DOCUMENT RESUME

ED 347 208 TM 018 740

AUTHOR	Ankenmann, Robert D.; Stone, Clement A.
TITLE	A Monte Carlo Study of Marginal Maximum Likelihood Parameter Estimates for the Graded Model.
SPONS AGENCY	Ford Foundation, New York, N.Y.
PUB DATE	Apr 92
CONTRACT	890-0572
NOTE	39p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 21-23, 1992).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Computer Simulation; *Estimation (Mathematics); Item Bias; *Mathematical Models; *Maximum Likelihood Statistics; Monte Carlo Methods; Sample Size; *Scoring; Statistical Distributions; Test Length
IDENTIFIERS	*Ability Estimates; *Graded Response Model; Item Parameters; MULTILOG Computer Program; One Parameter Model; Research Replication; Two Parameter Model

ABSTRACT

Effects of test length, sample size, and assumed ability distribution were investigated in a multiple replication Monte Carlo study under the 1-parameter (1P) and 2-parameter (2P) logistic graded model with five score levels. Accuracy and variability of item parameter and ability estimates were examined. Monte Carlo methods were used to evaluate marginal maximum likelihood estimates that the MULTILOG computer program produced for the 1P and 2P logistic graded models. Test lengths were 5, 10, and 20 items. Sample sizes were 125, 150, and 500 examinees for the 1P model; and 250, 500, and 1,000 examinees for the 2P model. Item bias and root mean squared error indicate that a minimum sample size of 500 examinees is required for accurate and stable estimates of the 2P graded model item parameters. For both 1P and 2P models, ability distribution and calibration sample size are not important factors in the estimation of ability parameters. Results are discussed in light of a study by S. P. Reise and J. Yu (1990). Together, the 2 studies provide a fairly complete picture of factors that may influence the use of 1P or 2P graded models. Eight figures and 5 tables present analysis results, and there is a 21-item list of references. (SLD)

A Monte Carlo Study of Marginal Maximum Likelihood Parameter Estimates for the Graded Model

by

Robert D. Ankenmann and Clement A. Stone

University of Pittsburgh

Paper presented at the 1992 Annual Meeting of the National Council on Measurement in Education. The preparation of this paper was supported, in part, by the Ford Foundation (grant number 890-0572) for the QUASAR project. The opinions expressed are those of the authors and not necessarily those of the Ford Foundation.

A Monte Carlo Study of Marginal Maximum Likelihood Parameter Estimates for the Graded Model

Robert D. Ankenmann and Clement A. Stone
University of Pittsburgh

With the emerging popularity of performance assessments, there is a rising interest in the use of tests that contain polychotomously scored items. The availability of computer programs such as BIGSTEPS (Wright & Linacre, 1990) and MULTILOG (Thissen, 1988) now renders a wide selection of polychotomous item response theory (IRT) models accessible to measurement practitioners. For example, the polychotomous models implemented by MULTILOG include Samejima's (1969) graded model, a version of Masters' (1982) partial credit model, a multiple choice model (Thissen & Steinberg, 1984), and Bock's (1972) nominal model.

Samejima's (1969) graded model, as implemented in MULTILOG, uses a marginal maximum likelihood (MML) estimation procedure via the EM algorithm. The MML procedure employed by MULTILOG assumes a structure for the ability distribution, typically N(0,1). Thus, the incidental parameter θ is not estimated jointly with item parameters, and asymptotic properties (e.g., consistency) of maximum likelihood (ML) estimates for the item parameters may apply even in small item sets (Mislevy & Stocking, 1989). After MML estimates of the item parameters are obtained, ML estimates of θ can be obtained. If either the IRT model or the assumed ability distribution is incorrect, the statistical properties of the MML estimates may fail to hold (Mislevy & Sheehan, 1989). Recent results from Stone (1990) indicate that skewed ability distributions, in particular, adversely affect MML parameter estimates in the two-parameter logistic IRT model, but the impact of non-normal ability distributions diminishes with increased test length or sample size.

Reise and Yu (1990) used MULTILOG to study the effects of sample size, true ability distribution, and true discrimination parameter distribution on parameter recovery in the two-parameter graded model for tests with 25 items and five score levels per item. They found that sample size had little effect on the recovery of ability parameters, but had an effect on the recovery of item parameters. Sample sizes of at least 500 examinees were recommended to achieve acceptable correlations and root mean squared errors, and sample sizes of 1,000 to 2,000 were recommended when item parameter recovery is crucial. It was concluded that item parameter estimation benefits from the use of highly discriminating items with examinees having heterogeneous ability. The recovery of ability parameters was found to improve as test length increased. Inconsistent effects of true ability distribution were observed. With respect to item parameter estimation, the uniform true ability conditions were found to be slightly superior to the normal and skewed conditions. However, for ability parameter estimation, the uniform true ability conditions yielded inferior estimates compared to those produced for the normal and skewed conditions. Looking at the correlations between true and estimated parameters that were reported by Reise and Yu (1990), it is interesting to note that for small sample sizes there were inconsistencies across true ability conditions. For example, sometimes the correlation corresponding to the normal true ability condition was substantially less than that corresponding to one of the non-normal conditions, and other times the reverse was observed. Such inconsistencies may have been due to the fact that only one set of data was generated and analyzed for each experimental condition (i.e., multiple replications were not employed).

The purpose of the present study was to expand on this research: by investigating the effects of test length, sample size, and assumed ability distribution in the context of a multiple replication Monte Carlo study; and by examining these factors under both the one-parameter (1P) and two-parameter (2P) logistic graded models with five score levels. Furthermore, this study was designed to examine the effect of small test lengths (e.g., 5 and 10 items) on the recovery of ability and item parameters in the graded model. Typically, a small number of open-ended tasks will constitute a performance assessment, whereas traditional multiple choice tests consist of a greater number of items. Specifically, this study examined the accuracy and variability of item parameter and ability estimates.

Method

Monte Carlo methods were used to evaluate the MML estimates that MULTILOG produced for the 1P and 2P logistic graded models with five score levels; that is, two Monte Carlo studies were conducted, one study for each model. The following methodology is described in terms of a single study and was applied to both the 1P and 2P investigations.

Three factors were manipulated: test length (5, 10, and 20 items), sample size (125, 250, and 500 examinees for the 1P model; 250, 500, and 1,000 examinees for the 2P model), and assumed distribution of ability (normal and skewed positive). A test consisting of 10 performance based items was viewed as what students can reasonably respond to in a class period. A test length of 5 was chosen to reflect small test lengths that may occur when tests of dichotomously scored items are restructured into testlets, where each testlet consists of several dichotomously scored items and so becomes treated as one polychotomously scored item (Thissen, Steinberg, & Mooney, 1989). The test length of 20 was chosen as an upper bound for the number of polychotomously scored items that might be administered in a single test. Two considerations governed the selection of the levels of sample size: A sufficiently large sample size was needed to ensure stable parameter estimates, and a suitably small sample size was required to determine the conditions under which parameter estimates become unstable. For the 1P graded model, stable estimates were achieved with a sample size of 500, and a sample size of 125 was required to yield less stable results. Because more parameters are estimated in the 2P graded model, a larger sample size (N=1,000) was required to achieve stable estimates, and a sample size of 250 was small enough to produce less stable estimates. Skewed distributions are commonly found in educational settings when tests do not match the ability of the examinees. Therefore, a positively skewed ability distribution was chosen to represent the condition of non-normal ability to assess the quality of MML parameter estimates under violation of the normality assumption.

All simulated item responses were created as follows: (a) an ability parameter was randomly generated from an assumed distribution, (b) this randomly generated ability parameter and the defined item parameters were used with the graded model to calculate the corresponding probabilities and cumulative probabiiities of scoring at each of the five score levels, and (c) these cumulative probability values were compared to a randomly generated number from a uniform [0,1] distribution. The simulated item response was defined as the highest score level at which the random number was less than or equal to the associated cumulative probability. In Samejima's graded model, these cumulative probabilities must increase as the score level increases. Iteration of this procedure produced the simulated data set corresponding to a particular experimental condition; for a 5-item test written by 125 examinees the data set would consist of 125 simulated item response vectors, each with 5 scores.

Normal ability distributions were generated using the IMSL function RNNOA. The skewed ability distributions represented deviations from a normal distribution and were derived by using a power method described by Fleishman (1978). This method involves the transformation of a standard normal deviate, Z, as follows: $Z' = a + bZ + cZ^2 + dZ^3$; where a, b, c, and d are power method weights. To produce a skewed distribution (skewness=0.75 and kurtosis=0.0) the following coefficients were used: a = -0.1736300195, b = 1.1125146004, c = 0.1736300195, and d = -0.0503344487.

