	0.004
(S.E.)	(0.152)	(0.089)	(0.072)	(0.054)	(0.045)	(0.037)	(0.026)
Item 19 Bias	0.115	0.068	0.084	0.027	-0.007	0.048	-0.001
(S.E.)	(0.138)	(0.082)	(0.067)	(0.047)	(0.041)	(0.034)	(0.023)
Item 20 Bias	0.147	0.093	0.022	0.029	0.008	0.069	-0.011
(S.E.)	(0.150)	(0.084)	(0.071)	(0.049)	(0.043)	(0.034)	(0.025)

Table 12

Item Bias for b2-parameters Across Moderate Negative Conditions

	n=100	n=250	n=500	n=750	n=1000	n=1500	n=3000
Item 1 Bias	0.246	0.871	1.517	0.967	1.272	1.256	1.359
(S.E.)	(1.413)	(0.582)	(0.172)	(0.146)	(0.151)	(0.124)	(0.101)
Item 2 Bias	0.527	0.990	1.362	0.981	1.244	1.236	1.279
(S.E.)	(0.244)	(0.208)	(0.125)	(0.125)	(0.116)	(0.091)	(0.073)
Item 3 Bias	0.419	0.879	1.471	0.945	1.241	1.234	1.317
(S.E.)	(0.290)	(0.084)	(0.094)	(0.053)	(0.052)	(0.042)	(0.042)
Item 4 Bias	0.399	0.855	1.551	0.934	1.278	1.258	1.368
(S.E.)	(0.289)	(0.313)	(0.242)	(0.193)	(0.206)	(0.166)	(0.130)
Item 5 Bias	0.378	0.824	1.512	0.929	1.242	1.237	1.335
(S.E.)	(0.189)	(0.122)	(0.127)	(0.081)	(0.080)	(0.066)	(0.063)
Item 6 Bias	0.234	0.687	1.697	0.873	1.246	1.227	1.383
(S.E.)	(0.242)	(0.180)	(0.589)	(0.120)	(0.130)	(0.107)	(0.093)
Item 7 Bias	-0.351	0.578	1.407	0.979	1.250	1.246	1.296
(S.E.)	(3.311)	(4.785)	(0.131)	(0.120)	(0.117)	(0.094)	(0.075)
Item 8 Bias	0.433	0.876	1.474	0.945	1.245	1.238	1.324
(S.E.)	(0.195)	(0.133)	(0.118)	(0.079)	(0.090)	(0.071)	(0.062)
Item 9 Bias	0.295	0.772	1.551	0.896	1.233	1.219	1.342
(S.E.)	(0.415)	(0.171)	(0.177)	(0.112)	(0.117)	(0.094)	(0.082)
Item 10 Bias	2.003	2.502	3.371	1.146	1.532	1.277	1.347
(S.E.)	(1.866)	(2.384)	(1.963)	(1.231)	(1.279)	(0.694)	(0.187)
Item 11 Bias	0.531	0.987	1.398	0.996	1.255	1.250	1.305
(S.E.)	(0.168)	(0.072)	(0.069)	(0.042)	(0.040)	(0.033)	(0.032)
Item 12 Bias	0.399	0.817	1.462	0.955	1.246	1.247	1.327
(S.E.)	(0.211)	(0.265)	(0.116)	(0.086)	(0.084)	(0.070)	(0.059)
Item 13 Bias	-0.310	0.531	1.405	0.936	1.224	1.220	1.301
(S.E.)	(1.319)	(1.307)	(0.498)	(0.054)	(0.051)	(0.042)	(0.041)
Item 14 Bias	-0.042	0.317	1.454	0.958	1.249	1.246	1.318
(S.E.)	(1.449)	(6.147)	(0.136)	(0.107)	(0.111)	(0.092)	(0.074)
Item 15 Bias	0.492	0.932	1.431	0.967	1.242	1.237	1.306
(S.E.)	(0.236)	(0.088)	(0.087)	(0.056)	(0.055)	(0.044)	(0.041)
Item 16 Bias	0.397	0.940	1.695	0.962	1.341	1.288	1.441
(S.E.)	(0.585)	(0.762)	(0.479)	(0.429)	(0.467)	(0.338)	(0.266)
Item 17 Bias	0.420	0.866	1.474	0.944	1.238	1.236	1.323
(S.E.)	(0.193)	(0.116)	(0.113)	(0.074)	(0.076)	(0.061)	(0.055)
Item 18 Bias	0.361	0.793	1.516	0.914	1.230	1.222	1.331
(S.E.)	(0.193)	(0.114)	(0.127)	(0.072)	(0.079)	(0.066)	(0.061)
Item 19 Bias	0.429	0.867	1.474	0.942	1.241	1.235	1.322
(S.E.)	(0.182)	(0.102)	(0.104)	(0.064)	(0.063)	(0.053)	(0.048)
Item 20 Bias	0.533	0.993	1.391	0.997	1.253	1.249	1.303
(S.E.)	(0.167)	(0.067)	(0.071)	(0.042)	(0.039)	(0.031)	(0.031)

Table 13

Item Bias for b2-parameters Across Extreme Negative Conditions

	n=100	n=250	n=500	n=750	n=1000	n=1500	n=3000
Item 1 Bias	1.024	2.428	3.008	2.386	3.321	3.104	3.239
(S.E.)	(2.893)	(1.536)	(0.403)	(0.358)	(0.582)	(0.401)	(0.301)
Item 2 Bias	1.512	2.105	2.424	2.023	2.607	2.475	2.545
(S.E.)	(0.345)	(0.457)	(0.232)	(0.228)	(0.366)	(0.247)	(0.172)
Item 3 Bias	1.319	2.194	2.723	2.225	3.021	2.842	2.975
(S.E.)	(0.941)	(0.330)	(0.227)	(0.138)	(0.224)	(0.163)	(0.106)
Item 4 Bias	1.640	2.656	3.135	2.480	3.471	3.263	3.428
(S.E.)	(0.701)	(1.050)	(0.579)	(0.545)	(0.779)	(0.566)	(0.428)
Item 5 Bias	1.672	2.315	2.882	2.321	3.234	3.034	3.174
(S.E.)	(0.890)	(0.379)	(0.318)	(0.212)	(0.328)	(0.239)	(0.160)
Item 6 Bias	5.061	3.512	9.064	2.605	3.993	3.549	3.737
(S.E.)	(5.960)	(3.052)	(5.498)	(0.357)	(1.204)	(0.469)	(0.309)
Item 7 Bias	0.324	1.579	2.558	2.103	2.745	2.610	2.681
(S.E.)	(4.218)	(4.759)	(0.297)	(0.284)	(0.427)	(0.304)	(0.234)
Item 8 Bias	1.480	2.265	2.731	2.244	3.055	2.884	3.003
(S.E.)	(0.450)	(0.355)	(0.248)	(0.221)	(0.346)	(0.245)	(0.170)
Item 9 Bias	1.365	2.406	3.070	2.396	3.434	3.193	3.348
(S.E.)	(2.433)	(0.693)	(1.667)	(0.295)	(0.483)	(0.335)	(0.239)
Item 10 Bias	2.826	4.991	4.777	4.680	6.374	5.428	4.529
(S.E.)	(19.556)	(3.618)	(6.820)	(2.597)	(3.026)	(2.742)	(2.216)
Item 11 Bias	1.498	2.087	2.466	2.057	2.631	2.526	2.612
(S.E.)	(0.248)	(0.128)	(0.108)	(0.095)	(0.138)	(0.106)	(0.061)
Item 12 Bias	1.455	2.183	2.733	2.207	3.003	2.851	2.976
(S.E.)	(0.398)	(0.447)	(0.265)	(0.269)	(0.425)	(0.292)	(0.172)
Item 13 Bias	-0.591	0.212	0.974	1.974	2.391	2.600	2.807
(S.E.)	(1.859)	(2.969)	(2.342)	(1.032)	(1.762)	(0.843)	(0.109)
Item 14 Bias	-0.510	0.446	2.674	2.217	2.999	2.813	2.942
(S.E.)	(10.250)	(15.212)	(0.536)	(0.285)	(0.484)	(0.333)	(0.230)
Item 15 Bias	1.588	2.143	2.559	2.118	2.780	2.648	2.754
(S.E.)	(0.794)	(0.272)	(0.160)	(0.136)	(0.219)	(0.152)	(0.102)
Item 16 Bias	2.107	3.173	3.624	2.875	4.191	3.806	3.865
(S.E.)	(1.861)	(1.810)	(1.082)	(1.163)	(1.443)	(1.045)	(0.780)
Item 17 Bias	1.493	2.250	2.768	2.253	3.063	2.899	3.022
(S.E.)	(0.422)	(0.314)	(0.234)	(0.198)	(0.299)	(0.221)	(0.151)
Item 18 Bias	1.543	2.309	2.981	2.324	3.250	3.027	3.185
(S.E.)	(0.764)	(0.334)	(0.922)	(0.208)	(0.339)	(0.229)	(0.161)
Item 19 Bias	1.481	2.236	2.734	2.235	3.038	2.858	2.995
(S.E.)	(0.359)	(0.250)	(0.198)	(0.161)	(0.255)	(0.174)	(0.123)
Item 20 Bias	1.488	2.069	2.444	2.050	2.597	2.498	2.587
(S.E.)	(0.256)	(0.127)	(0.103)	(0.090)	(0.133)	(0.103)	(0.055)

b3-parameters

Tables 14, 15 and 16 contain the item BIAS and standard error of the item BIAS for the *b3*-parameters across all sample sizes for the normal distribution, MN, and EN conditions. As with the test level results, unexpected results were obtained for n = 500 for all three distributions. The largest BIAS for all items with the normal distribution conditions was with n = 100. As with the *b1*- and *b2*-parameters, the MN and EN conditions were more similar to one another and for both distribution conditions, and BIAS increased as sample size increased.

