Recovery of Marginal Maximum Likelihood Estimates in the Two-Parameter Logistic Response Model: An Evaluation of MULTILOG

Clement A. Stone
University of Pittsburgh

Marginal maximum likelihood (MML) estimation of the logistic response model assumes a structure for the distribution of ability (θ). If this assumption is incorrect, the statistical properties of MML estimates may not hold. Monte carlo methods were used to evaluate MML estimation of item parameters and maximum likelihood (ML) estimates of θ in the two-parameter logistic model for varying test lengths, sample sizes, and assumed θ distribution. 100 datasets were generated for each of the combinations of factors, allowing for item-level analyses based on means across replications. MML estimates of item difficulty were generally precise and stable in small samples, short tests, and under varying distributional assumptions of θ. When the true distribution of θ was normal, MML estimates of item discrimination were also generally precise and stable. ML estimates of θ were generally precise and stable, although the distribution of θ estimates was platykurtic and truncated at the high and low ends of the score range. *Index terms: marginal maximum likelihood, monte carlo, MULTILOG, two-parameter logistic response model.*

Alternatives to joint maximum likelihood (JML) estimation of the two- and three-parameter logistic response model in LOGIST (Wingersky, 1983) have recently become available. BILOG (Mislevy & Bock, 1983) implements marginal maximum likelihood (MML) procedures (Bock & Aitkin, 1981) as well as a Bayesian marginal modal method described by Mislevy (1986). MULTILOG (Thissen, 1986) also implements MML procedures. These estimation procedures have been compared with regard to their efficiency and accuracy (see, e.g., Mislevy & Stocking, 1989; Qualls & Ansley, 1985; Yen, 1987). A general finding of this research is that MML and Bayesian estimation procedures are advantageous, particularly when the item set and/or examinee sample is small.

The MML procedures in BILOG and MULTILOG assume a structure for the distribution of θ, 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 apply even in small item sets (Mislevy & Stocking, 1989). Once MML estimates of the item parameters are obtained, ML estimates of θ can be obtained. However, if either the logistic response model or the assumed distribution of θ is incorrect, the statistical properties of MML estimates may not hold (Mislevy & Sheehan, 1989).

Drasgow (1989) recently evaluated MML estimates using a Fletcher-Powell (1963) algorithm and the two-parameter logistic response model. Item responses were simulated for tests comprised of 5, 10, 15, and 25 items in conjunction with groups of 200, 300, 500, and 1,000 examinees sampled from a N(0,1) distribution. The item parameters for the study were selected from responses to the Job Descriptive Index (Drasgow & Hulin, 1988), which represented a moderately easy test (i.e., item difficulties were centered below the mean of the θ distribution). Ten replications for each of the four sample sizes and for each of the four test lengths were examined. The measurement of average recovery of item parameters and item response functions

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 1, March 1992, pp. 1-16

0146-6216/92/010001-16$2.05

(IRFs) was possible by examining individual item parameter estimates across replications. This approach contrasts with most monte carlo research involving the logistic response model in which a single replication is used and estimation accuracy is examined across items. Drasgow concluded that, for less extreme item parameters ($.80 \leq a_j \leq 1.40$; $-1.50 \leq b_j \leq 1.50$), unbiased parameter estimates with "reasonably small" standard errors were achieved with 200 examinees and five items. For more extreme item parameters ($a_j < .80$, $a_j > 1.40$; $b_j < -1.50$, $b_j > -1.50$), however, estimates were biased and had "large" standard errors.

Seong (1990) evaluated the recovery of item and θ parameters in BILOG when the true θ distribution did not correspond to the prior specification. Five replications of item responses were generated assuming a two-parameter logistic response model, a test of 45 items, 100 or 1,000 examinees, and three true θ distributions (normal, positively, or negatively skewed). For $N = 1{,}000$, matching of the prior to the true θ distribution provided more accurate estimates of item and θ parameters. Although θ estimation also improved for $N = 100$ when the prior on the θ distribution matched the true θ distribution, the results for the item parameters were inconsistent. For example, given a dataset with a true negatively skewed θ distribution, item discrimination estimates were found to be more accurate when the prior was specified as normal as opposed to negatively skewed. Also, given a dataset with a true positively skewed θ distribution, item difficulty estimates were found to be most accurate when the prior was specified as negatively skewed.

The present study extended this research in several ways. Assessing the average recovery of item parameters and IRFs across 5 or 10 replications might yield unstable results; thus, the present study used 100 replications. In comparison with Drasgow's research, the present study analyzed the recovery of θ parameters and assessed the impact of the true distribution of θ. Although Seong's (1990) work addressed these as well, the present study used small samples that were less extreme, a platykurtic θ distribution, and shorter tests.

Method

Model and Model Parameters

Monte carlo methods were used to evaluate MML parameter estimates produced by the EM algorithm (Dempster, Laird, & Rubin, 1977) in MULTILOG for the two-parameter logistic response model. Item discriminations (a_j) and difficulties (b_j) were specified for J test items, and ability parameters (θ_n) were defined for N examinees. Item parameters are typically specified in one of two ways—either by using estimates from calibrating a particular test or by randomly sampling item parameters (e.g., uniformly sampling b_j on the interval −2.0 to 2.0). Rather than base the present study on a randomly generated set of item parameters, results from calibrating a 20-item math achievement test based on the two-parameter logistic model were used. These item parameters are given in Table 1. Note

Table 1
Item Parameters for the 20-Item Test

Item	a_j	b_j
1	.935	−.429
2	.941	−.019
3	1.960	.849
4	2.520	.666
5	1.670	1.360
6	.830	−2.180
7	.716	−2.040
8	2.360	−.338
9	1.810	−.427
10	2.370	−.352
11	.724	.040
12	1.900	.283
13	1.680	.604
14	3.000	.427
15	2.450	1.180
16	1.360	.536
17	1.130	2.430
18	1.370	1.820
19	2.350	1.480
20	2.210	1.730
Mean	1.714	.381
SD	.688	1.177

that in MULTILOG the scaling factor in the logistic model is 1.0, not 1.7. Thus, for example, a mean of 1.7 for the item discriminations corresponds to a mean of 1.0 in LOGIST.