Item discriminations (a_j) and item thresholds (b_{1j}, b_{2j}, b_{3j}, and b_{4j}) for J test items were the defined item parameters. Researchers typically define these parameters in one of two ways: by using estimates from a particular test calibration, or by randomly sampling item parameters. Although random assignment may provide more general results, a disadvantage is that an unusual distribution of a_j or b_{ij} parameters could occur in a test of short length. As well, the combination of a_j and b_{ij} parameters for a particular item could be quite unrealistic. Therefore, in the present study, response data from a 10-item subset of the QUASAR Cognitive Assessment Instrument (QCAI) (Lane, 1991)--a multi-form mathematics test consisting of open-ended reasoning and problem solving tasks--were used to determine the defined item parameters. The items in the subset were chosen to reflect as broad a range of difficulty as was possible: three items were moderately easy (items 1, 2, and 3); three items were moderately to very difficult (items 8, 9, and 10); and the remaining items were of moderate difficulty. MULTILOG was used to calibrate 1P and 2P graded model item parameter estimates for this subset of 10 items. These estimates served as the defined (true) item parameters for the study: they were used to generate simulated data sets, and they were also used as the true parameters against which the estimated parameters were compared. Their values and distributional information are given in Table 1. Note that the 5-item test was obtained by using every even numbered item from the 10 items and the 20-item test was obtained by duplicating the set of 10 items.

Insert Table 1

Basing a Monte Carlo study on estimates from a test calibration of real data may only be valid to the extent that the calibration is itself valid. Several procedures were used to determine whether the data used to define the item parameters conformed to 1P and 2P graded models. Unidimensionality of the QCAI was assessed through the use of confirmatory factor analysis (Lane, Stone, Ankenmann, & Liu, 1992). A one factor model fit the data, providing evidence that the test measured a single mathematics dimension. It was assumed that if the entire test measured a single dimension then a subset of items would also measure a single dimension. For each of the items the goodness of fit between the real data and the graded model was examined by comparing the proportion of examinees who responded to each of the response levels in the real versus simulated data (see Table 2). The simulated data contained the expected responses given that the model was true; that is, based on the defined item parameters. Chi-square statistics of observed versus expected proportions for each of the 1P and 2P models were calculated for each item based on sample sizes of 1,695. The largest chi-square value was χ^2 (4, 1695) = 8.46, $p > .05$. Therefore, none of the chi-square statistics were significant, and it was concluded that both the 1P and 2P logistic graded models adequately fit the data. Note that the observed versus expected proportions were compared at one level of examinee ability (from -5 to 5), rather than subdividing the examinees by ability into five or six levels and then comparing the proportions at each ability level. Accurate classification of the examinees by ability, based on estimates for the tests considered in this study (lengths of 5 and 10 items), was impossible.

Insert Table 2

To justify the use of a 1P graded model with the data, a statistical comparison of the 1P and 2P models estimated by MULTILOG was performed. Because the 1P and 2P models are hierarchical (i.e., the 2P model estimates all the parameters of the 1P model plus additional parameters), the two models may be compared statistically by comparing the "negative twice the loglikelihood" statistic reported by MULTILOG for each model. The difference between the statistics for the hierarchical models is distributed as chi-square (Thissen, Steinberg, & Gerrard, 1986) and may be used to calculate the significance of specifying additional parameters in the model. The difference of the "negative twice the loglikelihood" between the 1P and 2P models was χ^2 (9) = 8.4, $p >$ .40. Because the difference chi-square was not significant, the additional item discrimination parameters estimated in the 2P model did not significantly improve model-data fit. Therefore, although a 2P model could be estimated, a 1P model was also appropriate.

For each of the 18 different experimental conditions associated with a particular Monte Carlo study--three levels of test length, three levels of sample size, and two levels of true ability distribution--100 data sets were generated. For example, 100 data sets were generated for the experimental condition consisting of 125 examinees, 5 items, and normal true ability distribution. A total of 1,800 data sets were analyzed. A different "seed" (starting value for the random number generator) was used for each of the 18 experimental conditions. The results may be less comparable across conditions but they are less dependent upon specific seed values and the sampling results are independent of each other.

Simulated data sets were calibrated using MULTILOG. To minimize computer time, the defined model parameters that were used to generate simulated data sets were used as the starting values for item parameter estimation with MULTILOG. Maximum likelihood estimates of ability were then obtained by again running MULTILOG, but with item parameters fixed at their estimated values. It could be argued that the use of true values as start values may spuriously avoid the problem of local maxima. However, this was not considered to be a major concern, because a noted strength of the EM algorithm is that the choice of starting values is not critical (Bock, 1991).

Before the results from MULTILOG could be compared against true values, it was necessary that a common metric underlie both the estimated and true values of the item and ability parameters. The estimates from MULTILOG were placed on the same metric as the true values using the computer program EQUATE (Baker, 1991). This program obtains equating coefficients (slope and intercept adjustments) using Stocking's and Lord's (1983) procedure of minimizing the difference between the test characteristic curves for the items that are common to the target test and the test to be transformed. In the case of a Monte Carlo study, the target test consists of the known parameter values, and the number of common items is equal to the number of items in the data set being analyzed. After the equating coefficients are obtained, a simple linear transformation is performed on the parameter estimates to place them on the same scale as the true values (Baker, 1991). Because a linear transformation is used to perform the equating, the probabilities of scoring at each score level remain the same whether rescaled or non-rescaled ability and item parameter estimates are substituted into the IRT model.

The defined ability and item parameters that were used to generate a simulated data set were also used as true parameters against which estimated ability and item parameters were compared. The MML item parameter and ability estimates were evaluated using two criteria: the bias of the estimate, and the root mean squared error (RMSE) of the estimate. Recovery of item parameter values was assessed by averaging information across 100 replications. The use of multiple replications allowed for analyses based on statistics computed across replications as opposed to most IRT Monte Carlo research which utilizes a single data set and analyses based on statistics computed across items. Results for a single data set can be particularly misleading when the sample size is small or the test length is short.

Bias in each a_j was assessed by examining the difference between the mean of $\hat{a}_j$ across 100 replications and a_j:

$$\text{bias in } a_j = \left(\sum_{k=1}^{100} \hat{a}_{jk} \right) / 100 - a_j \, , \tag{1}$$

where k references the replication and j the item. Bias in b_{ij} was similarly assessed:

$$\text{bias in } b_{ij} = \left(\sum_{k=1}^{100} \hat{b}_{ijk} \right) / 100 - b_{ij} \, , \tag{2}$$

where j and k are defined as above and i references the score category boundary. Item parameter recovery was also assessed by examining the RMSE for each a_j or b_{ij} across 100 replications. The formulae are presented below:

$$\text{RMSE } a_j = [(1/100) \sum_{k=1}^{100} (\hat{a}_{jk} - a_j)^2]^{1/2} \, , \tag{3}$$

$$\text{RMSE } b_{ij} = [(1/100) \sum_{k=1}^{100} (\hat{b}_{ijk} - b_{ij})^2]^{1/2} \, . \tag{4}$$

The recovery of ability estimates was also assessed by examining bias and RMSE; however, this information was averaged across subjects within each replication. The formulae are presented below:

$$\text{bias in } \theta_k = (1/N) \sum_{n=1}^{N} (\hat{\theta}_n - \theta_n) \, , \tag{5}$$

$$\text{RMSE } \theta_k = [(1/N) \sum_{n=1}^{N} (\hat{\theta}_n - \theta_n)^2]^{1/2} \, , \tag{6}$$

where k references the replication number (between 1 and 100), n references the examinee number, and N is the sample size. By examining both bias and RMSE, it was possible to consider the accuracy and variability of the point estimates. The use of bias as a measure of accuracy and RMSE as a measure of variability of point estimates precluded the need to employ correlations as evaluation ~riteria. It was felt that correlations, which only indicate the extent to which rank ordering is maintained, would be inferior to the more direct and informative measures of accuracy and variability.

Results

Ancillary Results: 2P Study

The following ancillary results from the MULTILOG analyses are given in Table 3: the average number of iterations, the average posterior mean and standard deviation of the quadrature distribution at the final iteration, and the average and standard deviation of the slope and intercept equating coefficientr.