The greatest amount of BIAS was observed for the EN distribution followed in turn by the MN distribution and the normal distribution conditions. And while generally the size of BIAS decreased as the sample size increased for the normal distributions, BIAS actually increased as sample size increased for the EN and MN distribution conditions. Further, the increase noted for the MN and EN distribution conditions was more extreme for the EN distribution than for the MN distribution but less extreme than for the *b1*- and *b2*-parameters. BIAS was less than or equal to 0.05 with nine exceptions for the normal distribution, $n \geq 750$, but for no items for both the EN and MN distribution conditions.

Further inspection of the full set of BIAS values reveals that the amount of BIAS was also dependent upon item: items which had *b3*- and *b4*-parameters very close in value had much larger BIAS, particularly when sample size was small. As with the *a*-, *b1*- and *b2*-parameters, the standard error of the BIAS generally decreased as the sample size increased across the 20 items for all three distribution conditions, and in this case, the standard errors tended to be close in value or larger than their corresponding bias only for the normal distribution conditions where $n \geq 100$. For the MN and EN distribution conditions most standard errors were smaller than the BIAS values.

Consequently, when the value of the BIAS was divided by it’s standard error for $n \geq 250$, the results suggest that the BIAS values were not significantly different from zero for the normal distribution conditions. In contrast the standard errors for the BIAS across EN and MN distribution conditions tended to be less that their corresponding BIAS, resulting in the ratio of the BIAS to its standard error being large, suggesting that the BIAS was significantly different from zero for these conditions.

Table 14

Item Bias for b3-parameters Across Normal Conditions

	n=100	n=250	n=500	n=750	n=1000	n=1500	n=3000
Item 1 Bias	0.208	0.060	0.074	0.028	-0.004	0.053	-0.001
(S.E.)	(1.124)	(0.178)	(0.098)	(0.107)	(0.092)	(0.072)	(0.053)
Item 2 Bias	0.156	0.079	-0.050	0.010	0.001	0.074	-0.029
(S.E.)	(0.396)	(0.426)	(0.199)	(0.249)	(0.201)	(0.158)	(0.122)
Item 3 Bias	0.135	0.069	0.062	0.027	-0.001	0.052	-0.003
(S.E.)	(0.140)	(0.072)	(0.064)	(0.042)	(0.036)	(0.029)	(0.022)
Item 4 Bias	0.127	0.075	0.047	0.035	-0.005	0.055	-0.008
(S.E.)	(0.241)	(0.220)	(0.116)	(0.131)	(0.112)	(0.085)	(0.068)
Item 5 Bias	0.156	0.094	0.018	0.033	0.010	0.066	-0.011
(S.E.)	(0.163)	(0.111)	(0.079)	(0.066)	(0.058)	(0.047)	(0.035)
Item 6 Bias	0.095	0.045	0.111	0.024	-0.016	0.037	0.004
(S.E.)	(0.147)	(0.087)	(0.071)	(0.054)	(0.046)	(0.036)	(0.026)
Item 7 Bias	1.073	0.136	0.007	0.033	0.006	0.063	-0.018
(S.E.)	(3.369)	(4.926)	(0.135)	(0.164)	(0.137)	(0.105)	(0.081)
Item 8 Bias	0.134	0.079	0.042	0.026	0.004	0.059	-0.008
(S.E.)	(0.165)	(0.121)	(0.077)	(0.077)	(0.066)	(0.048)	(0.037)
Item 9 Bias	0.102	0.052	0.112	0.029	-0.015	0.039	0.003
(S.E.)	(0.157)	(0.108)	(0.078)	(0.063)	(0.056)	(0.045)	(0.034)
Item 10 Bias	0.271	0.146	0.177	0.031	-0.030	0.022	0.012
(S.E.)	(0.430)	(0.382)	(0.116)	(0.090)	(0.071)	(0.055)	(0.043)
Item 11 Bias	-0.042	0.113	-0.036	0.035	0.021	0.093	-0.027
(S.E.)	(1.279)	(0.122)	(0.096)	(0.078)	(0.065)	(0.052)	(0.041)
Item 12 Bias	-5.590	-2.968	-0.388	-0.176	-0.007	0.079	-0.019
(S.E.)	(6.633)	(6.682)	(2.302)	(1.893)	(0.573)	(0.075)	(0.056)
Item 13 Bias	0.675	0.184	0.069	0.024	-0.003	0.048	0.000
(S.E.)	(1.240)	(0.796)	(0.062)	(0.039)	(0.035)	(0.029)	(0.021)
Item 14 Bias	-5.022	-1.510	0.039	0.023	-0.002	0.056	-0.009
(S.E.)	(11.168)	(9.068)	(0.097)	(0.108)	(0.093)	(0.076)	(0.053)
Item 15 Bias	0.149	0.085	0.029	0.027	0.005	0.063	-0.010
(S.E.)	(0.165)	(0.099)	(0.075)	(0.059)	(0.051)	(0.041)	(0.030)
Item 16 Bias	0.122	0.074	0.103	0.037	-0.014	0.052	0.005
(S.E.)	(0.348)	(0.365)	(0.159)	(0.198)	(0.167)	(0.131)	(0.094)
Item 17 Bias	0.126	0.090	0.017	0.025	0.011	0.071	-0.014
(S.E.)	(0.593)	(0.121)	(0.082)	(0.073)	(0.062)	(0.051)	(0.036)
Item 18 Bias	0.145	0.084	0.049	0.025	0.001	0.057	-0.007
(S.E.)	(0.149)	(0.086)	(0.068)	(0.052)	(0.046)	(0.035)	(0.026)
Item 19 Bias	0.150	0.091	0.024	0.029	0.008	0.067	-0.010
(S.E.)	(0.152)	(0.094)	(0.076)	(0.055)	(0.048)	(0.040)	(0.029)
Item 20 Bias	0.158	0.100	-0.002	0.030	0.015	0.080	-0.017
(S.E.)	(0.160)	(0.093)	(0.077)	(0.058)	(0.051)	(0.041)	(0.030)

Table 15

Item Bias for b3-parameters Across Moderate Negative Conditions

	n=100	n=250	n=500	n=750	n=1000	n=1500	n=3000
Item 1 Bias	0.629	0.921	1.492	0.971	1.267	1.253	1.346
(S.E.)	(1.414)	(0.573)	(0.156)	(0.132)	(0.137)	(0.113)	(0.091)
Item 2 Bias	0.571	1.034	1.279	0.987	1.222	1.219	1.242
(S.E.)	(0.289)	(0.227)	(0.123)	(0.148)	(0.119)	(0.099)	(0.075)
Item 3 Bias	0.479	0.911	1.448	0.959	1.245	1.237	1.312
(S.E.)	(0.268)	(0.076)	(0.083)	(0.047)	(0.046)	(0.038)	(0.038)
Item 4 Bias	0.480	0.937	1.440	0.963	1.256	1.243	1.316
(S.E.)	(0.238)	(0.208)	(0.152)	(0.126)	(0.133)	(0.105)	(0.085)
Item 5 Bias	0.538	0.987	1.394	0.991	1.251	1.247	1.297
(S.E.)	(0.176)	(0.084)	(0.077)	(0.055)	(0.049)	(0.041)	(0.037)
Item 6 Bias	0.369	0.807	1.516	0.915	1.241	1.231	1.341
(S.E.)	(0.197)	(0.120)	(0.131)	(0.077)	(0.081)	(0.067)	(0.059)
Item 7 Bias	1.380	1.149	1.391	0.982	1.247	1.243	1.289
(S.E.)	(3.286)	(4.939)	(0.123)	(0.117)	(0.112)	(0.089)	(0.073)
Item 8 Bias	0.493	0.945	1.423	0.972	1.247	1.240	1.305
(S.E.)	(0.181)	(0.111)	(0.095)	(0.067)	(0.069)	(0.054)	(0.048)
Item 9 Bias	0.396	0.814	1.520	0.915	1.236	1.225	1.330
(S.E.)	(0.426)	(0.150)	(0.146)	(0.094)	(0.100)	(0.080)	(0.069)
Item 10 Bias	0.641	1.009	1.842	0.912	1.265	1.214	1.337
(S.E.)	(1.366)	(0.501)	(0.897)	(0.237)	(0.252)	(0.157)	(0.154)
Item 11 Bias	0.602	1.097	1.313	1.029	1.254	1.254	1.269
(S.E.)	(0.269)	(0.073)	(0.056)	(0.046)	(0.037)	(0.033)	(0.027)
Item 12 Bias	-4.804	-2.534	1.307	0.954	1.244	1.241	1.273
(S.E.)	(7.383)	(9.179)	(0.783)	(1.237)	(0.058)	(0.048)	(0.040)
Item 13 Bias	1.100	1.181	1.498	0.940	1.225	1.221	1.301
(S.E.)	(1.217)	(1.149)	(0.386)	(0.053)	(0.049)	(0.041)	(0.040)
Item 14 Bias	-4.808	-3.084	1.426	0.968	1.247	1.244	1.306
(S.E.)	(11.713)	(15.524)	(0.121)	(0.098)	(0.098)	(0.080)	(0.064)
Item 15 Bias	0.518	0.971	1.406	0.984	1.248	1.244	1.300
(S.E.)	(0.170)	(0.079)	(0.075)	(0.051)	(0.048)	(0.039)	(0.036)
Item 16 Bias	0.434	0.965	1.567	0.969	1.298	1.264	1.381
(S.E.)	(0.414)	(0.521)	(0.337)	(0.299)	(0.324)	(0.236)	(0.188)
Item 17 Bias	0.533	0.998	1.385	0.994	1.249	1.246	1.293
(S.E.)	(0.172)	(0.089)	(0.077)	(0.056)	(0.050)	(0.042)	(0.037)
Item 18 Bias	0.486	0.930	1.427	0.971	1.244	1.240	1.307
(S.E.)	(0.169)	(0.083)	(0.083)	(0.053)	(0.051)	(0.042)	(0.040)
Item 19 Bias	0.528	0.980	1.397	0.987	1.251	1.244	1.302
(S.E.)	(0.168)	(0.079)	(0.076)	(0.048)	(0.046)	(0.036)	(0.034)
Item 20 Bias	0.569	1.038	1.364	1.016	1.258	1.255	1.294
(S.E.)	(0.166)	(0.066)	(0.062)	(0.040)	(0.035)	(0.029)	(0.028)