Experimental Factors

The following three factors were manipulated: test length (J = 10, 20, and 40 items), sample size (N = 250, 500, and 1,000 examinees), and the true distribution of θ [N(0,1), positively skewed, and symmetric but platykurtic]. 20- and 40-item tests were considered because they represent test lengths frequently found in psychological and educational applications (Yen, 1987). For a more extreme case, a test comprised of 10 items was also examined. For the 20-item test, the item parameters in Table 1 were used. For the 10-item test, every even numbered item from the set of 20 was used ($\bar{a}_j$ = 1.89 SD_{a_j} = .73; $\bar{b}_j$ = .26 SD_{b_j} = 1.14). The 20-item test was duplicated in order to produce the 40-item test. The skewed and platykurtic distributions represent deviations from a normal distribution and were derived by using a power method described by Fleishman (1978). This method involves transforming 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 = .75 and kurtosis = 0.0), the following coefficients were used: a = -.1736, b = 1.1125, c = .1736, and d = -.0503. To produce the symmetric but platykurtic distribution (skewness = 0.0 and kurtosis = -1.0), the following coefficients were used: a = 0.0, b = 1.2210, c = 0.0, and d = -.0802.

Data Generation and Analysis

Using the defined item parameters, item response vectors were generated by randomly sampling θ from an assumed distribution, determining the probability of a correct response according to the two-parameter logistic response model given the item parameters, and comparing the probability with a random number sampled from a uniform [0,1] distribution. A simulated response was scored correct if the probability of a correct response was less than or equal to the sampled number. Simulated datasets were generated using GENIRV (Baker, 1982) and item parameter estimates were obtained using MULTILOG. To minimize computer time, model parameters were used as starting values in MULTILOG. ML estimates of θ were then obtained from MULTILOG, but with item parameters fixed at their estimated values. Note that using true values as starting values may be viewed as inappropriate because local maxima may be avoided. However, this factor was deemed inconsequential because one of the noted strengths of the EM algorithm is that the choice of starting values is not critical (Bock, 1991).

For each of the 27 combinations of experimental conditions (3 levels of test length, 3 levels of N, and 3 different true θ distributions), 100 different datasets were generated. For example, 100 different datasets were generated each with 250 item response vectors to a 10-item test under the assumption that abilities were distributed N(0,1). Thus, 2,700 datasets were analyzed using MULTILOG. The use of multiple replications allowed for analyses based on means across replications as opposed to analyses based on a single dataset, because results for a single dataset can be particularly misleading when N is small and/or the test length is short. Note that a different "seed" (starting value for the random number generator) was used to generate item responses for the 27 combinations of the three experimental conditions. Although the results may be less comparable across conditions, they are less dependent on specific seed values and the sampling results are independent of each other.

Before the results from MULTILOG could be compared against "true" values, it was important that a common metric underlie both the estimates and true values. In item response theory models, the measurement scale for item and θ parameter estimates is arbitrary. However, once the θ scale is determined, the scale for the items also is determined. To obtain the item and θ parameters using MMLE procedures in MULTI-

LOG, the likelihood function is integrated over a specified prior distribution for θ (typically unit normal). This isolates the estimation of item parameters from estimation of θ parameters. However, the integration is accomplished by approximating the normal density function with quadrature points and weights (Mislevy & Stocking, 1989). Thus, the form of the estimated θ distribution is defined by this quadrature distribution rather than the specified prior distribution, although in practice the two will differ only slightly. Consequently, the metric of the θ and item parameter estimates in MULTILOG is determined by the mean ($\bar{X}$) and SD (σ_x) of the final adjusted quadrature distribution (Baker, 1990).

In order for the MULTILOG item parameter estimates to be on the same scale as their true values, the following transformation equations were used:

$$a_j = a_k \, \sigma_{x_q} / \sigma_{\theta j} \tag{1}$$

and

$$b_j = \frac{\bar{a}_k}{\bar{a}_j} b_k + \left(\bar{\theta}_j - \frac{\bar{a}_k}{\bar{a}_j} \bar{x}_q \right) \quad , \tag{2}$$

where j, k, and q reference the underlying true metric, the MULTILOG estimates, and the final adjusted quadrature distribution, respectively. Similarly, metric information for the item parameter estimates is imparted to the θ estimates and the following transformation was applied to the MULTILOG θ estimates:

$$\theta_j = \frac{\bar{a}_k}{\bar{a}_j} \theta_k + \left(\bar{b}_j - \frac{\bar{a}_k}{\bar{a}_j} \bar{b}_q \right) \quad , \tag{3}$$

where j and k are defined as above. Note that Equations 1–3 are based on equations described by Loyd and Hoover (1980) for equating. For more detail on the above transformation equations, see Baker (1990).

Recovery of item parameter values was assessed by averaging information across the 100 replications. Bias in each a_j was assessed by examining the difference between the mean of $\hat{a}_j$ and a_j across 100 replications:

$$\text{Bias in } a_j = \sum_{k=1}^{100} (\hat{a}_{jk} - a_j)/100 \quad , \tag{4}$$

where k references the replication and j references the item. Bias in b_j was analogously assessed:

$$\text{Bias in } b_j = \sum_{k=1}^{100} (\hat{b}_{jk} - b_j)/100 \quad . \tag{5}$$

Recovery was also assessed by examining the root mean squared error (RMSE) for each a_j (or σ_{a_j}) and each b_j (or σ_{b_j}) across the 100 replications:

$$\text{RMSE } a_j = \left[\sum_{k=1}^{100} (\hat{a}_{jk} - a_j)^2/100 \right]^{1/2} \tag{6}$$

and

$$\text{RMSE } b_j = \left[\sum_{k=1}^{100} (\hat{b}_{jk} - b_j)^2/100 \right]^{1/2} \quad . \tag{7}$$

By examining both bias and RMSE, it was possible to consider both the accuracy and variability of the point estimates. Mean absolute bias—the absolute value of the bias for each item averaged across test items—was computed for the item parameters across J items for each of the experimental conditions. The absolute value was used when summarizing across the test items so that positive and negative bias values did not cancel each other and thus misrepresent the true difference between the estimates and true values. However, the sign was retained for individual item bias results (Equation 4) to examine positive or negative bias at the item level.