Insert Table 3

Fewer iterations were required as sample size increased and as test length decreased from 20 items to a test of length 5 or 10. There were small differences between the number of iterations required for normal and positively skewed ability distributions; however, there did not appear to be a systematic pattern to the size or the direction of these differences across sample size and test length. The posterior mean moved further from 0 and the posterior standard deviation increased as sample size increased; the differences in the posterior means, however, were negligible. As sample size and test length increased, the mean of the slope equating coefficient moved further from 1 and the standard deviation decreased. The mean of the intercept equating coefficient tended to remain stable and close to 0. The standard deviation of the intercept equating coefficient decreased as sample size increased, and tended to increase slightly as test length increased. Although not reported in Table 3, it should be noted that the slope equating coefficient varied from 0.821 to 1.221 across 100 replications of a particular experimental condition, and the intercept equating coefficient varied fror:: -0.212 to 0.276 across another. These variations illustrate the importance of multiple replications and rescaling. A single replication could yield an extreme data set, not typical and therefore not representative; and the item parameter estima.es produced by a particular data set could be on a metric quite different from the true parameters, thus making the comparison of true and estimated parameters spurious.

Item Parameter Recovery: 2P Study

The signed bias in the slope and threshold parameteıs was calculated using equations (1) and (2), previously defined. To facilitate the interpretation of results the mean absolute bias of the slope parameter corresponding to a particular experimental condition was calculated by averaging the absolute bias of the slope parameters across the items that were common to the three test lengths. Similarly, the mean absolute bias of each of the four threshold parameters was averaged across common items. Averaging absolute bias across common items had two advantages: results were summarized, hence easier to assimilate; an results were reported for the same set of items across conditions, thus facilitating comparisons.

The mean absolute bias and RMSE for the slope parameters (a_j) are summarized in Figure 1. Bias and RMSE diminished as sample size increased. In general, the decrease was larger when sample size increased from 250 to 500 than when it increased from 500 to 1,000, a trend that was observed in both the normal and skewed conditions across all test lengths. For test lengths of 10 and 20 items, RMSE was very nearly the same but increased when test length dropped to 5 items. The amount of increase in RMSE from the 10- and 20-item conditions to the 5-item condition diminished as sample size increased. For a sample size of 1,000 the difference in RMSE between 5 items and 10 or 20 items was negligible. The same test length effect was observed in the bias of the slope parameters, but the test length by sample size interaction that was observed in the RMSE was less pronounced for the bias. There was no distributional effect on RMSE, and only a slight but negligible effect was observed for bias.

Insert Figure 1

Mean absolute bias in the threshold parameters (b_{1j}, b_{2j}, b_{3j}, and b_{4j}) are shown in Figure 2. The bias of the three lowest thresholds (b_{1j}, b_{2j}, and b_{3j}) was low and stable across test length and sample size conditions for normal ability distributions. For sample sizes of 500 and 1,000, and for all test lengths under the normal true ability condition, the bias of the highest threshold (b_{4j}) was also low and stable; however, for a sample size of 250 the bias was noticeably higher. This may be an artifact attributable to the extreme true value of the b_4 threshold in the eighth item (i.e., b_{48} = 3.458). No upper limit was imposed on the parameter estimates calibrated by MULTILOG. Therefore, under those simulation conditions which included the smallest sample size (N=250) it was not uncommon for a b_{48} threshold as large as 7.0 to appear in at least a few replications. Bias of the threshold estimates under the skewed true ability conditions was generally higher than for the normal true ability conditions. In addition, bias of the b_{1j}, b_{2j}, and b_{3j} thresholds under the skewed condition was higher for the 5-item test length than for the 10- or 20-item test lengths.

Insert Figure 2

Root mean squared error in the threshold parameters are reported in Figure 3. For all thresholds and at all levels of test length and ability distribution, RMSE decreased as sample size increased. As would be expected, the extreme thresholds (b_{1j}, b_{4j}) had higher RMSEs than the two middle thresholds (b_{2j}, b_{3j}). As was seen for the bias, the large RMSE observed in the b_{4j} thresholds for the experimental condition consisting of normal ability distribution, N=250, and 5-item test length, may be an artifact attributable to the fact that no upper limit was placed on parameter estimates; the presence of high estimates for the b_{48} threshold in a few of the replications would of course result in higher variability of the estimates. Neither test length nor ability distribution effecis were observed in the RMSEs of the b_{1j}, b_{2j}, or b_{3j} threshold parameters.

Insert Figure 3

Results concerning the direction of bias in each of the parameters, for all items, were tabulated. The proportion of negative bias values (i.e., proportion of times $a_j - \bar{a}_j < 0$ across J items, where $\bar{a}_j$ is the mean of the $\hat{a}_j$ across 100 replications) are given in Table 4. Systematic positive or negative bias is indicated by a disproportionate number of positive or negative bias values. Positive and negative bias in a_j were determined by looking at the signed bias value that was calculated for each item by equation (1).

Insert Table 4

Four trends were observed: (a) the proportion of negative a_j bias was generally low, indicating positive bias (i.e., overestimation); (b) as sample size increased, the proportion of negative a_j bias values increased; (c) as test length increased, the proportion of negative a_j bias values remained fairly constant; and (d) the proportion of negative a_j bias values for the skewed ability distributions was always greater than or equal to the proportions for the N(0,1) distributions. As Lord (1983) indicated, it is not surprising to find positive bias in the slope parameter estimates. However, the results reported here indicate that the positive bias can be reduced by increasing sample size.

Although not reported in tabular form, the direction of the bias in each of the threshold parameters (b_{1j}, b_{2j}, b_{3j}, and b_{4j}) was also examined. For the lowest threshold (b_{1j}) the bias tended to be negative, and when there was positive bias it appeared only when the true ability distribution was normal. For normal true ability distributions, the amount of negative bias tended to decrease as test length increased; for the 20-item test length and normal true ability conditions, bias was slightly positive across all sample sizes. Note that all of the true b_{1j} parameters were negative. Bias in the b_{2j} and b_{3j} thresholds was almost always positive; recall that six of the true b_{2j} parameters were negative, and only one true b_{3j} parameter was negative. For the b_{4j} thresholds there tended to be positive bias under the normal true ability distributions and negative bias under the skewed distributions; all of the true b_{4j} parameters were positive.

Ability Parameter Recovery: 2P Study

Bias and RMSE in ability parameter recovery were calculated using equations (5) and (6), respectively. As well, the correspondence between the true ability and estimated ability distributions were examined by using the following statistics: mean, standard deviation, skewness coefficient, and kurtosis coefficient. These statistics are presented in Table 5.

Insert Table 5

Several interesting trends are noteworthy. Under all conditions of test length, sample size, and true ability distribution, the distribution of ability estimates was always platykurtic and had a smaller standard deviation than the corresponding distribution of true ability. Under both conditions of true ability, normal and skewed positive, these deviations from the true distribution diminished as test length increased, but not as sample size increased. Also, for all of the normal true ability distributions, the estimated ability distributions were positively skewed. In all but one case (*N*=250, 10 items, N(0,1)) this deviation diminished with increased test length but not with increased sample size.

That the standard deviation was smaller in the estimated ability distributions than in the true ability distributions is probably due to the fact that the range in true traits was from -5 to 5 but considerably narrower in the estimated traits. This seems to indicate that there was an underestimation of extreme abilities, in an absolute sense. That is, the highest positive ability estimates were not as extreme (high) as their corresponding true values, and the lowest negative ability estimates were not as extreme (low) as their corresponding true values.

For the skewed true ability conditions, the estimated distributions exhibited a lesser degree of skew than the true distributions. The correspondence between estimated and true skewed distributions improved as test length increased, but remained constant as sample size increased. The fact that the amount of skew in the estimated distributions was less than in the true distributions may be due to the fact that MULTILOG assumes a normal N(0,1) prior on the ability distribution.

Bias and RMSE in ability parameter recovery are shown in Figures 4 and 5, respectively. In Figure 4, it can be seen that ability decreased as test length increased, for all sample sizes and true ability distributions, when looking at all ranges of ability except for $-5 \le \theta \le 5$. Note that the bias results for the whole ability range $-5 \le \theta \le 5$ are not particularly informative (i.e., all biases in this range are close to 0) due to the offsetting effect of positive and negative bias values which occur in the narrower ability ranges (e.g., $-2 \le \theta \le -1$ vs. $1 \le \theta \le 2$; $-1 \le \theta \le 0$ vs. $0 \le \theta \le 1$). The amount of bias and the difference in bias among the various test length conditions decreased for abilities in the range $-1 \le \theta \le 1$, where bias was less than 0.15 under all conditions of test length, sample size, and ability distribution; furthermore, the differences in bias among the various test length conditions was very small. As ability became more extreme, the amount of bias and the difference in bias among the various test lengths increased.

