An Empirical Study of the Effects of Small Datasets and Varying Prior Variances on Item Parameter Estimation in BILOG

Michael R. Harwell and Janine E. Janosky
University of Pittsburgh

Long-standing difficulties in estimating item parameters in item response theory (IRT) have been addressed recently with the application of Bayesian estimation models. The potential of these methods is enhanced by their availability in the BILOG computer program. This study investigated the ability of BILOG to recover known item parameters under varying conditions. Data were simulated for a two-parameter logistic IRT model under conditions of small numbers of examinees and items, and different variances for the prior distributions of discrimination parameters. The results suggest that for samples of at least 250 examinees and 15 items, BILOG accurately recovers known parameters using the default variance. The quality of the estimation suffers for smaller numbers of examinees under the default variance, and for larger prior variances in general. This raises questions about how practitioners select a prior variance for small numbers of examinees and items. *Index terms: BILOG, item parameter estimation, item response theory, parameter recovery, prior distributions, simulation.*

Mislevy (1986) presented a marginalized Bayesian procedure for estimating item parameters that is a direct extension of the approach of Bock and Aitkin (1981). A distinguishing characteristic of Mislevy's procedure is its ability to constrain item parameter estimates (e.g., discrimination) from assuming unreasonable values. The conceptual and mathematical advantages of the Bayesian approach over earlier procedures, along with preliminary empirical evidence (Mislevy, 1986), suggest that this procedure may solve many of the remaining estimation problems in item response theory (IRT) (e.g., poorly determined parameter estimates even for large numbers of examinees).

The implementation of the key components of the marginalized Bayesian procedure in the BILOG computer program (Mislevy & Bock, 1986) allows practitioners to take advantage of its comprehensive nature and mathematical elegance. However, relatively little is known about the performance of BILOG in estimating item parameters for small datasets (e.g., small numbers of items or examinees). Furthermore, little empirical evidence is available about the effects of different prior distribution variances on item parameter estimation in BILOG or the extent to which the default prior variance in BILOG is appropriate for small datasets.

The purpose of this study was to investigate the ability of BILOG to recover known item parameters for different numbers of items, examinees, and variances of the prior distributions of discrimination parameters for the two-parameter logistic IRT model. The intent was to determine the lower limit (in terms of numbers of examinees, items, and prior variance) at which the program satisfactorily recovers item parameters. The results should help to define limits on the appropriate use of BILOG and its default prior variance for small datasets. The two-parameter model was selected because of its applicability in a variety of settings (e.g., constructing multiple-choice items that respondents are unlikely to guess correctly) and because parameter estimation in this model is less intractable than

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 3, September 1991, pp. 279-291

0146-6216/91/030279-13$1.90

in the three-parameter model. Once the effects of small datasets and varying prior variances have been documented for the two-parameter model, the research focus would naturally be expected to shift to the more complex three-parameter model.

Characteristics of BILOG

Prior Distributions

A distinguishing characteristic of BILOG is its ability to impose prior distributions on the discrimination parameters. The default option in BILOG is designed to impose a prior distribution on discrimination parameters but no prior distribution for location parameters; it also permits prior distributions to be imposed on location parameters if desired. Ability (θ) is assumed to be a continuous random variable with a prior distribution. Imposing a prior distribution on discrimination parameters has the effect of constraining estimates of these parameters from assuming unreasonable values, which is a problem that has plagued item parameter estimation in IRT (Mislevy, 1986). However, this advantage depends on the specification of appropriate prior distributions and an appropriate IRT model.

One assumption underlying specification of a prior distribution for discrimination parameters is that the parameters are independently and identically distributed (Swaminathan & Gifford, 1985). It is also assumed that such parameters are exchangeable (i.e., the prior probability distribution for a particular parameter is no different from that of any other parameter of the same type). By comparison, the estimation of location parameters usually proceeds smoothly and there is less need to impose prior distributions on these parameters (Mislevy, 1986).

The variance of a prior distribution plays a key role in estimating parameters. A prior distribution is said to be informative if its variance is small. A small variance implies that the values of a parameter will be tightly clustered around the mean of the prior distribution, with some values of the parameter more likely than others. The primary effect of an informative prior on parameter estimation is to "shrink" the estimate toward the mean of the parameter's prior distribution by an amount that is proportional to the information contained in the prior distribution of that parameter (Mislevy & Stocking, 1989). Other things being equal, smaller prior variances lead to estimates that are nearer the mean of the prior than do larger variances. This makes it less likely that the estimates will assume unreasonable values.

The discrimination parameters (a_j) in most testing settings are typically greater than zero. This suggests that the distribution of a_j can be modeled by a unimodal and positively skewed distribution such as the lognormal. For computational simplicity, BILOG performs a logarithmic transformation of the form $\alpha_j = \log a_j$. This results in each α_j having a normal distribution with a density that is proportional to

$$\exp\{-.5[(\alpha_j - \mu_\alpha)/\sigma_\alpha]^2\} \quad . \tag{1}$$

Following Lindley and Smith (1972), μ_α and σ_α^2 are known as hyperparameters. Imposing (normal) prior distributions on the α_j makes it unlikely that unreasonable estimates will be obtained. This may also improve the estimation of location and θ parameters (Swaminathan & Gifford, 1985).