Table 16

Item Bias for b3-parameters Across Extreme Negative Conditions

	n=100	**n=250**	**n=500**	**n=750**	**n=1000**	**n=1500**	**n=3000**
Item 1 Bias	2.225	2.501	2.914	2.324	3.197	2.999	3.125
(S.E.)	(2.686)	(1.473)	(0.368)	(0.327)	(0.535)	(0.364)	(0.276)
Item 2 Bias	1.445	1.890	2.138	1.836	2.251	2.180	2.221
(S.E.)	(0.290)	(0.334)	(0.161)	(0.174)	(0.231)	(0.165)	(0.111)
Item 3 Bias	1.591	2.165	2.637	2.172	2.904	2.748	2.868
(S.E.)	(0.805)	(0.267)	(0.192)	(0.124)	(0.192)	(0.144)	(0.090)
Item 4 Bias	1.551	2.315	2.697	2.195	2.927	2.787	2.901
(S.E.)	(0.439)	(0.648)	(0.367)	(0.340)	(0.505)	(0.369)	(0.277)
Item 5 Bias	1.493	2.065	2.447	2.052	2.623	2.520	2.600
(S.E.)	(0.263)	(0.172)	(0.120)	(0.115)	(0.172)	(0.129)	(0.080)
Item 6 Bias	1.569	2.363	2.948	2.357	3.317	3.096	3.263
(S.E.)	(0.829)	(0.356)	(0.394)	(0.207)	(0.344)	(0.244)	(0.178)
Item 7 Bias	2.720	2.788	2.502	2.065	2.673	2.547	2.612
(S.E.)	(4.174)	(4.815)	(0.274)	(0.261)	(0.390)	(0.280)	(0.215)
Item 8 Bias	1.476	2.139	2.549	2.120	2.786	2.656	2.748
(S.E.)	(0.305)	(0.259)	(0.187)	(0.166)	(0.257)	(0.187)	(0.125)
Item 9 Bias	1.920	2.356	2.989	2.320	3.260	3.049	3.186
(S.E.)	(1.486)	(0.625)	(0.892)	(0.250)	(0.416)	(0.286)	(0.201)
Item 10 Bias	6.763	4.551	8.421	2.870	4.128	3.636	3.598
(S.E.)	(6.187)	(4.630)	(6.166)	(1.034)	(1.295)	(0.675)	(0.476)
Item 11 Bias	1.504	1.910	2.155	1.861	2.218	2.171	2.208
(S.E.)	(0.190)	(0.086)	(0.071)	(0.069)	(0.087)	(0.068)	(0.032)
Item 12 Bias	-1.444	-0.429	2.183	1.821	2.352	2.288	2.353
(S.E.)	(5.079)	(7.977)	(1.155)	(1.583)	(0.263)	(0.187)	(0.074)
Item 13 Bias	2.778	3.589	3.629	2.250	3.106	2.735	2.788
(S.E.)	(1.507)	(2.420)	(1.477)	(0.821)	(1.304)	(0.663)	(0.104)
Item 14 Bias	-5.384	-7.390	2.434	2.151	2.859	2.698	2.810
(S.E.)	(15.095)	(43.014)	(2.162)	(0.249)	(0.419)	(0.293)	(0.200)
Item 15 Bias	1.497	2.077	2.463	2.056	2.641	2.533	2.621
(S.E.)	(0.481)	(0.178)	(0.127)	(0.116)	(0.179)	(0.127)	(0.082)
Item 16 Bias	1.866	2.789	3.185	2.553	3.599	3.315	3.369
(S.E.)	(1.246)	(1.333)	(0.810)	(0.848)	(1.072)	(0.772)	(0.576)
Item 17 Bias	1.500	2.051	2.409	2.031	2.567	2.478	2.554
(S.E.)	(0.262)	(0.178)	(0.123)	(0.120)	(0.170)	(0.126)	(0.080)
Item 18 Bias	1.462	2.125	2.558	2.125	2.778	2.643	2.749
(S.E.)	(0.292)	(0.184)	(0.148)	(0.126)	(0.198)	(0.140)	(0.090)
Item 19 Bias	1.489	2.082	2.464	2.063	2.651	2.533	2.624
(S.E.)	(0.265)	(0.157)	(0.119)	(0.108)	(0.161)	(0.114)	(0.076)
Item 20 Bias	1.507	2.006	2.336	1.980	2.449	2.370	2.439
(S.E.)	(0.227)	(0.099)	(0.087)	(0.075)	(0.109)	(0.086)	(0.042)

b4-parameters

Tables 17, 18 and 19 contain the item BIAS and standard error of the item BIAS for the *b4*-parameters across all sample sizes for the normal distribution, MN, and EN conditions. As with the test level results, unexpected results were obtained for n = 500 for all three distributions. The largest BIAS for all items with the normal distribution conditions was with n = 100. As with the *b1*-, *b2*- and *b3*-parameters, the MN and EN conditions were more similar to one another and for both distribution conditions, and BIAS increased as sample size increased.

The greatest amount of BIAS was observed for the EN distribution followed in turn by the MN distribution and the normal distribution conditions. And while generally the size of BIAS decreased as the sample size increased for the normal distributions, BIAS actually increased as sample size increased for the EN and MN distribution conditions. Further, the increase noted for the MN and EN distribution conditions was more extreme for the EN distribution than for the MN distribution but less extreme than for the *b1*-, *b2*- and *b3*-parameters. BIAS was less than or equal to 0.05 with four exceptions for the normal distribution, $n \geq 750$, but for no items for both the EN and MN distribution conditions.

Further inspection of the full set of BIAS values reveals that the amount of BIAS was also dependent upon item: items which had *b3*- and *b4*-parameters very close in value had much larger BIAS, particularly when sample size was small. As with the *a*-, *b1*- and *b2*-parameters, the standard error of the BIAS generally decreased as the sample size increased across the 20 items for all three distribution conditions, and in this case, the standard errors tended to be close in value or larger than their corresponding bias only for the normal distribution conditions where $n \geq 100$. For the MN and EN distribution conditions most standard errors were smaller than the BIAS values.

Consequently, when the value of the BIAS was divided by it's standard error for $n \geq 250$, the results suggest that the BIAS values were not significantly different from zero for the normal distribution conditions. In contrast the standard errors for the BIAS across EN and MN distribution conditions tended to be less that their corresponding BIAS, resulting in the ratio of the BIAS to its standard error being large, suggesting that the BIAS was significantly different from zero for these conditions.