Results

Ancillary results from the MULTILOG analyses are given in Table 2; these include the average number of iterations (cycles), the average posterior mean and SD of the quadrature distribution at the final iteration, and the number of times item discriminations (a_j) exceeded 4.5 ($a_{4.5}$). All information in the tables and figures represent data summarized across 100 replications. Table 2 shows that fewer iterations were

Table 2
Average Number of Cycles (AC), Posterior Mean and SD, and Number of $a_j > 4.5$ ($a_{4.5}$) for $N = 250$, 500, and 1,000

Test Length and Distribution	$N = 250$				$N = 500$				$N = 1{,}000$			
		Posterior				Posterior				Posterior		
	AC	Mean	SD	$a_{4.5}$	AC	Mean	SD	$a_{4.5}$	AC	Mean	SD	$a_{4.5}$
10-Item Test												
N(0,1)	19	.003	1.010	18	13	.001	1.013	3	10	-.001	1.012	0
Skewed +	21	-.002	1.015	26	18	-.007	1.018	11	12	-.007	1.018	0
Platykurtic	20	.005	1.017	25	14	-.007	1.014	3	11	-.006	1.016	1
20-Item Test												
N(0,1)	22	-.006	1.045	12	17	-.002	1.042	1	13	-.005	1.043	0
Skewed +	22	-.001	1.043	14	16	.002	1.049	4	14	-.001	1.051	0
Platykurtic	21	.014	1.049	9	15	.005	1.050	0	13	.011	1.052	0
40-Item Test												
N(0,1)	29	-.006	1.137	6	24	.006	1.140	0	23	-.005	1.140	0
Skewed +	30	.012	1.131	8	28	.016	1.142	0	24	.011	1.146	0
Platykurtic	25	-.028	1.134	12	24	-.025	1.137	3	18	-.034	1.140	0

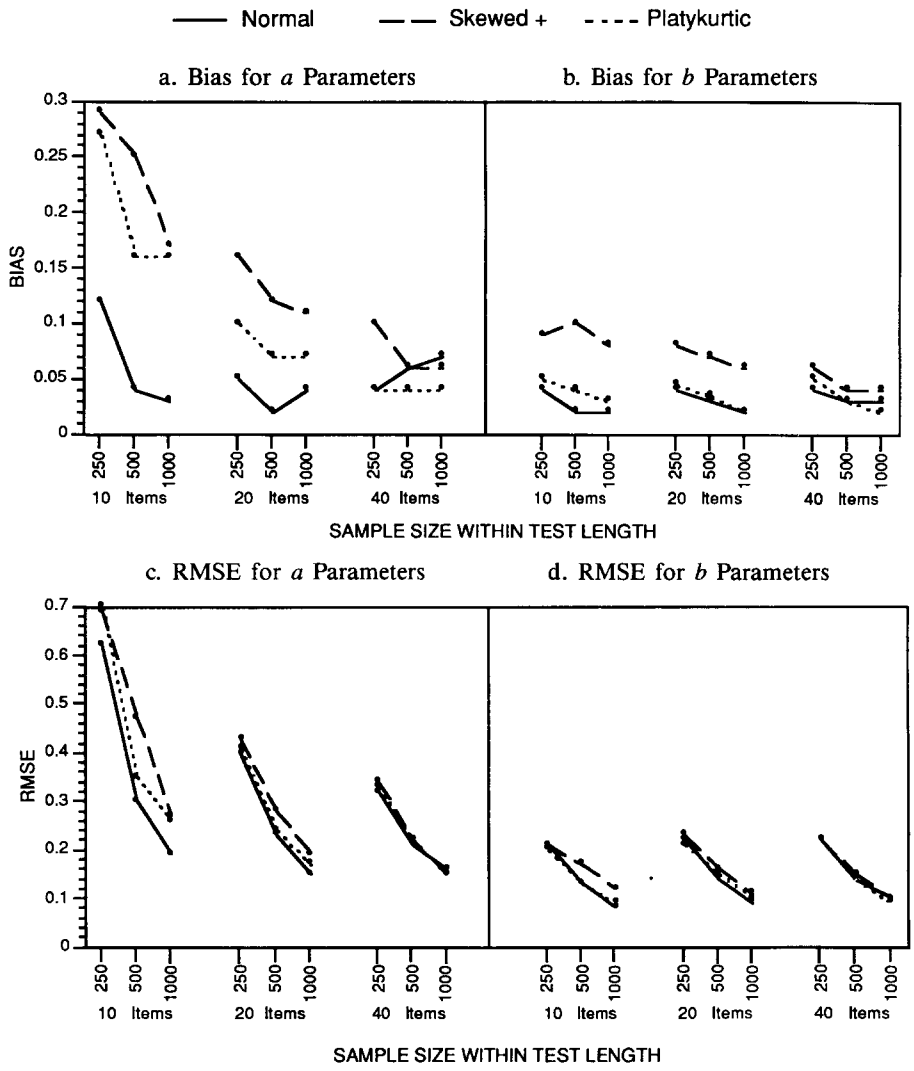
necessary as the number of examinees increased and as the number of items decreased. When the true distribution for θ was non-normal, small changes were observed in the number of iterations. In nearly every case, however, more iterations were required when the distribution was skewed positively. Given that item parameter values were used as starting values, the average number of iterations was surprisingly high. Although the posterior means generally approximated 0, the posterior SDs deviated further from 1.0 as the number of items increased. More extreme item discrimination estimates were found for both small sample sizes and short tests. Also, in general, more extreme parameter estimates were observed in the cases in which the θ distribution was skewed or platykurtic.

Recovery of Item Parameters

Test level results. Figure 1 provides a summary of the absolute value of bias and RMSE for the item parameters across the J items for each of the 27 experimental conditions. There appears to be an effect due to test length for the a parameters (Figure 1a). For the skewed and platykurtic cases, there was less bias in a as the number of items increased from 10 to 40 items. In addition, RMSE was reduced with increased test length (Figure 1c), although most of the impact was observed when the test increased from 10 to 20 items. There was also an effect due to sample size (N). However, the impact of N on bias was limited to the skewed and platykurtic cases and was primarily observed for tests comprised of 10 or 20 items in conjunction with N increasing from 250 to 500. Within each of the tests (10, 20, or 40 items), it was not surprising to find that RMSE was reduced as N increased. Finally, the shape of the true θ distribution affected the bias, but had relatively little effect on RMSE results. Except for the 40-item tests, bias was greater when the distributions were skewed or platykurtic as opposed to normal, and the results for the platykurtic case were better behaved than the skewed positive case, in general. As indicated above, this effect was mitigated by increasing N or the number of items.

Bias for the b parameters (Figure 1b) was negligible (very close to 0.00) regardless of N and the number of items. However, there appeared to be a marginal effect due to the shape of the true θ distribution, but only when comparing the skewed versus normal and platykurtic cases. In addition, bias decreased marginally as N increased. The only factor influencing the RMSE for b_j was N (Figure 1d). It was not surprising to

Figure 1
Mean Absolute Bias and Mean RMSE Across All Items



find that increasing N reduced RMSE values, because standard errors that describe the sampling distribution for estimates decrease as N increases.

Item level results. Figures 2 and 3 present bias and RMSE results at the item level for a low discriminating item (a_j = .83), an item with approximately average discriminating power (a_j = 1.9), a highly discriminating item (a_j = 3.0), an item of average difficulty (b_j = -.02), an easy item (b_j = -2.18), and a moderately difficult item (b_j = 1.82).