Insert Figure 4

Turning to Figure 5, the effect of test length on the RMSE of ability was similar to that of bias: RMSE of ability decreased as test length increased. For abilities in the range $\theta > 2$ the size of the RMSE and the difference in RMSE among different test lengths was larger than for abilities in the range $-2 \le \theta \le 2$. For all ability levels (i.e., $-5 \le \theta \le 5$) the RMSE for the 5-item test length condition was about 0.5. For the 20-item test length condition, considering only the five items that were common with the 5-item test length condition, the RMSE was about 0.3.

Insert Figure 5

Item Parameter and Ability Parameter Recovery: 1P Study

Results from the 1P study were examined in the same way as those that were presented for the 2P study. However, because the presentation and discussion of the figures would be lengthy, and few effects of the manipulated factors were observed, an overview of results is given below.

With respect to the slope parameter (a_j), mean absolute bias was negligible (less than 0.02) for all levels of test length, sample size, and true ability distribution. The RMSE in the slope parameter was also small (less than 0.05) across all levels of test length, sample size, and ability distribution. Although the values of bias and RMSE were small for all sample sizes, they did decrease as sample size increased.

Mean absolute bias in the three lowest threshold parameters (b_{1j}, b_{2j}, and b_{3j}) was small (less than 0.02) and fairly stable for all test lengths and sample sizes under the normal true ability condition. A slight sample size effect, in which bias decreased as sample size increased, was observed. Bias in the b_{2j} and b_{3j} thresholds, for the skewed true ability conditions, was also small (less than 0.02) and stable across all test lengths and sample sizes. Bias in the b_{1j} threshold under the skewed true ability condition ranged from 0.01 to 0.04, and a small test length effect was observed, in which bias was smallest for the 20-item test length conditions and larger for the 5- and 10-item conditions. As in the 2P study, bias in the extreme threshold b_{4j} was considerably less stable than for the other thresholds. For the normal true ability conditions, bias was considerably higher for sample sizes of N=125; and the bias associated with sample sizes of N=250 and N=500 was less than 0.02, and somewhat more stable across sample size and test length conditions. Bias in the b_{4j} threshold under the skewed true ability conditions was unstable and ranged from 0.01 to 0.05, and demonstrated no consistent differences among the various sample sizes and test lengths.

Root mean squared error in the threshold parameters of the 1P study showed trends that were very similar to those in the 2P study. For all thresholds, at all levels of test length and true ability distribution, RMSE decreased as sample size increased. The extreme thresholds (b_{1j}, b_{4j}) had higher RMSEs than the two middle thresholds (b_{2j}, b_{3j}).

Finally, bias and RMSE in the ability parameters of the 1P study were virtually the same as those reported for the 2P study. There was a test length effect in which bias and RMSE decreased as test length increased.

Discussion

This study examined the recovery of MML ability and item parameter estimates produced by MULTILOG under the 1P and 2P logistic graded models. Test length, sample size, and true ability distribution were manipulated factors. The accuracy and variability of item parameter and ability estimates were examined with bias and RMSE statistics. These results suggest several implications for measurement practitioners.

Item parameter bias and RMSE in the 2P study indicate that a minimum sample size of 500 examinees is required to obtain accurate and stable estimates of the 2P graded model item parameters. This conclusion is consistent with the Reise and Yu (1990) findings involving test lengths. For a sample size of 500, if the ability distribution is normal, test lengths as small as 5 items will yield slope and threshold parameter estimates that are just as accurate and stable as those produced by test lengths of 10 or 20 items. When the ability distribution is skewed, increasing the sample size to 1,000 examinees produces slope estimates that are as accurate and stable as those produced for normal true ability distributions. Based on Seong's (1990) work with dichotomous IRT models, increasing the number of quadrature points to 20 may also help to minimize the effect of non-normal true ability distributions. With regard to the threshold estimates, however, the same gain in accuracy is not obtained with a sample size of 1,000. Therefore, it is important to consider the nature of the ability distribution when deciding whether a sample size as low as 500 will be adequate to achieve accurate and stable item parameter estimates. For the 1P graded model a minimum sample size of 250 is required to obtain accurate and stable item parameter estimates.

For both the 1P and 2P models, ability distribution and calibration sample size are not important factors in the estimation of ability parameters. This conclusion, too, is consistent with Reise and Yu (1990). Sample size is not a factor in the estimation of θ because, as Seong (1990) noted, θ is estimated for each examinee separately without consideration of the sample size. As expected, the accuracy of ability parameter estimates increases and the variability decreases as test length increases. Comparing the bias and RMSE results from this 2P graded model study with those of Stone's (1990) 2P dichotomous model study, it is interesting to observe that for 5-item tests the 2P graded model with five score levels will yield ability estimates having the same accuracy and variability as those produced by the 2P dichotomous model having double the test length. However, this benefit to the graded model test items decreases as test length increases.

Some important differences between the results of the present study and those of the Reise and Yu (1990) study can be identified. The effect of skewed true ability distribution on threshold parameter estimation that is reported in this study was not observed by Reise and Yu. This may be due to the fact that the test length used by Reise and Yu was long, and fixed at 25 items. In the present study, the skewness effect was observed in the short 5- and 10-item test length conditions. Another important observation from the present study, that was not reported by Reise and Yu, is that extreme abilities are underestimated in an absolute sense (i.e., the range of the estimated ability distribution tends to be smaller than that of the true ability distribution). Typically, the normal true ability distributions ranged from $-5 \leq \theta \leq 5$, whereas the corresponding estimated ability distributions ranged from $-3 \leq \theta \leq 3$. Similarly, the positively skewed estimated ability distributions were truncated, as compared to the corresponding true ability distributions.

Another important difference between the present study and that of Reise and Yu (1990) concerns the rescaling of parameter estimates. As previously mentioned, the present study utilized Stocking's and Lord's (1983) procedure of minimizing the differences between the test characteristic curves, as implemented in Baker's (1991) EQUATE program, to place item parameter estimates on the same metric as true parameter values. By contrast, Reise and Yu did not place parameter estimates on the same metric as the true parameters, for the normal and skewed ability conditions. In the present study, the mean of the slope transformation coefficient (m) across replications was observed to be consistently close to 0.9 (see Table 3). Therefore, the rescaled slope parameter estimates, obtained using the formula $a_j^* = a_j / m$, would tend to be larger that the non-rescaled estimates. The mean of the intercept transformation coefficient (k) across replications was almost always very close to 0 (see Table 3). Therefore, the rescaled threshold parameter estimates, obtained using the formula $b_{ij}^* = m(b_{ij}) + k$, would almost always be smaller than the non-rescaled estimates.

It is important to note that the quality of the correlations reported by Reise and Yu (1990) were not compromised by the non-equivalence of true and estimated parameter metrics. The rescaling of parameter estimates employs a linear transformation, which would preserve rank ordering. Therefore, the rescaling of Reise's and Yu's parameter estimates to achieve metric equivalence with true parameters would probably not alter their conclusions. However, RMSEs are affected by rescaling. In particular, correct RMSEs can only be obtained when a rescaling of parameter estimates that places them on the same metric as the true parameters is performed. If the purpose of a study is simply to ascertain whether or not certain factors have an effect on parameter estimates, or to compare the effects of different estimation procedures (e.g., MML vs. JML), then the use of unequated parameter estimates may be adequate. However, if the purpose of the study is to understand the errors in estimation that occur because of certain factors, then the use of rescaled estimates becomes relevant.

The results of this study, in conjunction with the results of Reise and Yu (1990), provide a fairly complete picture of the factors which may influence the use of the 1P or 2P graded models. Considering the two studies, a variety of test lengths, sample sizes, assumed ability distributions, and true slope parameter distributions has been evaluated. Nonetheless, as with all Monte Carlo research, other studies are needed to establish the generality of the results. Also, other factors may be relevant to those who wish to use the graded model. As Reise and Yu noted, the effect of the number of score levels on the MML parameter estimates produced for the graded model should also be studied. Finally, it may not be the case that the factors identified and investigated in this study and in the Reise and Yu study have the same influence on parameter estimation in other polychotomous IRT models. Other models, such as the partial credit model (Masters, 1982), should also be studied.