Table 17

Item Bias for b4-parameters Across Normal Conditions

	n=100	n=250	n=500	n=750	n=1000	n=1500	n=3000
Item 1 Bias	0.185	0.092	-0.077	0.018	0.019	0.086	-0.027
(S.E.)	(0.405)	(0.390)	(0.207)	(0.233)	(0.192)	(0.152)	(0.118)
Item 2 Bias	0.154	0.080	-0.114	0.000	0.004	0.086	-0.043
(S.E.)	(0.557)	(0.577)	(0.271)	(0.329)	(0.264)	(0.210)	(0.163)
Item 3 Bias	0.175	0.122	-0.043	0.039	0.024	0.094	-0.026
(S.E.)	(0.217)	(0.146)	(0.110)	(0.085)	(0.075)	(0.058)	(0.047)
Item 4 Bias	0.134	0.080	0.010	0.034	0.000	0.065	-0.015
(S.E.)	(0.287)	(0.273)	(0.140)	(0.161)	(0.141)	(0.108)	(0.085)
Item 5 Bias	0.185	0.112	-0.034	0.037	0.022	0.085	-0.020
(S.E.)	(0.211)	(0.161)	(0.107)	(0.096)	(0.085)	(0.066)	(0.050)
Item 6 Bias	0.172	0.116	-0.033	0.033	0.019	0.087	-0.023
(S.E.)	(0.202)	(0.149)	(0.101)	(0.087)	(0.075)	(0.060)	(0.045)
Item 7 Bias	0.314	0.109	-0.036	0.025	0.010	0.071	-0.027
(S.E.)	(0.708)	(0.968)	(0.175)	(0.209)	(0.177)	(0.133)	(0.107)
Item 8 Bias	0.155	0.092	0.010	0.026	0.012	0.068	-0.014
(S.E.)	(0.196)	(0.154)	(0.092)	(0.095)	(0.080)	(0.059)	(0.045)
Item 9 Bias	0.144	0.097	0.019	0.026	0.006	0.066	-0.010
(S.E.)	(0.181)	(0.126)	(0.087)	(0.076)	(0.067)	(0.053)	(0.040)
Item 10 Bias	0.146	0.089	0.046	0.026	0.004	0.057	-0.007
(S.E.)	(0.162)	(0.108)	(0.076)	(0.064)	(0.054)	(0.045)	(0.032)
Item 11 Bias	0.387	0.116	-0.052	0.038	0.025	0.102	-0.032
(S.E.)	(1.207)	(0.150)	(0.112)	(0.092)	(0.078)	(0.062)	(0.049)
Item 12 Bias	-2.097	-1.182	-0.319	-0.096	-0.002	0.080	-0.019
(S.E.)	(7.441)	(6.731)	(1.986)	(1.741)	(0.483)	(0.076)	(0.057)
Item 13 Bias	0.172	0.098	0.046	0.026	0.004	0.057	-0.004
(S.E.)	(0.166)	(0.119)	(0.066)	(0.045)	(0.038)	(0.031)	(0.022)
Item 14 Bias	-4.590	-1.284	0.033	0.022	-0.001	0.057	-0.009
(S.E.)	(10.976)	(7.574)	(0.101)	(0.110)	(0.095)	(0.078)	(0.055)
Item 15 Bias	0.157	0.094	0.009	0.029	0.008	0.071	-0.012
(S.E.)	(0.176)	(0.116)	(0.083)	(0.068)	(0.059)	(0.048)	(0.034)
Item 16 Bias	0.105	0.004	0.001	0.017	-0.012	0.062	-0.016
(S.E.)	(0.431)	(0.480)	(0.200)	(0.254)	(0.215)	(0.163)	0.127
Item 17 Bias	0.143	0.096	-0.006	0.026	0.015	0.079	-0.018
(S.E.)	(0.509)	(0.145)	(0.093)	(0.087)	(0.075)	(0.059)	0.043
Item 18 Bias	0.169	0.120	-0.019	0.029	0.019	0.082	-0.021
(S.E.)	(0.186)	(0.125)	(0.093)	(0.081)	(0.067)	(0.052)	0.040
Item 19 Bias	0.171	0.112	-0.032	0.031	0.022	0.086	-0.023
(S.E.)	(0.205)	(0.143)	(0.104)	(0.085)	(0.073)	(0.058)	0.045
Item 20 Bias	0.176	0.107	-0.039	0.031	0.025	0.094	-0.027
(S.E.)	(0.201)	(0.126)	(0.100)	(0.080)	(0.067)	(0.054)	0.042

Table 18

Item Bias for b4-parameters Across Moderate Negative Conditions

	n=100	n=250	n=500	n=750	n=1000	n=1500	n=3000
Item 1 Bias	0.607	1.089	1.230	1.011	0.019	1.217	1.215
(S.E.)	(0.306)	(0.239)	(0.118)	(0.141)	(0.192)	(0.097)	(0.072)
Item 2 Bias	0.627	1.101	1.166	0.992	0.004	1.200	1.193
(S.E.)	(0.372)	(0.315)	(0.165)	(0.205)	(0.264)	(0.135)	(0.101)
Item 3 Bias	0.629	1.128	1.285	1.035	0.024	1.251	1.259
(S.E.)	(0.182)	(0.081)	(0.058)	(0.048)	(0.075)	(0.033)	(0.027)
Item 4 Bias	0.519	0.985	1.377	0.981	0.000	1.235	1.286
(S.E.)	(0.239)	(0.197)	(0.120)	(0.114)	(0.141)	(0.085)	(0.070)
Item 5 Bias	0.603	1.078	1.321	1.022	0.022	1.246	1.267
(S.E.)	(0.186)	(0.093)	(0.065)	(0.059)	(0.085)	(0.042)	(0.034)
Item 6 Bias	0.603	1.088	1.319	1.024	0.019	1.251	1.268
(S.E.)	(0.182)	(0.092)	(0.061)	(0.056)	(0.075)	(0.038)	(0.031)
Item 7 Bias	0.721	1.018	1.315	0.996	0.010	1.227	1.255
(S.E.)	(0.635)	(1.021)	(0.111)	(0.133)	(0.177)	(0.087)	(0.065)
Item 8 Bias	0.544	1.005	1.378	0.994	0.012	1.243	1.288
(S.E.)	(0.179)	(0.107)	(0.079)	(0.064)	(0.080)	(0.044)	(0.041)
Item 9 Bias	0.526	0.984	1.394	0.990	0.006	1.243	1.295
(S.E.)	(0.177)	(0.100)	(0.085)	(0.060)	(0.067)	(0.045)	(0.039)
Item 10 Bias	0.497	0.973	1.430	0.977	0.004	1.238	1.302
(S.E.)	(0.180)	(0.115)	(0.095)	(0.065)	(0.054)	(0.051)	(0.058)
Item 11 Bias	0.633	1.126	1.275	1.034	0.025	1.247	1.251
(S.E.)	(0.277)	(0.081)	(0.058)	(0.052)	(0.078)	(0.036)	(0.028)
Item 12 Bias	-3.194	-2.441	1.321	0.956	-0.002	1.240	1.271
(S.E.)	(7.142)	(8.944)	(0.777)	(1.203)	(0.483)	(0.049)	(0.040)
Item 13 Bias	0.533	0.974	1.423	0.972	0.004	1.236	1.298
(S.E.)	(0.192)	(0.149)	(0.084)	(0.048)	(0.038)	(0.034)	(0.033)
Item 14 Bias	-4.653	-2.868	1.419	0.970	-0.001	1.244	1.303
(S.E.)	(11.630)	(14.495)	(0.117)	(0.097)	(0.095)	(0.078)	(0.063)
Item 15 Bias	0.544	1.006	1.382	1.000	0.008	1.248	1.295
(S.E.)	(0.172)	(0.078)	(0.070)	(0.051)	(0.059)	(0.036)	(0.033)
Item 16 Bias	0.507	0.988	1.390	0.976	-0.012	1.233	1.299
(S.E.)	(0.333)	(0.315)	(0.176)	(0.181)	(0.215)	(0.141)	(0.105)
Item 17 Bias	0.571	1.044	1.352	1.011	0.015	1.248	1.280
(S.E.)	(0.173)	(0.093)	(0.071)	(0.057)	(0.075)	(0.040)	(0.036)
Item 18 Bias	0.589	1.060	1.347	1.019	0.019	1.254	1.282
(S.E.)	(0.177)	(0.079)	(0.063)	(0.051)	(0.067)	(0.037)	(0.031)
Item 19 Bias	0.607	1.088	1.320	1.026	0.022	1.251	1.270
(S.E.)	(0.177)	(0.083)	(0.061)	(0.054)	(0.073)	(0.037)	(0.029)
Item 20 Bias	0.607	1.108	1.312	1.033	0.025	1.258	1.274
(S.E.)	(0.169)	(0.074)	(0.059)	(0.045)	(0.067)	(0.032)	(0.026)