From Figure 2, it can be seen that the effects of the factors on bias depend on the particular item parameter. For the low discriminating item (Figure 2a), bias was negligible irrespective of the number of test items, the true distribution for θ, and N. For the average discriminating item (Figure 2b), bias was greater when the distribution of θ deviated from N(0,1) for tests comprised

Figure 2
Absolute Value of Bias for Specific Item Parameters

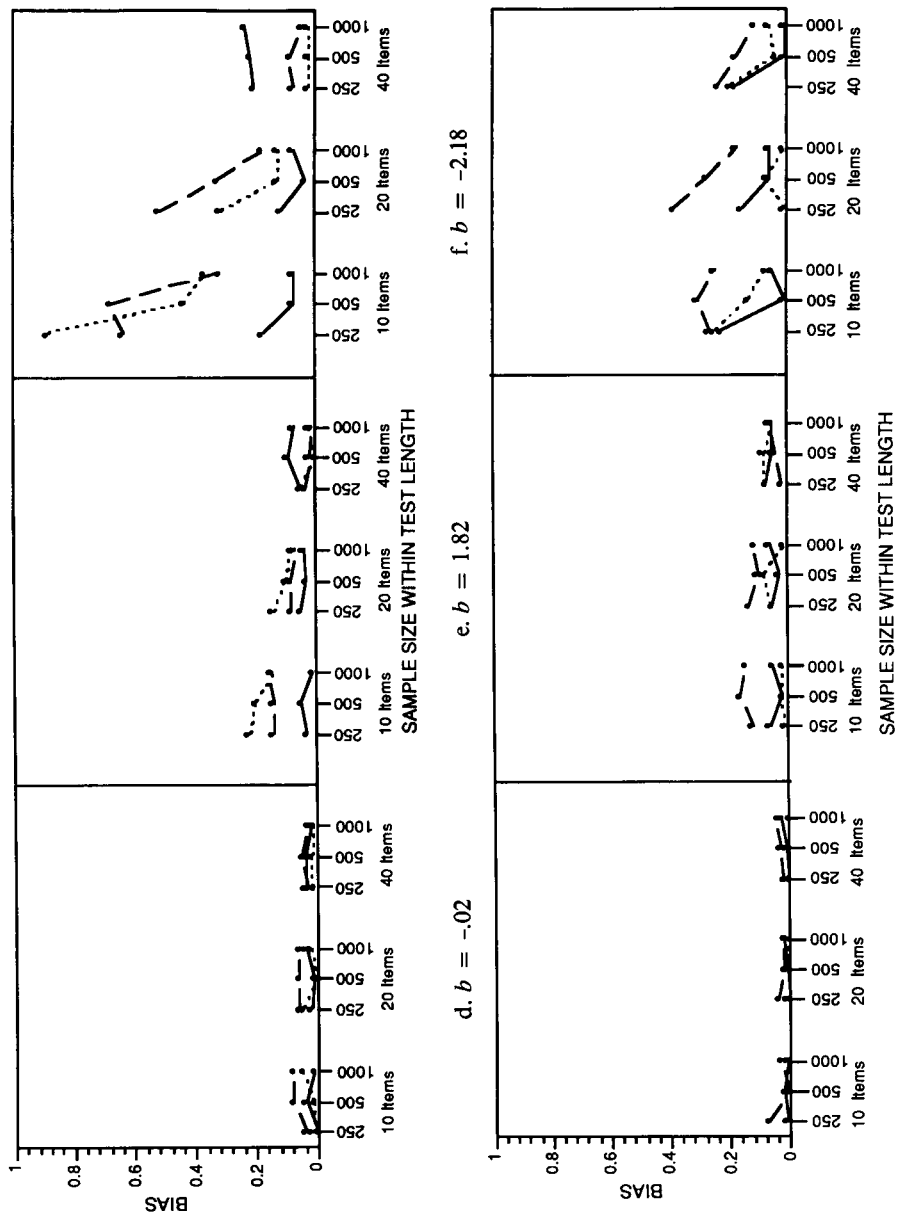
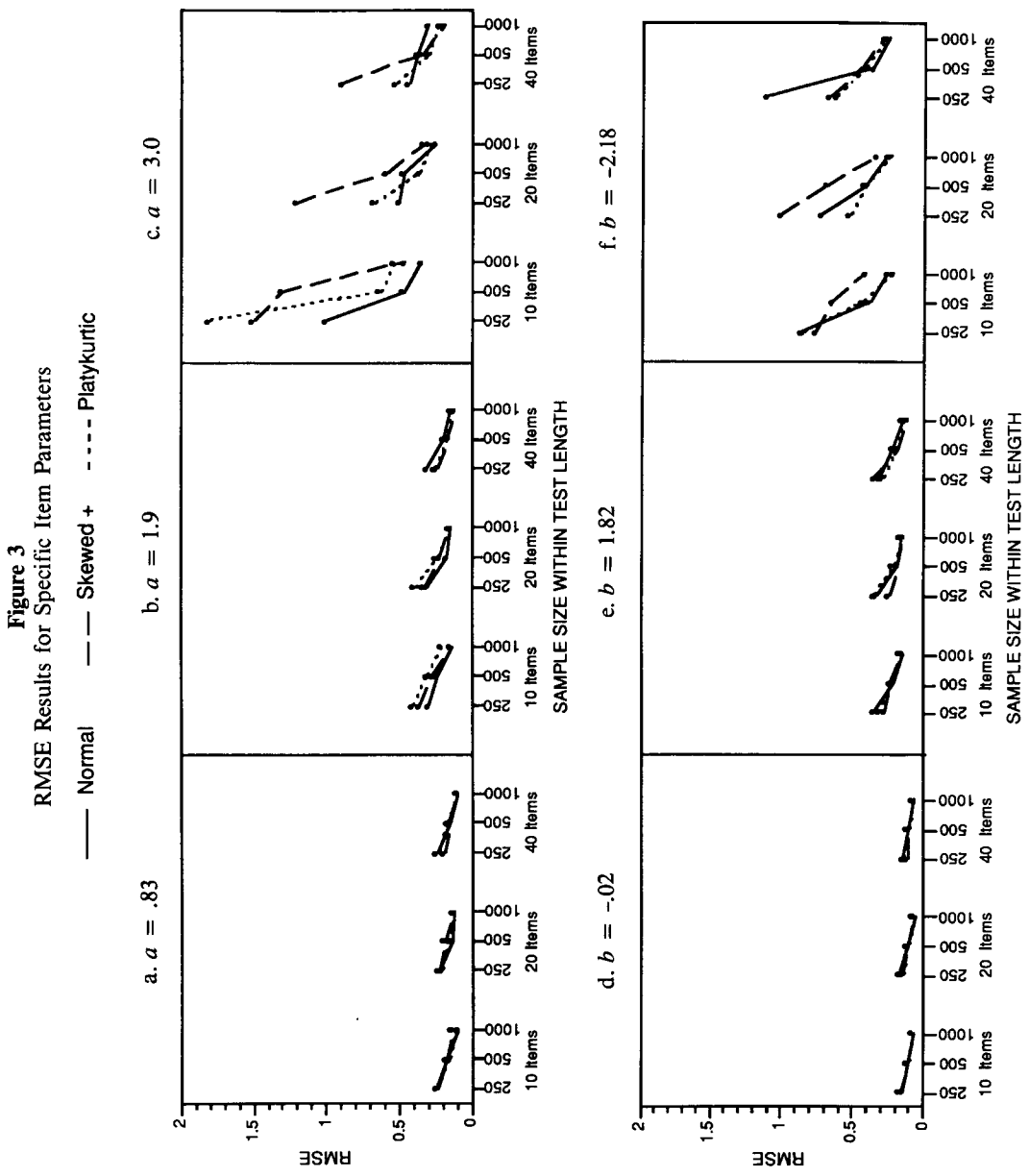