References

Baker, F. B. (1991). *Equating under the graded response model.* Unpublished manuscript, University of Wisconsin.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R. D. (1991, April). *Item parameter estimation.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43,* 521-532.

Lane, S. (1991, April). *The conceptual framework for the development of a mathematics assessment instrument for QUASAR.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1992, April). *Empirical evidence for the reliability and validity of performance assessments.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

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

Lord, F. M. (1983). Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika, 48,* 205-217.

Masters, G. A. (1982). Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in the estimation of item parameters. *Psychometrika, 54,* 661-679.

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

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27,* 133-144.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

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

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

Stone, C. A. (1990, April). *An evaluation of marginal maximum likelihood estimates via the EM algorithm in the 2-parameter logistic response model.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.

Thissen, D. (1988). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory* (Version 5.1). Mooresville, IN: Scientific Software.

Thissen, D., & Steinberg, L. (1984). A model for multiple choice items. *Psychometrika, 49,* 501-519.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118-128.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26,* 247-260.

Wright, B. D., & Linacre, J. M. (1990). *A user's guide to BIGSTEPS.* Chicago, IL: MESA.

Table 1

Item Parameters for the Monte Carlo Study

	One-parameter Model 10-item Test					Two-parameter Model 10-item Test				
Item	a_j	b_{1j}	b_{2j}	b_{3j}	b_{4j}	a_j	b_{1j}	b_{2j}	b_{3j}	b_{4j}
1	1.284	-1.100	-0.887	-0.156	0.791	1.516	-1.003	-0.810	-0.145	0.720
2	1.284	-1.925	-0.972	0.352	1.403	1.266	-1.940	-0.979	0.355	1.414
3	1.284	-1.221	-0.455	0.214	1.116	1.353	-1.182	-0.441	0.207	1.080
4	1.284	-2.072	-0.110	0.782	1.679	1.251	-2.123	-0.113	0.802	1.720
5	1.284	-1.926	-0.015	1.039	1.510	1.280	-1.930	-0.015	1.041	1.513
6	1.284	-1.180	-0.070	0.808	1.715	1.932	-0.939	-0.054	0.647	1.364
7	1.284	-1.171	0.259	0.818	1.421	1.164	-1.257	0.278	0.877	1.529
8	1.284	-0.887	0.048	0.712	3.337	1.213	-0.905	0.053	0.733	3.458
9	1.284	-1.271	0.116	1.359	2.582	0.982	-1.546	0.135	1.650	3.160
10	1.284	-0.140	0.649	1.109	1.917	1.225	-0.138	0.663	1.131	1.957
avg	1.284	-1.289	-0.144	0.704	1.747	1.318	-1.296	-0.128	0.730	1.792
sd	0	0.575	0.500	0.452	0.734	0.254	0.604	0.494	0.507	0.870

5-item Test: Every even numbered item from the 10-item test was used.

avg	1.284	-1.241	-0.091	0.753	2.010	1.377	-1.209	-0.086	0.734	1.983
sd	0	0.791	0.580	0.271	0.764	0.311	0.819	0.587	0.280	0.859

20-item Test: The items from the 10-item test were repeated.

avg	1.284	-1.289	-0.144	0.704	1.747	1.318	-1.296	-0.128	0.730	1.792
sd	0	0.559	0.487	0.440	0.715	0.248	0.588	0.481	0.494	0.847

Table 2

Model-Data Fit

		Proportion of Responses at Each Score Level							
Item	Data Source	0	1	2	3	4	chi-sq	df	p
1	observed	.244	.041	.163	.225	.328			
	1P model	.233	.042	.168	.210	.347	4.75	4	$p > .30$
	2P model	.228	.051	.162	.212	.347	8.35	4	$p > .05$
2	observed	.124	.152	.306	.216	.201			
	1P model	.138	.153	.293	.199	.216	7.62	4	$p > .10$
	2P model	.139	.152	.291	.202	.215	7.24	4	$p > .10$
3	observed	.232	.160	.160	.198	.249			
	1P model	.222	.161	.169	.201	.247	1.69	4	$p > .70$
	2P model	.223	.163	.166	.201	.248	1.16	4	$p > .80$
4	observed	.107	.375	.208	.154	.156			
	1P model	.105	.365	.198	.158	.174	4.71	4	$p > .30$
	2P model	.104	.363	.202	.160	.171	3.73	4	$p > .40$
5	observed	.125	.371	.239	.083	.182			
	1P model	.120	.356	.253	.088	.183	3.23	4	$p > .50$
	2P model	.120	.358	.251	.089	.183	2.81	4	$p > .50$
6	observed	.241	.244	.204	.158	.153			
	1P model	.224	.242	.206	.172	.156	4.28	4	$p > .30$
	2P model	.228	.243	.203	.163	.163	2.57	4	$p > .60$
7	observed	.241	.332	.126	.110	.191			
	1P model	.241	.304	.131	.115	.209	7.69	4	$p > .10$
	2P model	.241	.303	.130	.121	.206	8.46	4	$p > .05$
8	observed	.295	.216	.155	.305	.029			
	1P model	.289	.224	.154	.302	.031	0.98	4	$p > .90$
	2P model	.291	.219	.157	.303	.029	0.23	4	$p > .99$
9	observed	.216	.310	.273	.139	.062			
	1P model	.230	.299	.263	.146	.063	3.37	4	$p > .40$
	2P model	.223	.302	.269	.145	.061	1.28	4	$p > .80$
10	observed	.457	.188	.098	.127	.130			
	1P model	.434	.184	.110	.129	.143	6.49	4	$p > .10$
	2P model	.433	.183	.109	.130	.144	6.79	4	$p > .10$

Note: All data sets had a sample size of 1,695.

Table 3

Results From 2P Graded Model MULTILOG Analyses

	5-item Test		10-item Test		20-item Test	
	N(0,1)	Skewed +	N(0,1)	Skewed +	N(0,1)	Skewed +
N = 250						
avg. cycles	19	17	17	17	26	27
post. mean	0.000	0.002	-0.002	0.004	-0.012	-0.028
post. sd	1.002	1.001	1.009	1.008	1.061	1.074
m mean[a]	0.980	0.990	0.980	0.978	0.925	0.914
m sd	0.074	0.075	0.051	0.053	0.047	0.038
k mean[b]	0.010	-0.010	0.003	-0.020	0.018	0.019
k sd	0.080	0.073	0.065	0.070	0.094	0.107
N = 500						
avg. cycles	16	15	14	15	21	23
post. mean	0.000	0.000	0.001	-0.002	0.008	-0.005
post. sd	1.001	1.000	1.010	1.011	1.067	1.080
m mean	0.996	1.009	0.980	0.984	0.923	0.917
m sd	0.053	0.053	0.038	0.039	0.032	0.027
k mean	-0.007	-0.014	-0.001	-0.006	-0.001	0.002
k sd	0.056	0.049	0.050	0.053	0.066	0.071
N = 1,000						
avg. cycles	13	13	13	14	20	21
post. mean	0.000	0.000	0.001	0.002	0.003	-0.001
post. sd	1.001	1.000	1.011	1.010	1.072	1.081
m mean	0.995	1.014	0.976	0.988	0.920	0.915
m sd	0.034	0.038	0.027	0.029	0.020	0.018
k mean	-0.003	-0.011	0.005	-0.012	-0.002	-0.003
k sd	0.040	0.036	0.032	0.038	0.048	0.055

[a] *m* denotes the slope adjustment equating coefficient.

[b] *k* denotes the intercept adjustment equating coefficient.