Table 19

Item Bias for b4-parameters Across Extreme Negative Conditions

	n=100	n=250	n=500	n=750	n=1000	n=1500	n=3000
Item 1 Bias	1.426	1.763	1.947	1.717	2.002	1.968	1.994
(S.E.)	(0.290)	(0.290)	(0.127)	(0.151)	(0.162)	(0.118)	(0.083)
Item 2 Bias	1.342	1.599	1.768	1.586	1.790	1.794	1.793
(S.E.)	(0.350)	(0.429)	(0.174)	(0.218)	(0.218)	(0.163)	(0.117)
Item 3 Bias	1.496	1.850	2.059	1.802	2.109	2.075	2.099
(S.E.)	(0.176)	(0.085)	(0.064)	(0.065)	(0.081)	(0.059)	(0.029)
Item 4 Bias	1.514	2.129	2.456	2.039	2.620	2.517	2.604
(S.E.)	(0.336)	(0.457)	(0.258)	(0.243)	(0.364)	(0.263)	(0.195)
Item 5 Bias	1.495	1.910	2.174	1.871	2.249	2.197	2.236
(S.E.)	(0.206)	(0.115)	(0.080)	(0.084)	(0.106)	(0.082)	(0.048)
Item 6 Bias	0.199	1.914	2.174	1.871	2.244	2.191	2.235
(S.E.)	(0.199)	(0.108)	(0.081)	(0.073)	(0.099)	(0.077)	(0.041)
Item 7 Bias	1.667	2.012	2.223	1.892	2.331	2.252	2.297
(S.E.)	(0.797)	(1.094)	(0.185)	(0.183)	(0.253)	(0.184)	(0.138)
Item 8 Bias	1.482	2.039	2.375	2.008	2.545	2.450	2.517
(S.E.)	(0.252)	(0.197)	(0.135)	(0.126)	(0.191)	(0.138)	(0.089)
Item 9 Bias	1.494	2.068	2.441	2.038	2.622	2.507	2.585
(S.E.)	(0.277)	(0.206)	(0.152)	(0.126)	(0.209)	(0.146)	(0.095)
Item 10 Bias	1.478	2.198	2.542	2.127	2.778	2.643	2.720
(S.E.)	(0.319)	(0.272)	(0.185)	(0.179)	(0.262)	(0.176)	(0.129)
Item 11 Bias	1.493	1.837	2.033	1.787	2.068	2.043	2.063
(S.E.)	(0.175)	(0.085)	(0.064)	(0.065)	(0.076)	(0.062)	(0.028)
Item 12 Bias	1.722	-0.532	2.163	1.808	2.332	2.270	2.334
(S.E.)	(5.527)	(8.221)	(1.238)	(1.635)	(0.252)	(0.180)	(0.071)
Item 13 Bias	1.545	2.287	2.572	2.083	2.691	2.552	2.642
(S.E.)	(0.362)	(0.477)	(0.197)	(0.163)	(0.282)	(0.193)	(0.076)
Item 14 Bias	-5.921	-2.784	2.374	2.131	2.823	2.667	2.774
(S.E.)	(16.244)	(17.350)	(2.539)	(0.239)	(0.401)	(0.282)	(0.191)
Item 15 Bias	1.510	2.030	2.378	2.002	2.520	2.429	2.504
(S.E.)	(0.397)	(0.145)	(0.107)	(0.100)	(0.149)	(0.105)	(0.066)
Item 16 Bias	1.610	2.256	2.564	2.123	2.789	2.640	2.678
(S.E.)	(0.603)	(0.730)	(0.425)	(0.443)	(0.568)	(0.420)	(0.309)
Item 17 Bias	1.504	1.974	2.280	1.942	2.388	2.320	2.376
(S.E.)	(0.224)	(0.137)	(0.097)	(0.099)	(0.132)	(0.097)	(0.059)
Item 18 Bias	1.502	1.958	2.251	1.925	2.347	2.280	2.332
(S.E.)	(0.209)	(0.110)	(0.083)	(0.081)	(0.111)	(0.081)	(0.044)
Item 19 Bias	1.505	1.906	2.165	1.870	2.236	2.186	2.225
(S.E.)	(0.197)	(0.100)	(0.076)	(0.074)	(0.096)	(0.072)	(0.040)
Item 20 Bias	1.507	1.903	2.140	1.858	2.202	2.157	2.192
(S.E.)	(0.185)	(0.081)	(0.067)	(0.065)	(0.082)	(0.064)	(0.029)

Summary

Item-level analyses provide some insight into the test-level results. While reporting BIAS at the test level and RMSEs can tell us something about the ability to recover total test scores using the GRM and MULTILOG, these test level results do not provide information about which items may or may not be problematic. In contrast, the item level results provide in-depth information. Items with larger 'true' *a*-parameter values (above 1.20) tended to be overestimated to a larger degree than that those items with smaller 'true' *a*-parameters.

Three items were particularly problematic for *b*-parameter estimation; items 10, 12, and 14. For item 10, the *b1*- and *b2*-parameters were within 0.06 of one another and particularly when there were smaller sample sizes (n = 100, 250) MULTILOG produced estimates that were much larger than the 'true' values. Similarly, items 12 and 14 within 0.04 and 0.07 respectively, of one another and both the *b3*- and *b4*-parameters for these items were greatly overestimated. 'True' *b*-parameters that were lower in value on the θ scale were also poorly recovered in the skewed distributions

Chapter 6: Summary, Limitations and Future Directions

This chapter is organized in five sections. First, a summary of the research questions and methods used in this study is provided. Test-level and item-level results are then summarized and discussed in the second section and limitations of the current study are identified in the third section. Conclusions and implications for practice are discussed in the fourth section and directions for future research are provided in the last section.

Research Design and Methods Summary

Currently, one of the most popular methods used in calibrating polytomous data in education is the use of the GRM as executed in MULTILOG. However, with the expanding scope of use of polytomous item response theory (PIRT) in the social and health sciences not much time has been spent on discussing the possible effect of the non-normal distributions and small sample sizes common in these areas. Thus, the purpose of the current study was to conduct a simulation study to inform applied research regarding the use of PIRT with non-normal data, particularly when sample sizes are small, which is so often the case in clinical studies.

Four research questions were addressed:

1) Does the shape of the underlying θ distribution have an effect on test-level statistical outcomes for item and person parameter recovery under the GRM using MULTILOG?,

2) Does the shape of the underlying θ distribution have an effect on item-level statistical outcomes for item and person parameter recovery under the GRM using MULTILOG?,
3) Does sample size have an effect on test-level statistical outcomes for item and person parameter recovery under the GRM using MULTILOG?, and
4) Does sample size have an effect on item-level statistical outcomes for item and person parameter recovery under the GRM using MULTILOG?

Previous simulation studies suggested a minimum sample size of 500 for accurate parameter estimation under the GRM (Reise & Yu, 1990). However, the recommended samples sizes were not met for many of the studies in which PIRT was used in the social and heath science areas. Consequently, a range of seven sample sizes (100, 250, 500, 750, 1,000, 1,500, and 3,000) crossed three distribution shapes (normal (to act as a baseline), moderate negatively skewed (MN), and extremely negatively skewed (EN)) were considered. The number of replications of each of the 21 conditions was 1,000. RMSEs and test-level BIAS were calculated across items to assess the effect for sample size and distribution shape on total test scores and item-level BIAS and standard error of item BIAS were calculated to assess the effect of sample size and distribution shape at the item level.

Results Summary

Test level. Aside from θ estimates, the EN distribution conditions produced the poorest results overall. At the test level, recovery of the *a*-parameters showed the most consistent improvement as the sample size increased across all distribution conditions, and *b3*- and *b4*-location parameters were more accurately recovered than *b1*- and *b2*- location parameters. Test-level BIAS results revealed that for the EN and MN distribution conditions, locations parameters were in general underestimated.

The test-level results indicated that the shape of the underlying θ distribution does in fact have an effect on the accuracy of parameter estimation. The results also indicated that the θ distribution factor interacted with sample size and the value of the ‘true’ parameter. In general, and as expected, the normal distribution conditions produced better test- and item-level results than either skewed distribution across the seven sample sizes.

In addition, as with other simulation results (de Ayala, 2009; Reise & Yu, 1990), test-level results for the study showed that generally, as sample size increased, the accuracy of the recovered parameters increased at n = 750, after which accuracy tended to be constant for the normal distribution and at n = 1,000 after which the accuracy tended to be constant for the MN and EN distributions. Further, the accuracy of the estimated parameters is acceptable for the normal distribution condition when $n \geq 750$ but no acceptable sample size was found in this study for the MN or the EN conditions.

de Ayala (2009) and Reise and Yu (1990) suggested a minimum sample size of 500 for accurate parameter estimation using the GRM. However, unexplained results were obtained in the present study and attempts to correct the situation were futile (see *Limitations*). But the results of the present study suggest a minimum sample size of 750 with normally distributed data. Given extreme distributions of ability are found in personality and health research (see Bolt et al., 2004 and Cooke et al., 2001 for examples) additional research is needed to determine how to best handle situations in which is the distribution of the latent trait is extremely skewed.

Item level. As might be expected, the item level results agreed with the test level results but provided the reason and clarity for the test level results. For example, the *a*-parameter estimates for the extreme skewed distribution for each item were uniformly large, thus accounting for the large *a*-parameter estimates at the test level. The *a*-parameter estimates for the items with larger 'true' *a*-parameter values (above 1.20) tended to be overestimated to a larger degree than items with smaller 'true' *a*-parameters.

Three items were particularly problematic for *b*-parameter estimation: items 10, 12, and 14. These items each had two adjacent *b*-parameters which 'true' values that were very close in value. This caused problems with calibration of the data. In addition, across all 20 items, with *b*-parameters with 'true' values at locations along the θ scale where there was very little response data had much larger item BIAS. By looking at the test-level results *b*-parameters appear to be poorly recovered yet when item-level BIAS is investigated it can be seen that this is a result of three problematic items causing error in estimation.

The results of item-level analyses have not been provided in the literature to this point, possibly because of the number of pages needed to present the results. But the results of the present study reveal the item level data do shed some light on the test-level results. While reporting BIAS at the test level can tell us how accurate the score is overall and RMSE can tell us something about the precision of the total test score, authors should provide summary information about the characteristics of the items used to obtain the total score and if it was necessary to remove items that were found to be problematic.

Limitations of Present Study

Many of the replications for the conditions with small sample size and with the EN distribution condition did not converge even though default option for the number of calibration cycles was increased to 500 (see Table 4, Chapter 3). For example, of the 1,000 replications for the EN distribution condition and n = 100, 51.4% did not meet the convergence criterion set for this study (0.001). Increasing the criterion to 0.01 decreased this percentage substantially with at most 34.0% of the replications not converging with the EN distribution condition and n = 250. Due to this problem with convergence, as mentioned in Chapter 3, the outcome measures provided me be too large, too small or correct.

As mentioned above, unexplained values were obtained when estimating the *a*-parameter when n = 500 for all three distributions and at some, but not all, of the *b*-locations for the MN and EN distributions. As well, the ability estimates for the MN distribution with n = 3,000 were not as expected. To investigate these situations more thoroughly, three more datasets with 1,000 replications with n = 500 were generated and the analyses repeated for each. Results of these analyses were not consistent and inconclusive. There is no readily apparent explanation for why this happened and therefore, the results with n = 500 were essentially disregarded.

Conclusions and Implications for Practice

The results suggest that when completing a PIRT analysis with small samples or non-normal data, it is necessary to interpret the results of an item calibration with caution, particularly when the distribution is markedly negatively skewed. It is necessary for researchers using PIRT item calibration to have a complete statistical description of their data before deciding on whether or not to proceed with the analysis.