Figure 3
RMSE Results for Specific Item Parameters



of 10 or 20 items, irrespective of N. More pronounced effects were also observed for the 10-item tests. This same pattern of results also can be seen for the highly discriminating item (Figure 2c), although the effects are more extreme because the parameter was more extreme. One anomaly was observed for the highly discriminating item when the test length was 40 and the θ distribution was N(0,1). The bias in these cases was considerably larger than cases in which the tests had fewer items or the θ distribution deviated from N(0,1), irrespective of N.

For the item of average difficulty (Figure 2d), negligible bias was observed when the distribution, N, and number of test items were varied. For the easy and moderately difficult items (Figures 2f and 2e, respectively), the skewed condition generally demonstrated greater bias than the N(0,1) and platykurtic distributions regardless of N for tests of 10 and 20 items. For $N = 250$, bias was generally higher for the easy item (Figure 2f) than for the other items (Figures 2d and 2e), regardless of the distributional assumption or the test length.

Figure 3 presents the RMSE results for the same items. As was observed in previous results, the RMSE decreased as N increased. Smaller RMSE values were observed for the average difficulty item (Figure 3d; $b = -.02$) and the item with low discrimination (Figure 3a). Greater RMSE values were observed for the highly discriminating item (Figure 3c) and the item that was extremely easy (Figure 3f). Unlike the bias results, when the distribution of θ was skewed RMSE was not systematically larger than for the other distributional conditions. Some of the larger RMSE results observed for the shorter tests and smaller samples sizes are probably due to a few cases in which very extreme estimates were produced in MULTILOG. Finally, the fact that bias and RMSE were greater for the easy item as opposed to the more difficult item may be due to the fact that the difficulty level of the easy item ($b_j = -2.18$) was more extreme than for the difficult item ($b_j = 1.82$). This may have been compounded by the fact that the test was centered above 0 (see Table 1).

Some interesting results regarding the direction of bias in item discrimination parameters were observed. The proportion of negative bias values (i.e., proportion of times $\bar{a}_j - a_j < 0$ across J items) are given in Table 3. Note that systematic positive or negative bias would be indicated by a disproportionate number of positive or negative values. Four trends can be observed from the table: (1) The proportion of negative values was generally low, indicating positive bias; (2) as N increased, the proportion of negative values increased; (3) for the N(0,1) case, as the number of test items increased the proportion of negative values increased; and, (4) the proportion of negative values in the skewed and platykurtic cases was generally less than the values observed for the N(0,1) case.

Table 3
Proportion of Negative Bias Values for a_j

Test Length and Distribution	$N = 250$	$N = 500$	$N = 1{,}000$
10-Item Test			
N(0,1)	.1	.2	.4
Skewed +	.3	.2	.3
Platykurtic	.1	.2	.2
20-Item Test			
N(0,1)	.25	.6	.85
Skewed +	.1	.2	.3
Platykurtic	.2	.4	.35
40-Item Test			
N(0,1)	.53	.88	.93
Skewed +	.28	.35	.43
Platykurtic	.28	.35	.55

As Lord (1983) indicated, it is not surprising to find positive bias in the discrimination parameter estimates. However, these results indicate that the bias can be reduced by increasing N and increasing test length. It should also be noted that the bias appears to be negative for large N and longer tests when the true distribution for θ is N(0,1). Thus, $\bar{a}_j$ appears to consistently underestimate a_j for the N(0,1) case, but consistently overestimate a_j for the two non-normal cases. The reason for this divergence is unclear. It

could be due partly to the absence of extreme estimates that tend to positively skew the sampling distribution for a_j, but this does not account for the observed negative bias.

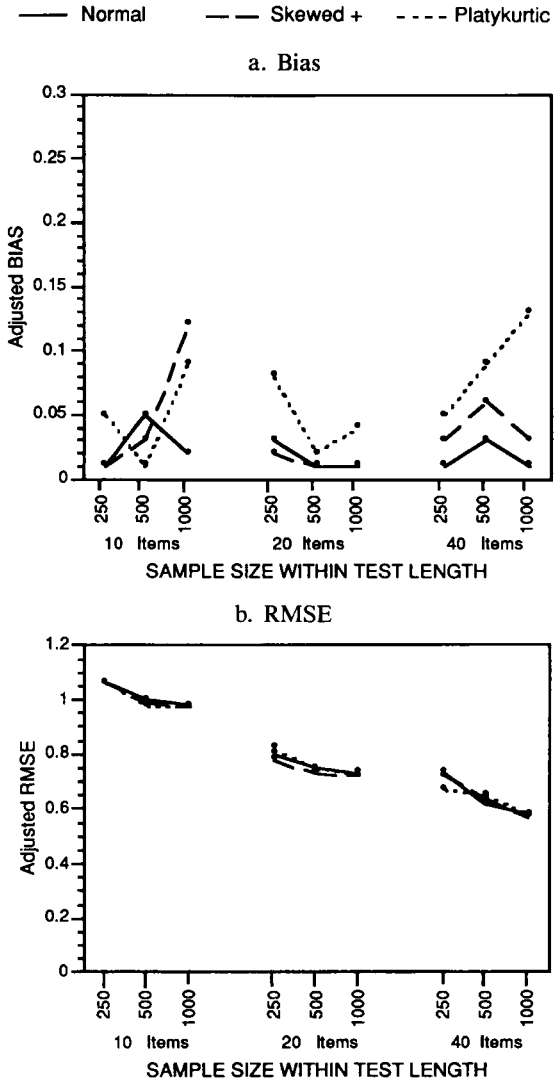
Although not reported in Table 3, the direction of the bias in b_j corresponded to the direction of the true value; that is, if the true value for b_j was negative, the bias was negative, and if the true value was positive, the bias was also positive. This finding applied particularly to the more extreme values ($b_j < -1.5$ and $b_j > 1.5$), irrespective of the true θ distribution, N, and test length. This bias causes scale expansion and is consistent with results described by Yen (1987).