Table 4

Proportion of Times the Bias of a_j was Negative

	N = 250	*N* = 500	*N* = 1,000
5-item Test			
N(0,1)	.00	.20	.40
Skewed +	.20	.40	.60
10-item Test			
N(0,1)	.10	.30	.20
Skewed +	.10	.40	.50
20-item Test			
N(0,1)	.00	.30	.40
Skewed +	.05	.40	.55

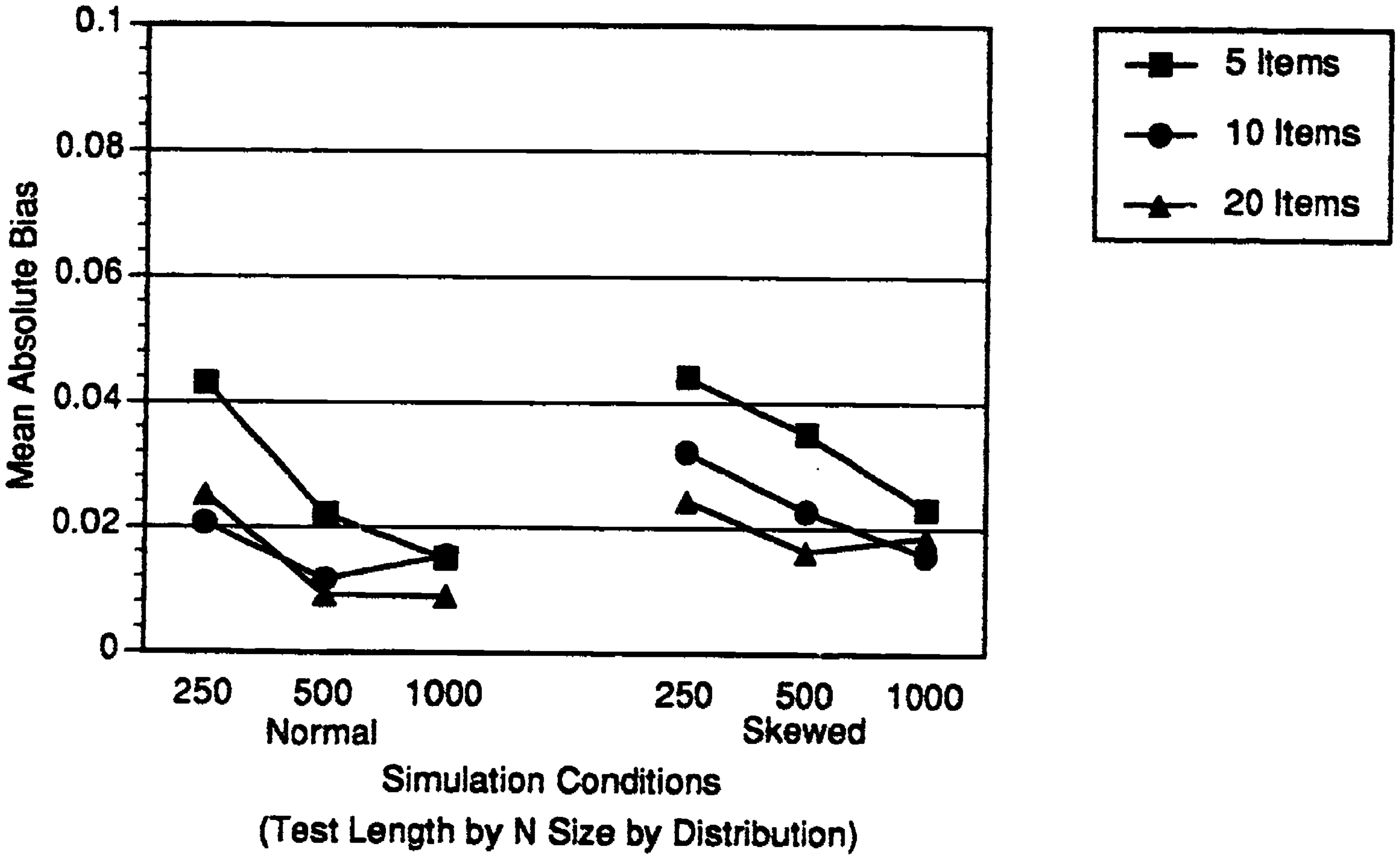
Table 5

Distributional Information for the True and Estimated Abilities

	N = 250		*N* = 500		*N* = 1,000	
	N(0,1)	Skewed	N(0,i)	Skewed	N(0,1)	Skewed
5-item Test						
True Traits						
mean	-0.001	0.000	-0.004	-0.003	-0.003	0.003
sd	0.991	0.999	1.001	0.996	0.999	1.000
skew	0.014	0.749	-0.001	0.736	0.003	0.751
kurtosis	-0.032	-0.015	-0.057	-0.023	-0.011	0.011
Est. Traits						
mean	0.017	0.000	0.000	-0.005	0.005	-0.003
sd	0.834	0.842	0.843	0.854	0.839	0.858
skew	0.067	0.352	0.061	0.330	0.066	0.342
kurtosis	-0.340	-0.334	-0.358	-0.382	-0.395	-0.382
10-item Test						
True Traits						
mean	0.005	-0.006	-0.004	0.000	0.005	-0.003
sd	0.998	0.994	1.002	1.002	0.995	1.001
skew	-0.010	0.744	0.012	0.750	-0.002	0.753
kurtosis	0.002	-0.018	0.029	-0.008	-0.028	0.014
Est. Traits						
mean	0.004	-0.014	0.002	-0.005	0.009	-0.008
sd	0.892	0.890	0.890	0.895	0.886	0.897
skew	0.008	0.457	0.036	0.461	0.027	0.451
kurtosis	-0.263	-0.315	-0.252	-0.308	-0.292	-0.306
20-item Test						
True Traits						
mean	0.001	0.000	0.004	0.003	0.002	-0.001
sd	0.999	1.001	1.001	1.002	0.999	1.000
skew	-0.001	0.746	0.010	0.740	-0.002	0.757
kurtosis	-0.061	-0.056	0.011	-0.041	0.005	0.017
Est. Traits						
mean	0.008	-0.003	0.007	0.000	0.003	-0.002
sd	0.924	0.922	0.925	0.927	0.924	0.926
skew	0.018	0.567	0.023	0.560	0.014	0.570
kurtosis	-0.207	-0.243	-0.212	-0.262	-0.217	-0.234

FIGURE 1

Mean Absolute Bias and RMSE in Slope Parameters in the 2-Parameter Study (Averaged Across the Common Items)



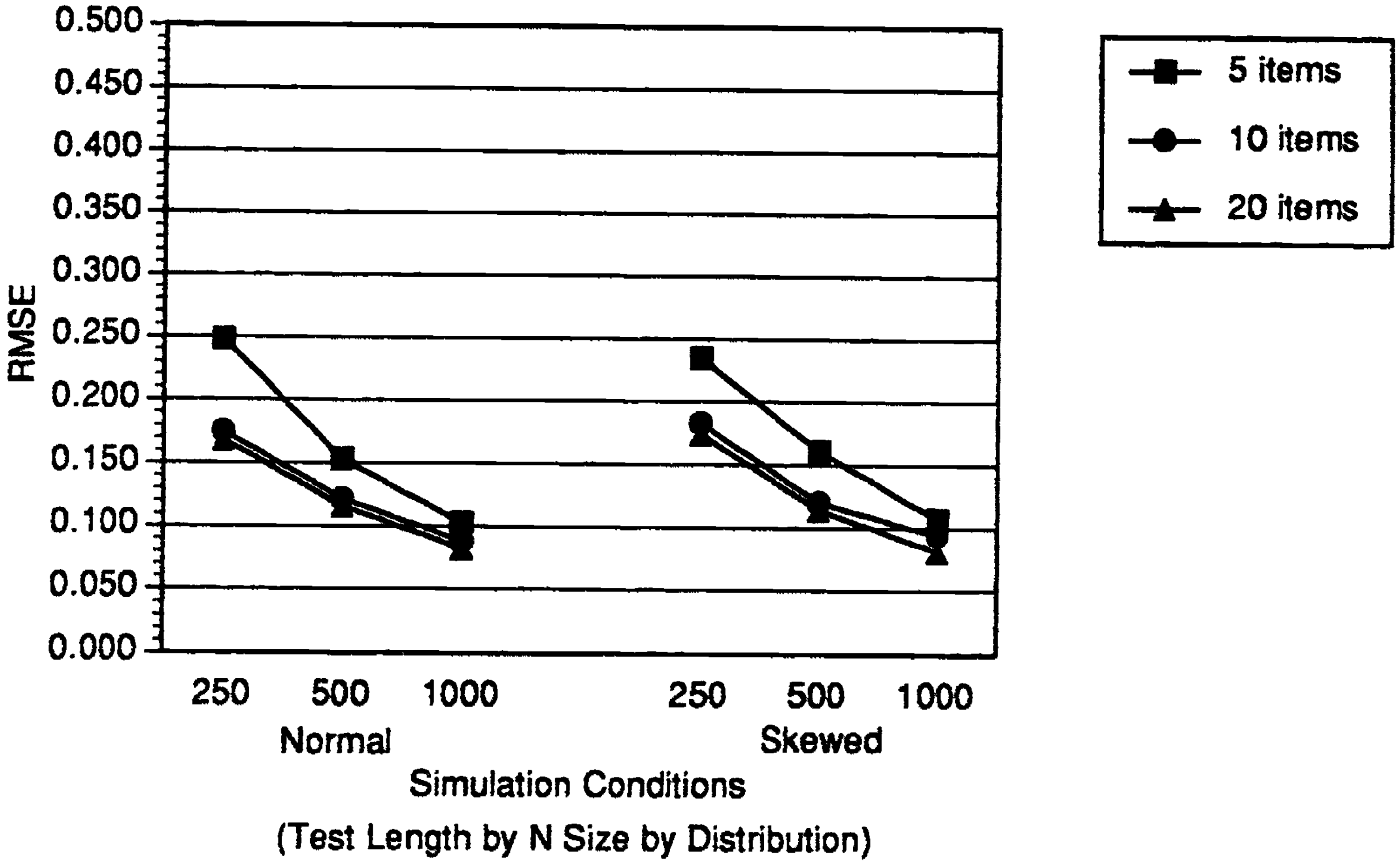


FIGURE 2

Mean Absolute Bias in Threshold Parameters for the 2-Parameter Study
(Averaged Across the Common Items)