When using the GRM with MULTILOG to calibrate the items, the sample size should be at least 750 if scores on the latent trait of interest are normally distributed. Results derived from samples that are moderately or extremely negatively skewed may be unsatisfactory. It is essential that researchers are thorough in their initial assessment of the data to be calibrated.

Lastly, as mentioned above, while reporting BIAS at the test level can tell us how accurate the score is overall and RMSE can tell us something about the precision of the total test score, authors should provide summary information about the characteristics of the items used to obtain the total score and how items found to be problematic were handled.

Future Directions

As indicated in the identification of limitations, it is recommended that another program (such as SAS and R) be used to generate data with conditions similar to those used in this study to possibly aid in explanation of the unexpected results obtained in the present study. Additionally, the current study considered a 20 item, 5 point likert-type scale assessment. Given that test length can affect calibration, the number of items should be varied. Also, the greater the number of score categories for an item, the greater the number of item parameters estimated, which in turn requires larger sample sizes. Given 3- and 7-point likert type items are often used in social science or health science assessment studies, the influence of sample size as well as distribution shape should be assessed with the intent of determining the minimum sample size required and the maximum skewness allowed.

References

Bahry, L. M., & Gotzmann, A. (2011, April). *The effect of levels of skewness on item parameter recovery.* Poster session presented at the Annual Conference of the National Council onMeasurement in Education, New Orleans, LA.

Bock, R. D. (1976). Basic issues in the measurement of change. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement.* London: Wiley & Sons, 75-76.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46(4),* 443-459. doi:10.1007/BF02293801

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6(4),* 431-444. doi:10.1177/014662168200600405

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the Psychopathy Checklist – Revised.

Psychological Assessment, *16(2),* 155-168. doi: 10.1037/1040-3590.16.2.155

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36(4),* 523-562. doi:10.1207/S15327906MBR3604_03

Cook, K. F., Teal, C. R., Bjorner, J. B., Cella, D., Chang, C., Crane, P. K., Gibbons, L. E., Hays, R. D., McHorney, C. A., Ocepek-Welikson, K., Raczek, A. E., Teresi, J. A., & Reeve, B. B. (2007). IRT health outcomes data analysis project: An overview and summary. *Quality of Life Research, 16,* 121-132. doi: 10.1007/s11136-007-9177-5

Cooke, D. J., Michie, C., & Kosson, D. S. (2001). Psychopathy and ethnicity: Structural, item, and test generalizability of the Psychopathy Checklist – Revised (PCL – R) in Caucasian and African American participants. *Psychological Assessment, 13(4),* 531-543. doi:10.1037/10403590.13.4.531

de Ayala, R. J. (2009). *The theory and practice of item response theory.* NY: Guilford Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39(1),* 1-38. Retrieved from http://www.jstor.org.login.ezproxy.library.ualberta.ca/action/show Publication?journalCode=jroyastatsocise2

Dodd, B. G. (1984). *Attitude scaling: A comparison of the graded response and partial credit latent trait models* (Doctoral dissertation). Available from Proquest Dissertation and Theses database. (AAT No. 8421690)

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13,* 77-90. doi:10.1177/014662168901300108

Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *The Psychopathy Checklist: Youth Version.* Toronto, Canada: Multi-Health Systems.

Gumpel, T. P. (1999). Use of item response theory to develop a measure of first-grade readiness. *Psychology in the Schools*, *36(4),* 285-293. doi: 10.1002/(SICI)1520-6807(199907)36:4<285::AID-PITS2>3.0.CO;2-M

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

Han, K. T. (2010). WinGen (Version 3) [Computer Software]. Retrieved from http://www.hantest.net/wingen

Hare, R. D. (2003). *Hare Psychopathy Checklist- Revised* (2nd ed.). North Tonawanda, NY: Multi-Health Systems.

Harwell, M., Stone, C.A., Hsu, T., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement*, *20,* 101-125. doi:10.1177/014662169602000201

Hays, R. D., Liu, H., Spritzer, K., & Cella, D. (2007). Item response theory analysis of physical functioning items in the medical outcomes study. *Medical Care, 45(5),* S32-S38. doi:10.1097/01.mlr.0000246649.43232.82

Kang, T., Cohen, A. S., & Sung, H. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33,* 499-518. doi:10.1177/0146621608327800

Linacre, J. M. (2002). WINSTEPS: Rasch-model computer program. (Version 3.36) [Computer software]. Chicago: MESA Press.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23(2),* 157-162. doi:10.1111/j.17453984.1986.tb00241.x

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174. doi: 10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of the EM algorithm. *Applied Psychological Measurement, 16,* 159-176. doi:10.1177/014662169201600206

Muraki, E., & Bock, R. D. (1997). PARSCALE-3: IRT based test scoring and item analysis for graded items and rating scales [Computer software]. Chicago: Scientific Software International.

Reise, S. P. & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81(2),* 93-103. doi:10.1207/S15327752JPA8102_01

Reise, S. P. & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27(2),* 133-144. doi:10.1111/j.17453984.1990.tb00738.x

Roche, A. F., Wainer, H., & Thissen, D. (1975). *Skeletal maturity: The knee joint as biological indicator.* New York: Plenum.

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education, 21*(1), 65 88. doi: 10.1080/08957340701796415

SAS Institute. (2006). SAS, Version 9.2. [Computer software].

Schrum, C. L., & Salekin, R. T. (2006). Psychopathy in adolescent female offenders: An item response theory analysis of the Psychopathy Checklist: Youth Version. *Behavioral Sciences and the Law, 24,* 39-63. doi:10.1002/bsl.679

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14,* 299-311. doi:10.1177/014662169001400307

Si, C. B. (2002). *Ability estimation under different item parameterization and scoring models.* . (Doctoral dissertation). Available from Proquest Dissertation and Theses database. (AAT No. 3118761)

Sinar, E. F., & Zickar, M. J. (2002). Evaluating the robustness of graded response model and classical test theory parameter estimates to deviant items. *Applied Psychological Measurement, 26,* 181-191. doi:10.1177/01421602026002005

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16,* 1-16. doi: 10.1177/014662169201600101

Tate, R. L. (1995). Robustness of the school-level IRT model. *Journal of Educational Measurement, 32*, 145-162. doi:10.1111/j.1745-3984.1995.tb00460

Thissen, D., Chen, W. -H., & Bock, D. (2003). Multilog (version 7). Lincolnwood, IL: Scientific Software International [Computer software].

Toland, M. D. (2008). *Determining the accuracy of item parameter standard error of estimates in BILOG-MG 3.* (Doctoral dissertation). Available from Proquest Dissertation and Theses database. (AAT No. 3317288)

Walton, K. E., Roberts, B. W., Krueger, R. F., Blonigen, D. M., & Hicks, B. M. (2008). Capturing abnormal personality with normal personality inventories: An item response theory approach. *Journal of Personality, 76(6),* 1623-1648. doi: 10.1111/j.1467-6494.2008.00533.x

Wang, W., & Chen, C. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement, 65,* 376-404. doi:10.1177/0013164404268673

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 111-154). Westport, CT: American Council on Education.

Appendix 'A'

```
rep1.MLG -
               ESTIMATION OF ITEM PARAMETERS FOR N=100
>PROBLEM RANDOM, INDIVIDUAL, NEXAMINEES=100, NITEMS=20,
NCHARS=9,
         NGROUPS=1, DATA='C:\Multilog\rep1.dat';
>TEST ALL, GRADED,
NC=(5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5);
>EST NC=500, ICRIT=0.001, CCRIT=0.001;
>SAVE;
>END;
5
12345
11111111111111111111
22222222222222222222
33333333333333333333
44444444444444444444
55555555555555555555
(9A1,20A1)
```

Appendix 'B'

```
rep1.MLG -
              ESTIMATION OF THETA SCORES FOR N=100
>PROBLEM SCORE, INDIVIDUAL, NEXAMINEES=100, NITEMS=20,
NCHARS=9,
         NGROUPS=1, DATA='C:\Multilog\rep1.dat';
>TEST ALL, GRADED,
NC=(5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5);
>EST NC=500, ICRIT=0.001, CCRIT=0.001;
>SAVE;
>END;
5
12345
11111111111111111111
22222222222222222222
33333333333333333333
44444444444444444444
55555555555555555555
(9A1,20A1)
```