Recovery of Test Response Functions

It has been argued (Hulin, Lissak, & Drasgow, 1982) that examining the item parameters separately may not be as important as examining the joint difference between the parameters for an item and its estimates. For example, it is possible to examine the area between the IRF given the item parameters and the IRF given the item parameter estimates. This is equivalent to examining the difference between the probability of correctly answering an item given true parameter values and the probability based on estimated item parameters and θs. Alternatively, the difference between true scores and estimated true scores based on the sum of these probabilities can also be examined (Yen, 1987). This latter comparison examines the extent to which the true test response function (TRF) is reproduced by parameter estimates.

Figure 4 presents the absolute value of the bias and RMSE results when comparing the true TRF with the estimated TRF. Note that in order to compare the results across the various test length conditions, it was necessary to simulate equivalent test lengths. Therefore, the bias and RMSE values for the 20-item tests were adjusted to a test length of 10 by dividing the values by 2, and bias and RMSE values for the 40-item tests were divided by 4. Figure 4a indicates that there is not a clear relationship between any of the factors and the bias in estimating TRFs, although larger bias values are associated somewhat with skewed and

Figure 4
Absolute Adjusted Bias and RMSE Results for Estimated True Scores



platykurtic distributions. However, the adjusted RMSE values (Figure 4b) decreased both as N increased and test length increased. There appears to be no systematic change in RMSE values when the distribution for θ is varied.

Recovery of θ Parameters

Bias and RMSE in recovering θ parameters were calculated using Equations 4 and 6 after

Table 4
Distributional Statistics for Estimated and True θs for N(0,1) Skewed (SK), and Platykurtic (PLK) Distributions

Test Length, θ Type, and Statistic	$N = 250$			$N = 500$			$N = 1,000$		
	N(0,1)	SK	PLK	N(0,1)	SK	PLK	N(0,1)	SK	PLK
10-Item Test									
True θ									
Mean	.01	−.01	−.01	.02	−.01	−.01	−0.00	.02	.01
SD	1.00	.99	1.00	.99	.99	1.00	1.01	1.01	.99
Skewness	.01	.78	.01	−.01	.76	.03	−0.00	.74	−.01
Kurtosis	.01	.08	−.98	−.05	.01	−1.01	−.06	.01	−1.00
Est. θ									
Mean	.03	.04	.05	.02	.04	.03	.03	.01	0.00
SD	.83	.83	.82	.82	.82	.82	.83	.83	.81
Skewness	.21	.34	.18	.20	.38	.20	.22	.36	1.8
Kurtosis	−.69	−.64	−.86	−.74	−.68	−.88	−.75	−.70	−.88
20-Item Test									
True θ									
Mean	0.00	−0.00	−0.00	−0.00	.01	.01	−.01	−.01	−.01
SD	1.01	1.00	1.00	.99	1.01	1.00	1.00	1.00	1.00
Skewness	.01	.77	−.01	−.01	.75	−.01	−0.00	.76	.01
Kurtosis	.01	.04	−1.01	.01	−.01	−1.00	−0.00	.01	−.99
Est. θ									
Mean	.01	.03	.06	.02	.03	.03	.03	.04	.06
SD	.90	.89	.89	.88	.89	.89	.88	.88	.88
Skewness	.23	.44	.14	.22	.46	.15	.23	.47	.16
Kurtosis	−.49	−.44	−.81	−.50	−.45	−.84	−.52	−.46	−.84
40-Item Test									
True θ									
Mean	0.00	.01	−.01	−0.00	−0.00	0.00	−0.00	0.00	0.00
SD	1.01	1.00	1.01	1.00	1.00	1.00	1.00	1.00	1.00
Skewness	−0.00	.75	0.00	−.01	.75	−0.00	−.01	.75	−0.00
Kurtosis	−.06	−.02	−1.00	.01	−.01	−1.00	−.02	0.00	−1.00
Est. θ									
Mean	.01	.03	−.02	.03	.05	−.02	.01	.04	−.04
SD	.95	.94	.95	.93	.94	.94	.93	.93	.92
Skewness	.18	.47	.11	.19	.50	.11	.20	.52	.11
Kurtosis	−.43	−.36	−.83	−.45	−.37	−.85	−.45	−.38	−.88

substituting θ for a_j. In addition, the correspondence between the true θ and estimated θ distributions were examined. Results are presented in Table 4.

Several interesting trends are noteworthy. For all cases, the distribution of θ estimates was always platykurtic and demonstrated a smaller SD than the true distribution. For all but the platykurtic conditions, these deviations from the true distribution diminished as the number of test items increased but not as N increased. In addition, when considering the unskewed true distributions [N(0,1) and platykurtic], the estimated θ distributions were positively skewed. As above, this deviation diminished with increased test length but not with increased N.

The fact that the SD was smaller in the estimated than the true distribution is probably due to the fact that the range in true θs was from −4 to 4, but considerably narrower in the estimated distribution. When simulating an item vector given $\theta < -2.5$ or $\theta > 2.5$, the likely

result is a vector of 0s and 1s, respectively. However, in MULTILOG, all 0 or 1 vectors are given the same θ estimate in spite of the fact that the true θs may range from −2.5 to −4 or from 2.5 to 4. Thus, the estimated θ distribution is truncated, which in turn reduces its SD.

For the true skewed and platykurtic θ distributions, the estimated distributions failed to exhibit the same degree of skewness or kurtosis, although the kurtosis was captured to a greater extent than the skewness. In addition, the correspondence between the estimated and true skewed distributions improved with increased test length and to some degree with increased *N*. These observations should not be surprising. The fact that the same degree of skewness was not demonstrated in the estimated distribution as in the true distribution is due to the fact that MULTILOG assumes a N(0,1) prior on the θ distribution. On the other hand, the platykurtic distribution was symmetric and more compatible with a normal prior.