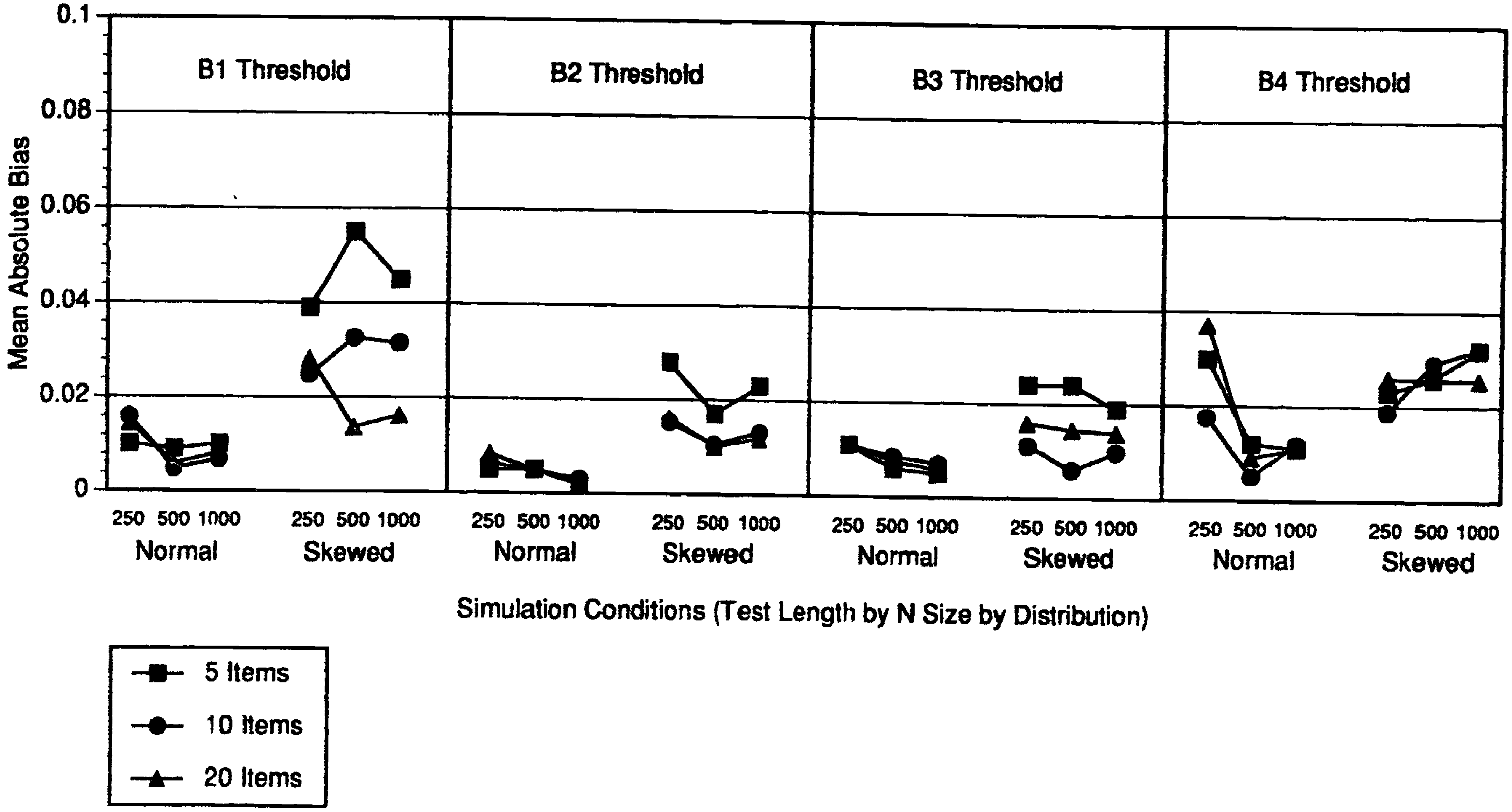


FIGURE 3

RMSE in Threshold Parameters in the 2-Parameter Study
(Averaged Across the Common Items)

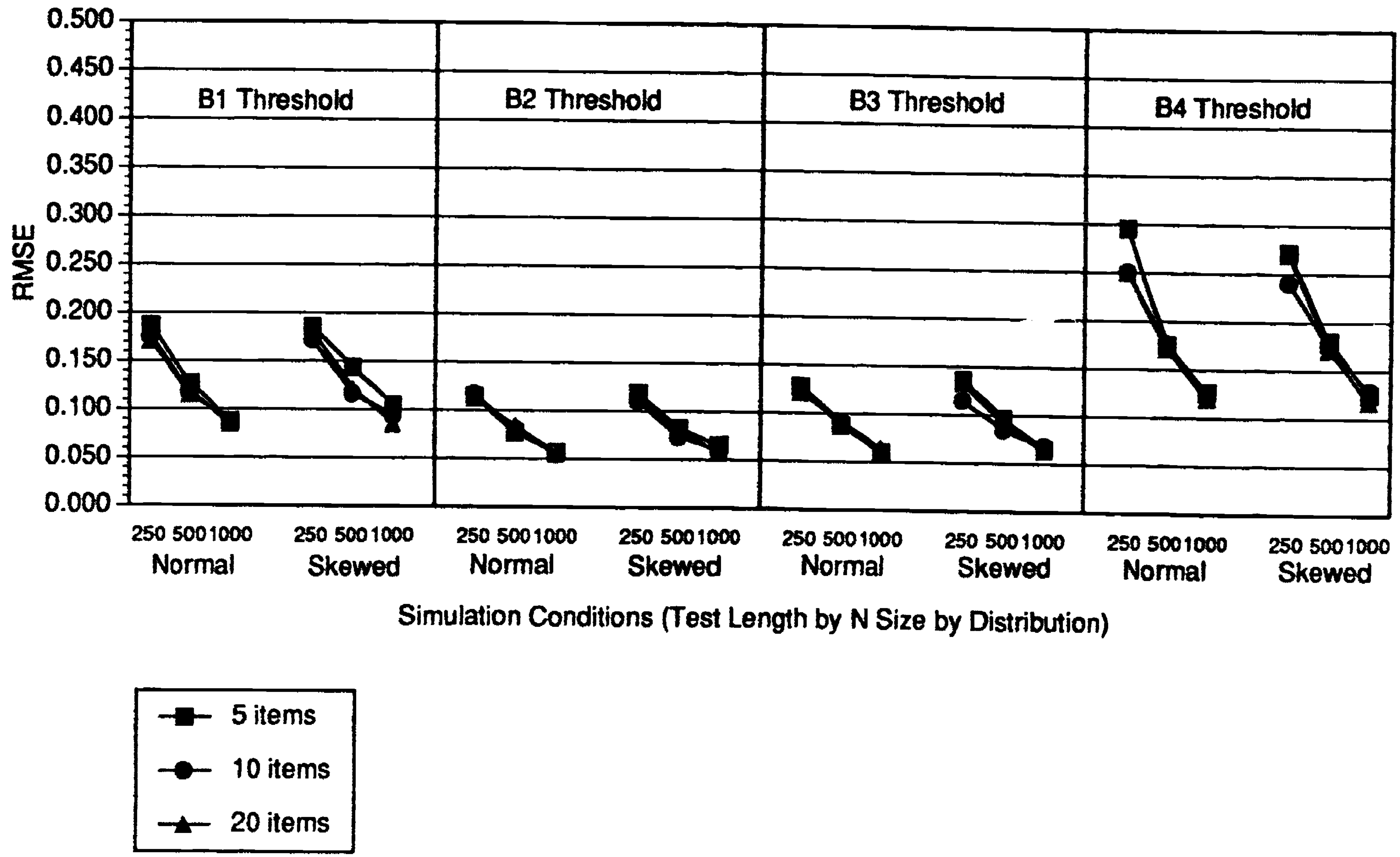


FIGURE 4

Bias in Ability Parameters in the 2-Parameter Study
(For Various Ranges of Ability)