Appendix 'C'

```
%macro results
(start,stop,cond,item,sample,item2,type);
ods listing close;

data pirt.truepars&cond;
infile "C:\sasfiles\&item\pars.wgi" firstobs=1
dlm='09'x;
input item model $ cats truea trueb1 trueb2 trueb3
trueb4;
run;

data pirt.estpars&cond;
%do value=&start %to &stop;
infile "C:\sasfiles\results\&cond\rep&value..par"
firstobs=1 obs=&item2;
input @6 aest 7.5 @17 b1est 8.5 @28 b2est 9.5 @40 b3est
9.5 @53 b4est 8.5;
item=_N_;
rep=&value;
output;
%end;
run;

proc sort;
by item;
run;

data work.allparsBIAS&cond;
     merge pirt.truepars&cond pirt.estpars&cond;
     by item;
diff_a_true_est=truea-aest;
diff_b1_true_est=trueb1-b1est;
diff_b2_true_est=trueb2-b2est;
diff_b3_true_est=trueb3-b3est;
diff_b4_true_est=trueb4-b4est;
run;

proc means data=pirt.allparsBIAS&cond;
```

```
var diff_a_true_est diff_b1_true_est diff_b2_true_est
diff_b3_true_est diff_b4_true_est;
by item;
output out=pirt.itemBIAS&cond;
run;

proc sort data=work.allparsBIAS&cond;
by rep;
run;

proc means data=work.allparsBIAS&cond noprint;
var diff_a_true_est diff_b1_true_est diff_b2_true_est
diff_b3_true_est diff_b4_true_est;
by rep;
output out=work.BIAS&cond;
run;

data work.BIAS&cond;
     set work.BIAS&cond;
if _STAT_="MEAN";
run;

ods pdf file="C:\sasfiles\results\BIASpars&cond..pdf";
proc means data=work.BIAS&cond;
var diff_a_true_est diff_b1_true_est diff_b2_true_est
diff_b3_true_est diff_b4_true_est;
title "BIAS for the parameters for &cond and &type";
run;
ods pdf close;

data work.allpars&cond;
     merge pirt.truepars&cond pirt.estpars&cond;
     by item;
absdiff_a_true_est=abs(truea-aest);
absdiff_b1_true_est=abs(trueb1-b1est);
absdiff_b2_true_est=abs(trueb2-b2est);
absdiff_b3_true_est=abs(trueb3-b3est);
absdiff_b4_true_est=abs(trueb4-b4est);
square_a_true_est=(truea-aest)**2;
square_b1_true_est=(trueb1-b1est)**2;
square_b2_true_est=(trueb2-b2est)**2;
```

```
square_b3_true_est=(trueb3-b3est)**2;
square_b4_true_est=(trueb4-b4est)**2;
run;

proc sort data=work.allpars&cond;
by rep;
run;

proc means data=work.allpars&cond noprint;
var square_a_true_est square_b1_true_est
square_b2_true_est square_b3_true_est
square_b4_true_est;
by rep;
output out=work.averagemlg&cond;
run;

data work.averagemlg&cond;
     set work.averagemlg&cond;
if _STAT_="MEAN";
avg_a_true_est=sqrt(square_a_true_est);
if avg_a_true_est=. then avg_a_true_est=0;
avg_b1_true_est=sqrt(square_b1_true_est);
avg_b2_true_est=sqrt(square_b2_true_est);
avg_b3_true_est=sqrt(square_b3_true_est);
avg_b4_true_est=sqrt(square_b4_true_est);
if avg_b1_true_est=. then avg_b1_true_est=0;
if avg_b2_true_est=. then avg_b2_true_est=0;
if avg_b3_true_est=. then avg_b3_true_est=0;
if avg_b4_true_est=. then avg_b4_true_est=0;
run;

ods pdf file="C:\sasfiles\results\RMSEpar&cond..pdf";
proc means data=work.averagemlg&cond;
     var avg_a_true_est avg_b1_true_est avg_b2_true_est
avg_b3_true_est avg_b4_true_est;
title "Root Mean Square Errors of the parameters for
&cond and &type";
run;
ods pdf close;

data pirt.truetheta&cond;
```

```
infile "C:\sasfiles\TTheta\&type\&sample..wge"
firstobs=1 dlm='09'x;
input obs ttheta;
run;

data pirt.esttheta&cond;
%do value=&start %to &stop;
infile "C:\sasfiles\thetaresults\&cond\rep&value..sco"
firstobs=1;
input @5 esttheta 6.3 @16 se 5.3 @23 obs 4.0;
rep=&value;
output;
%end;
run;

data pirt.allthetaBIAS&cond;
     merge pirt.truetheta&cond pirt.esttheta&cond;
     by obs;
diff_theta_true_est=ttheta-esttheta;
proc sort data=pirt.allthetaBIAS&cond;
by rep;
run;

proc means data=pirt.allthetaBIAS&cond noprint;
var diff_theta_true_est;
by rep;
output out=pirt.BIAStheta&cond;
run;

data pirt.BIAStheta&cond;
     set pirt.BIAStheta&cond;
if _STAT_="MEAN";
run;

ods pdf
file="C:\sasfiles\results\BIASthetas&cond..pdf";
proc means data=pirt.BIAStheta&cond;
var diff_theta_true_est;
title "BIAS for thetas for &cond and &type";
run;
ods pdf close;
```

```
data pirt.alltheta&cond;
     merge pirt.truetheta&cond pirt.esttheta&cond;
     by obs;
absdiff_theta_true_est=abs(ttheta-esttheta);
square_theta_true_est=(ttheta-esttheta)**2;
run;

proc sort data=pirt.alltheta&cond;
by obs;
run;

proc means data=pirt.alltheta&cond noprint;
var square_theta_true_est;
by obs;
output out=pirt.averagetheta&cond;
run;

data pirt.averagetheta&cond;
     set pirt.averagetheta&cond;
if _STAT_="MEAN";
avg_theta_true_est=sqrt(square_theta_true_est);
if avg_theta_true_est=. then avg_theta_true_est=0;
run;

ods pdf file="C:\sasfiles\results\RMSEtheta&cond..pdf";
proc means data=pirt.averagetheta&cond;
     var avg_theta_true_est;
title "Root Mean Square Errors of theta for &cond and
&type";
run;
ods pdf close;
ods listing;

proc sort data=pirt.alltheta&cond;
by rep;
run;

proc datasets library=work nolist kill;
run;
%mend results;
```

Appendix 'D'

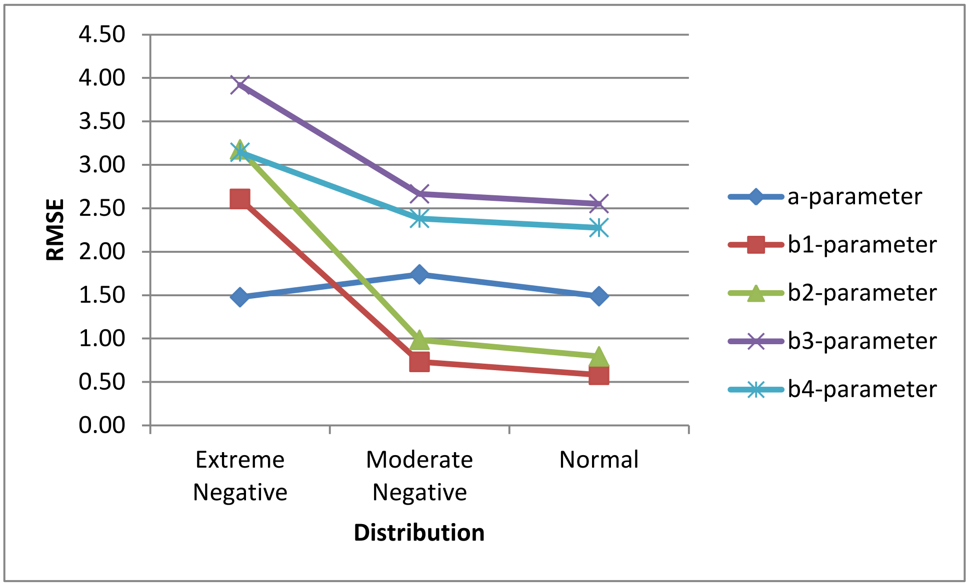

Figure 1. Root Mean Square Errors of parameters by distribution type for n=100.

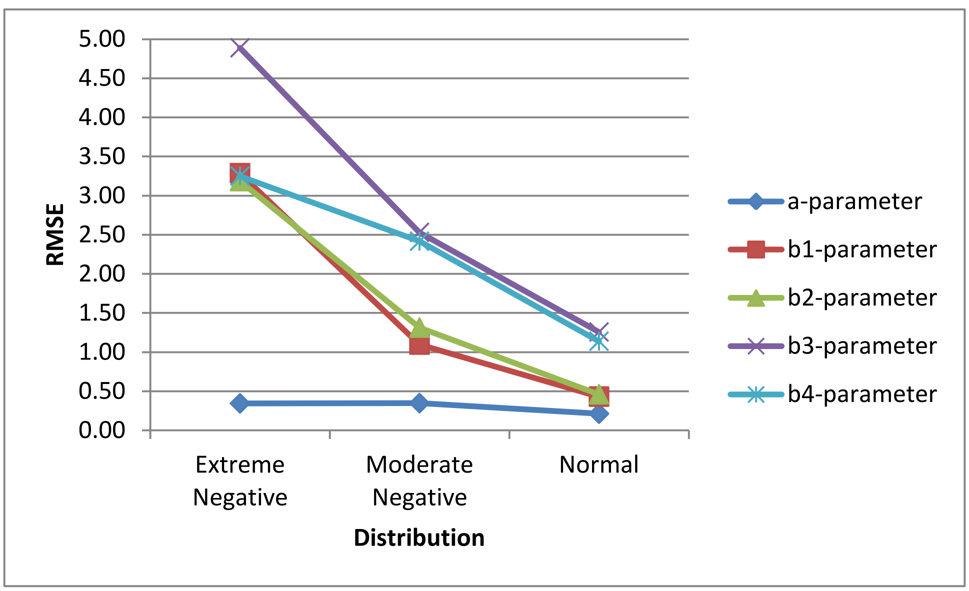

Figure 2. Root Mean Square Errors of parameters by distribution type for n=250.