The bias and RMSE results for estimated θs are given in Figures 5 and 6. Results are provided for the range −4 to 4, the entire true θ range, and in the following ranges: −2 to −1, −1 to 0, 0 to 1, 1 to 2, and > 2. These θ ranges allow for the examination of θ parameter recovery at varying levels of θ. It is not surprising to see that recovery is better at the middle ranges of θ ($-1 < \theta < 1$; Figures 5c and 5d, and 6c and 6d) than for more extreme θ values ($\theta < -1$ and $\theta > 1$). This is evidenced by smaller bias and RMSE values. In addition, better precision in the point estimates for θ are observed as the number of test items increases, but primarily for more extreme levels in θ ($\theta < -1$ and $\theta > 1$). As the test length increases, bias and RMSE decrease irrespective of *N* and the true distribution for θ. If the entire θ range (−4 to 4) is examined, RMSE decreases slightly as both the number of test items and *N* increase, but no impact on bias is observed. Finally, no impact on bias and RMSE is observed when simply considering changes in the true θ distribution. As Seong (1990) noted, *N* is not a factor in the estimation of θ because θ is estimated for each examinee separately with no consideration of the sample size.

An interesting trend in Figures 5 and 6 is that the effects on bias and RMSE for the more extreme levels of θ ($\theta < -1$ or $\theta > 1$) were not symmetric. For example, bias and RMSE in the θ range −2 to −1 was always greater than the bias and RMSE in the θ range 1 to 2. This asymmetry is probably due to the fact that the difficulty level of the test was centered slightly above a θ level of 0. For the 20-item test, mean difficulty was .38 (see Table 1).

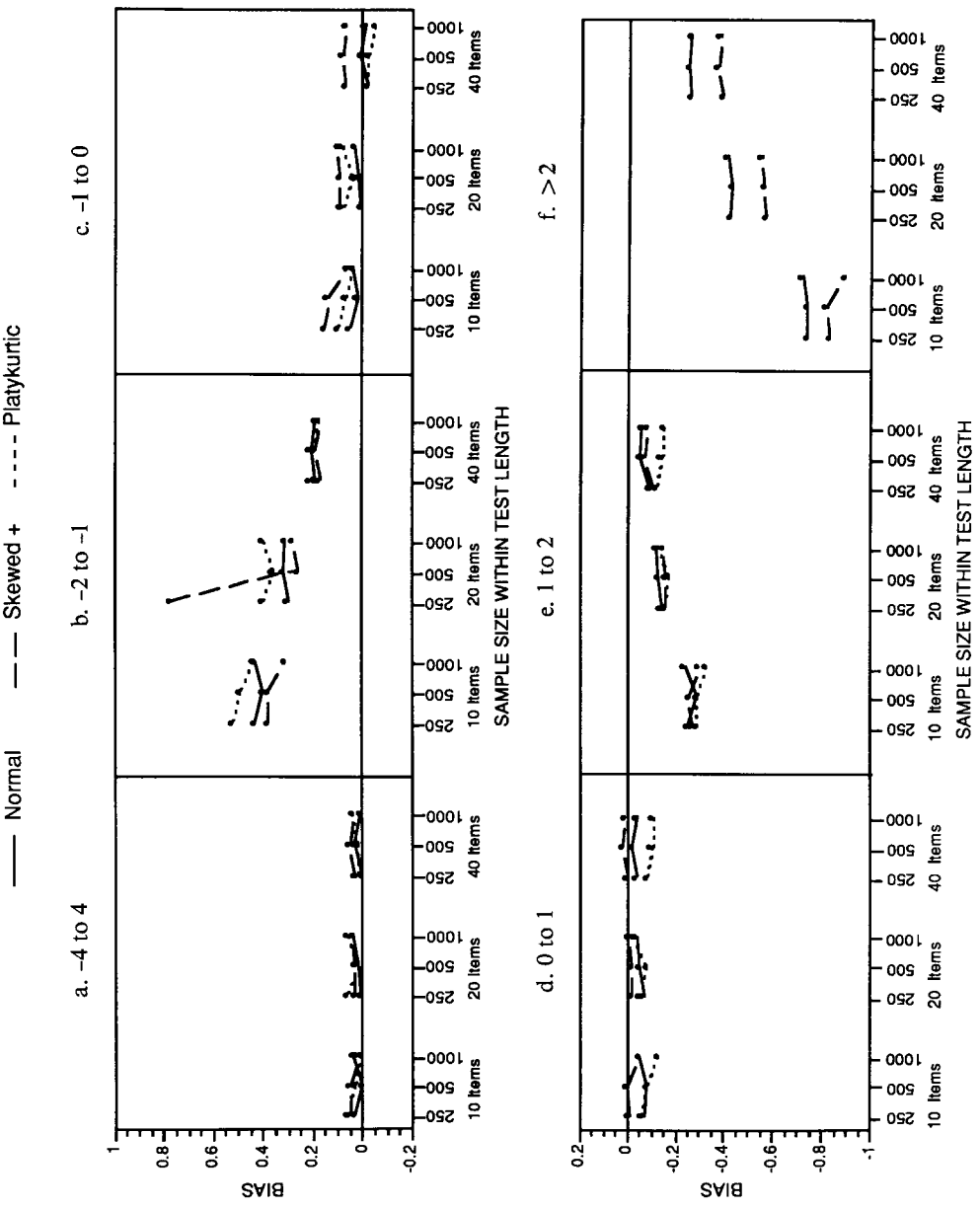
Another interesting finding that can be seen in the figures is that true θ was consistently underestimated for 10- and 20-item tests. This can be seen by examining the bias values across the various θ ranges. When $\theta > 0$, bias was negative; when $\theta < 0$, bias was positive. The point estimate of θ, given by the mean estimate across the 100 replications, was always less than the true θ when $\theta > 0$. When $\theta < 0$, the point estimate was greater than the true value, but because the values were negative, the true value was underestimated in an absolute sense. This effect was also generally observed in the 40-item tests, but more so for the extreme θ ranges.

Discussion

The results of this study offer several broad conclusions: (1) MML estimates of item difficulty were generally precise and stable in small samples ($N = 250$), short tests (10 items), and under varying distributional assumptions for θ (normal, skewed, and platykurtic); (2) For all the conditions, except $N = 250$ and test length = 10, MML estimates of item discriminations were generally precise and stable when the true distribution of θ was normal; and (3) ML estimates of θ were precise and stable, although extreme θ levels were consistently underestimated (i.e., negative bias for large positive θ values and positive bias for large negative θ levels).