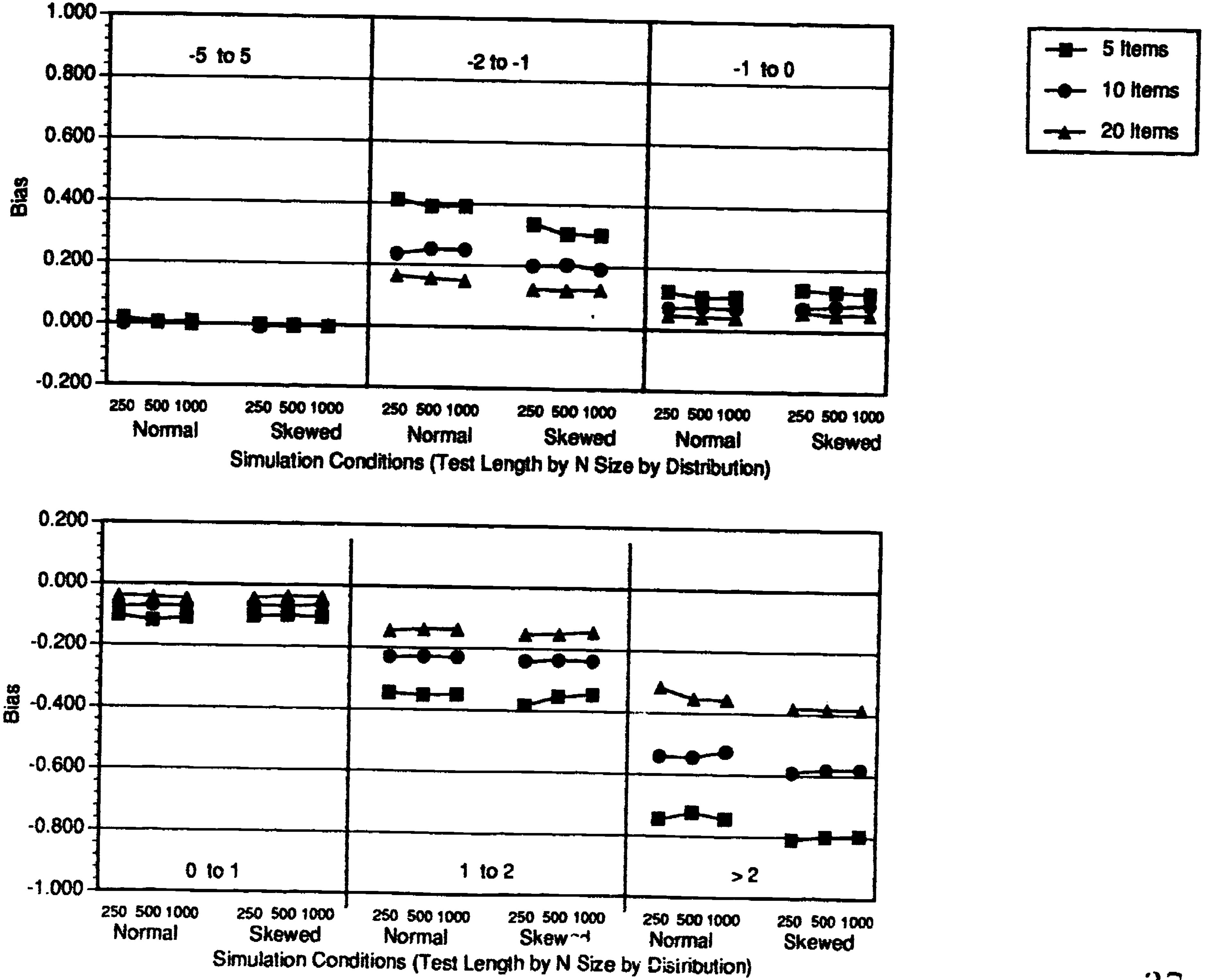
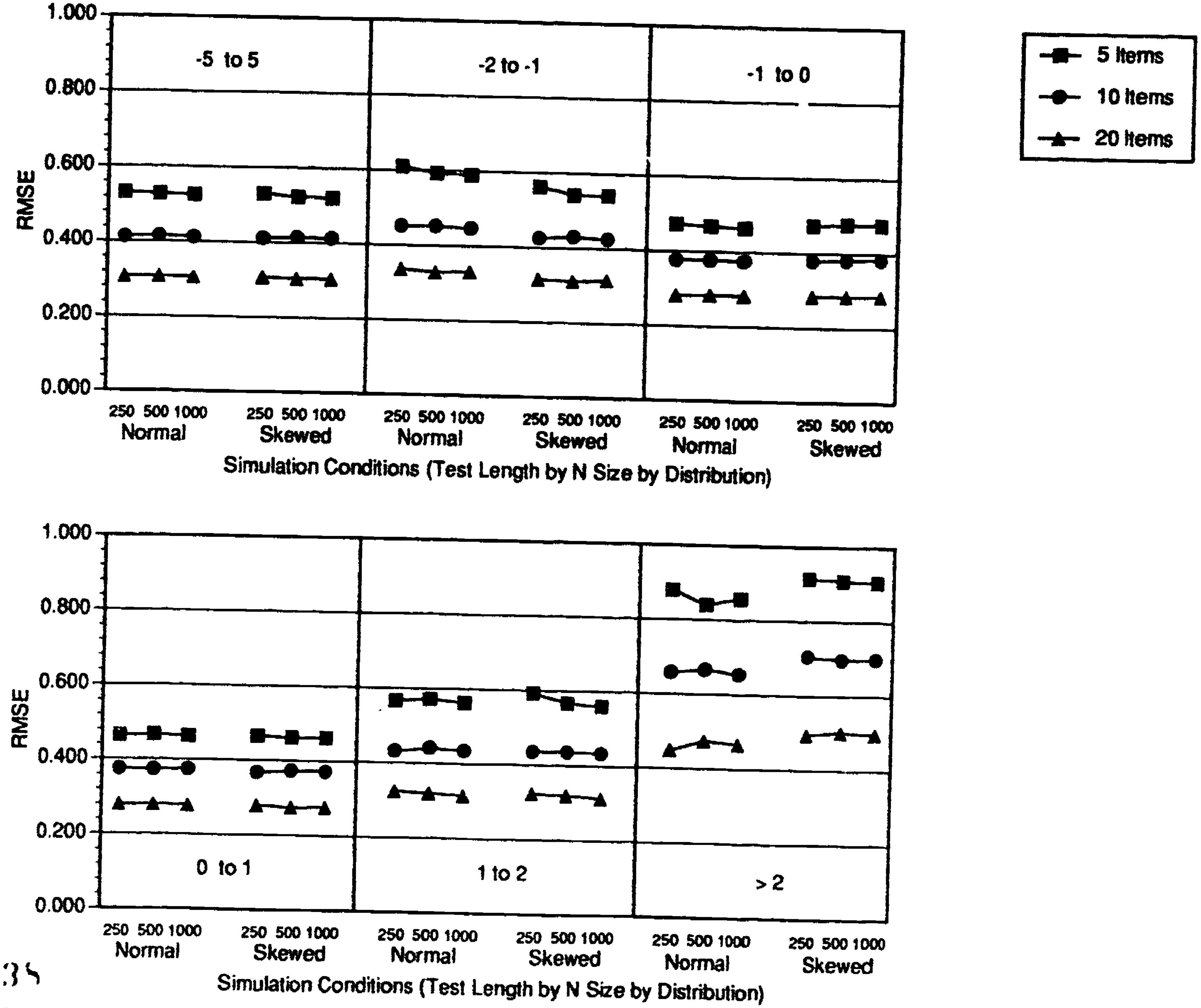


FIGURE 5

RMSE in Ability Parameters in the 2-Parameter Study
(For Various Ranges of Ability)


RMSE
-5 to 5
-2 to -1
-1 to 0
5 Items
10 Items
20 Items
Normal
Skewed
Simulation Conditions (Test Length by N Size by Distribution)
0 to 1
1 to 2
> 2