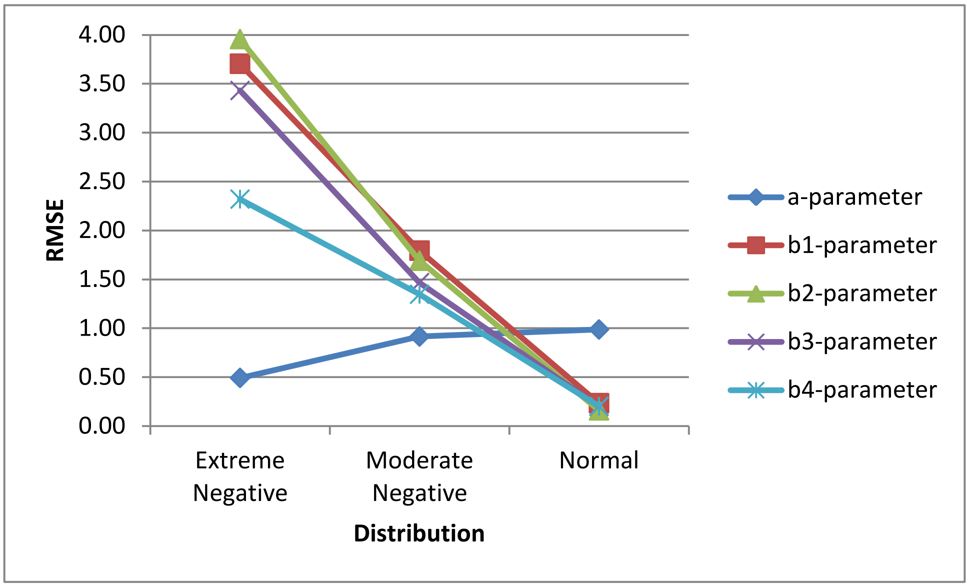

Figure 3. Root Mean Square Errors of parameters by distribution type for n=500.

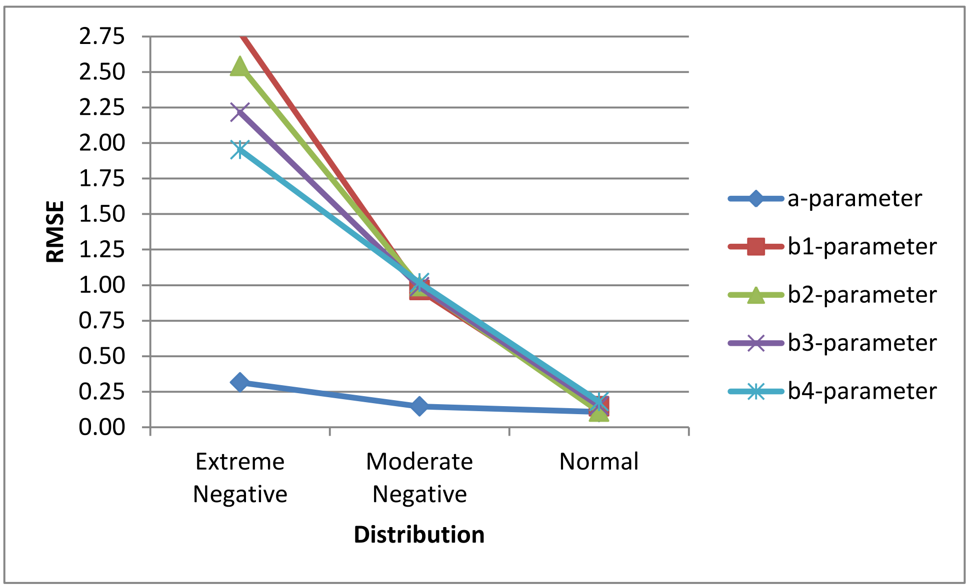

Figure 4. Root Mean Square Errors of parameters by distribution type for n=750.

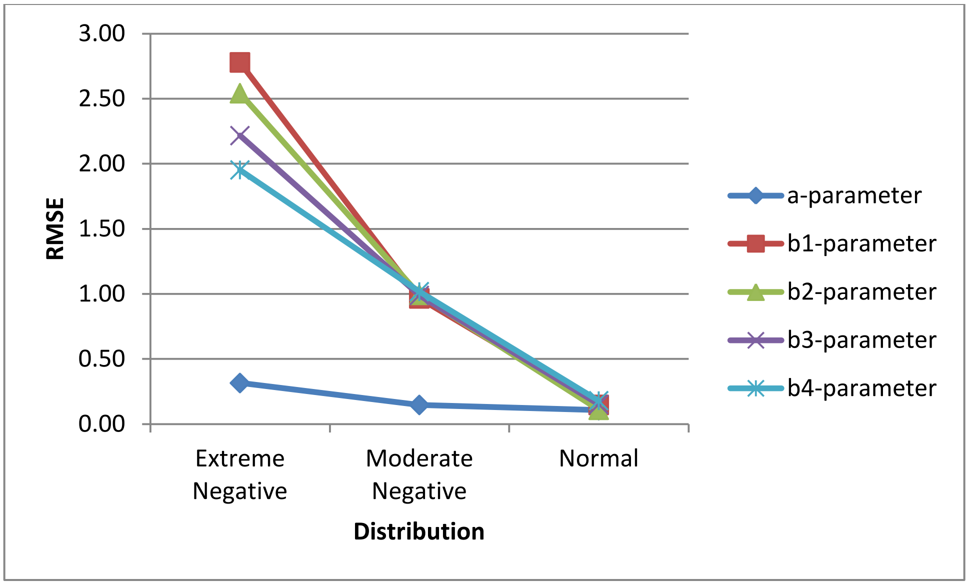

Figure 5. Root Mean Square Errors of parameters by distribution type for n=1000.

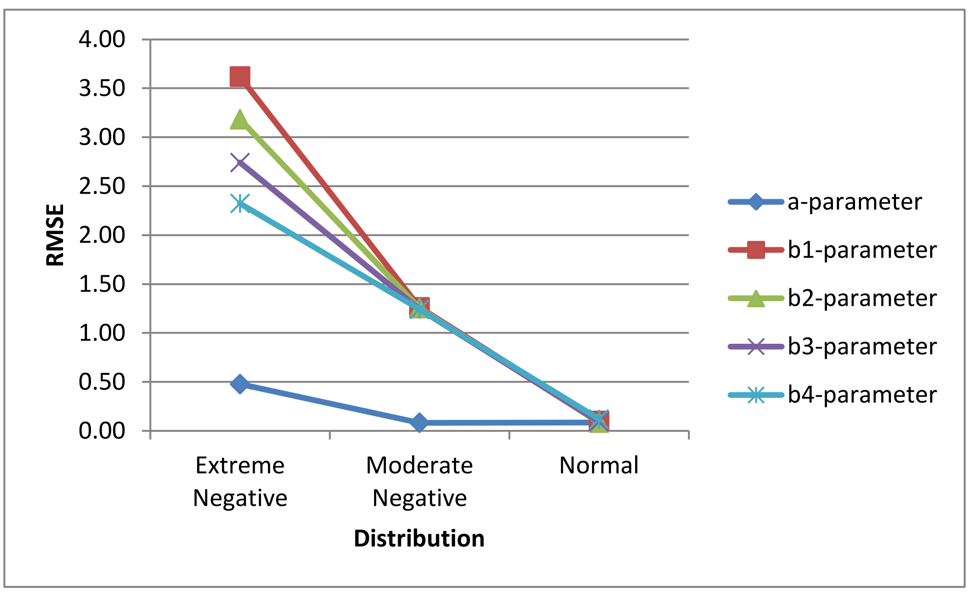

Figure 6. Root Mean Square Errors of parameters by distribution type for n=1500.

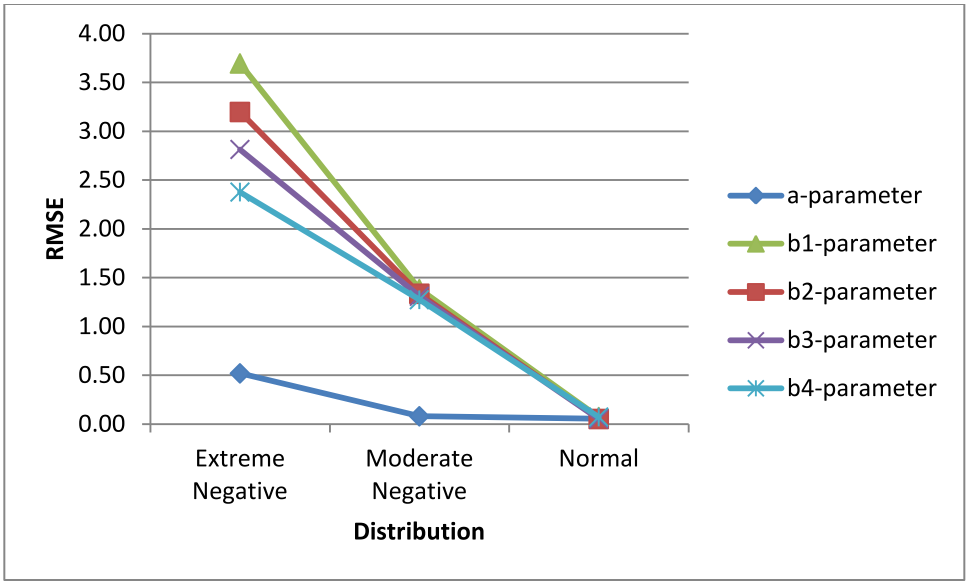

Figure 7. Root Mean Square Errors of parameters by distribution type for n=3000.