Although estimates of item difficulty were precise, more bias was found for 10- and 20-item tests when the distribution was skewed, regardless of *N*. As found by Seong (1990), when the test was long (40 or more items), the effect of the true

Figure 5
Bias Results at Various Ranges of θ



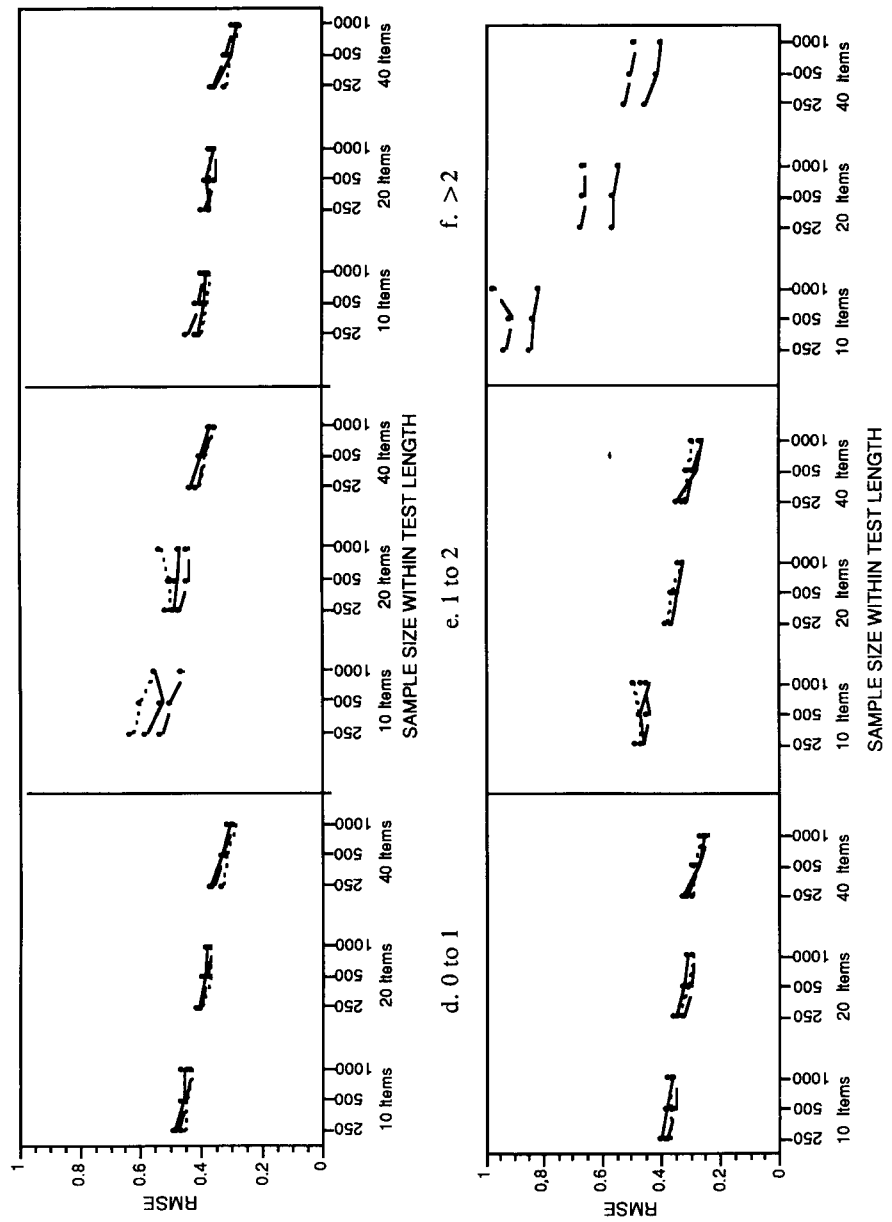


Figure 6
RMSE Results at Various Ranges of Ability
—— Normal — — Skewed + - - - - Platykurtic

θ distribution was negligible. In addition, the degree of bias was reduced as N increased. Finally, estimates of more extreme item difficulty parameters ($b_j > 1.5$ and $b_j < -1.5$) were generally exaggerated, resulting in an expanded scale for the item difficulties. As found by other researchers (e.g., Yen, 1987), large positive b_j were positively biased and large negative b_j were negatively biased.

True skewed and platykurtic θ distributions adversely affected the estimates of item discrimination estimates, with more pronounced effects being found with the skewed distributions. Incorporating more items into the test and/or increasing the sample size did, however, diminish the impact—both errors in estimating item discriminations and the variability surrounding the estimates decreased. Based on Seong's (1990) work, increasing the number of quadrature points to 20 should also help to minimize the effect of underlying non-normal θ distributions.

Errors in estimating θ were small, despite the rather large errors obtained in estimating item discriminations for small sample sizes and in tests comprised of 10 items. Although Seong (1990) reported that accuracy was improved when the specified prior distribution matched the true θ distribution, the effect of varying distributional conditions for θ in the present study was negligible. Test length seemed to be the most significant factor affecting θ estimation, but only at the more extreme ranges of θ. Increasing the length of the test did significantly reduce estimation errors and variability of the estimates. In addition, as found by Seong, increasing the number of quadrature points should increase the accuracy of estimates. Finally, the distribution of θ estimates was platykurtic and truncated at the high and low ends of the score range, reducing the SD of the distribution.

The generalizability of results is an issue in any simulation study. Thus, although the effects of sample size and test length have been studied by a number of researchers, more studies are needed in order to obtain definitive conclusions about the impact of the underlying θ distribution on the estimation of item and θ parameters. One feature of the present study, however, that increased the generality of the results was the use of 100 replications at each combination of the manipulated factors. Stone (1990) demonstrated that item parameter estimates across 100 replications are variable to the degree that the generality of results based on one or a small number of randomly generated datasets is likely to be compromised. This is especially true for conditions (e.g., small samples and/or non-normal θ distributions) that may be the very reason for a study.

References

Baker, F. B. (1982). *GENIRV: A program to generate item response vectors* [Computer program]. Madison WI: University of Wisconsin, Madison (Unpublished Manuscript).

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

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Dempster, A. P., Laird, N., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, 39,* (Series B), 1–38.

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

Drasgow, F., & Hulin, C. L. (1988). Cross-cultural measurement. *Interamerican Journal of Psychology, 21,* 1–24.

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

Fletcher, R., & Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. *Computer Journal, 2,* 163–168.

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

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 169–194.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville IN: Scientific Software.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in the estimation of item parameters. *Psychometrika, 54,* 661–680.

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

Qualls, A. L., & Ansley, T. N. (1985, April). *A comparison of item and ability parameter estimates derived from LOGIST and BILOG.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

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.

Stone, C. A. (1990, April). *IRT based monte carlo research: How many replications are necessary?* Paper presented at the Annual Meeting of the National Council of Measurement in Education, Chicago.

Thissen, D. (1986). *MULTILOG user's manual* [Computer program manual]. Mooresville IN: Scientific Software.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 45–56). Vancouver BC: Educational Research Institute of British Columbia.

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

Author's Address

Send requests for reprints or further information to Clement Stone, 110 OEH, University of Pittsburgh, Pittsburgh PA 15260, U.S.A.