Parameter Estimation

Consider the probability of a correct response under a two-parameter logistic IRT model for dichotomous item response data:

$$P_j(y_{ij} = 1 \mid \theta_i) = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]} = P_j(\theta_i) \quad , \tag{2}$$

for $i = 1, 2, 3, \ldots, n$ examinees, and $j = 1, 2, 3, \ldots, J$ items; $y_{ij} = 0,1$ is the binary response of examinee i to item j, θ_i is the ability of examinee i, a_j is the item discrimination parameter, and b_j is the item location parameter.

The Bayes marginalized estimation equations used in BILOG to estimate a_j and b_j for the two-parameter logistic model are

$$\alpha_j: \quad e_j^{\alpha} \sum_k^q [\bar{r}_{jk} - \bar{n}_{jk} P_j(X_k)] (X_k - b_j) - (\alpha_j - \mu_\alpha)/\sigma_\alpha^2 = 0 \quad , \tag{3}$$

and

$$b_j: -e_j^{\alpha} \sum_k^q [\bar{r}_{jk} - \bar{n}_{jk} P_j(X_k)] = 0 \tag{4}$$

(Harwell & Baker, in press). The solutions to Equations 3 and 4 require an integration over the prior distribution of θ to produce a marginalized likelihood. BILOG uses numerical quadrature to approximate this integration. X_k represents the finite θ value of the kth θ group. $\bar{r}_{jk}$ and $\bar{n}_{jk}$ represent "artificial" data related to the use of numerical quadrature (Harwell, Baker, & Zwarts, 1988).

The $(\alpha_j - \mu_\alpha)/\sigma_\alpha^2$ term in Equation 3 represents the contribution of the prior distribution imposed on α_j to the marginalized likelihood equations, and it requires estimates of the μ_α and σ_α^2 terms. No prior distribution is imposed on the b_j; thus, there is no term in Equation 4 representing the contribution of a prior distribution. BILOG permits the value of μ_α to be specified or estimated from the item response data with the "FLOAT" option. This reduces in BILOG to taking the arithmetic average of the J estimated α_j. Mislevy and Stocking (1989) suggest that practitioners should employ this option in most instances. The results of Baker (1990) suggest that for large datasets (e.g., 1,000 examinees and 40 items), it makes little difference whether or not the FLOAT option is used. Note that BILOG assumes that the means of the prior distributions are correct. The effect on parameter estimates when these means are incorrect is not known. The value of σ_α^2 cannot be estimated from the data; it must be specified, or the default value must be employed.

BILOG permits prior distributions with different variances to be specified for each item parameter. This may improve estimation in some instances. In this study, all of the prior distributions for the α_j were assumed to be normal with mean μ_α and variance σ_α^2.

Previous Research Using BILOG

Mislevy (1986) simulated data for 1,000 examinees on 20 items and reported that BILOG accurately recovered true item parameters. Yen (1987) compared the estimation capabilities of the BILOG and LOGIST (Wood, Wingersky, Barton, & Lord, 1982) computer programs for the three-parameter logistic IRT model. Yen examined the ability of BILOG to recover known parameters for tests of varying difficulty, different numbers of items (10, 20, 30, and 40), and different prior θ distributions. In all cases, the simulated item response data were based on 1,000 examinees. Normal priors with $\mu_b = 0$, $\sigma_b = 1$ were used for b_j. Normal priors with $\mu_\alpha = 0$, $\sigma_\alpha^2 = .5^2$ were used for log-transformed a_j. Yen found that BILOG capably recovered known item parameters as measured by several indices, including the correlation between true and estimated parameters and the root mean square deviation (RMSD), which was computed as the square root of the average of the squared differences between the estimated and true parameters.

Baker (1990) used a parameter-recovery study to examine the equating of BILOG results to an

underlying metric for the two-parameter logistic IRT model. Baker varied the true discrimination and location parameters to produce tests with designated properties (e.g., a difficult test with moderate discrimination). All results were based on tests of 40 items and 1,000 examinees. Baker also varied the variance of the prior distributions for the log-transformed discrimination parameters ($.75^2$, $.5^2$, and $.25^2$). The accuracy of parameter estimation was assessed by comparing the mean estimated and equated discrimination, location, and θ parameters to the mean of the true parameters. Baker's results indicated that BILOG accurately recovered item parameters. Different prior variances had little effect on the accuracy of parameter estimation.

Lim and Drasgow (1990) compared marginal maximum likelihood estimation (MMLE; with no prior distributions for item parameters) and Bayes modal estimation when assessing differential item functioning using a two-parameter logistic model. These authors examined the parameter recovery capabilities of BILOG for samples of 250 or 750 examinees on a 20-item test. Normal priors with $\mu_b = 0$, $\sigma_b = 2$ were used for b_j in Bayes modal estimation. Normal priors with $\mu_\alpha = 0$, $\sigma_\alpha^2 = 1$ were used for the log-transformed a_j. Unlike the other studies, Lim and Drasgow performed 50 replications for each combination of conditions, which permitted an investigation of the role of sampling error. They reported that Bayes modal estimates showed less estimation error than MMLE estimates when $n = 250$. The estimation error associated with the Bayes modal and MMLE estimates for $n = 750$ was similar.

Results for all of these studies are consistent with those of Swaminathan and Gifford (1985), who employed slightly different Bayesian methods to estimate parameters. However, none of these studies investigated the ability of BILOG to recover item parameters under conditions of small datasets and varying prior variances.

Method

This study investigated the ability of BILOG to recover item parameters for small numbers of examinees and items, and different variances of the prior distributions of a_j. Three factors were examined: number of examinees (75, 100, 150, 250, 500, or 1,000), number of items (15 or 25), and variance of the prior distributions for a_j (no prior, $.75^2$, $.5^2$, $.25^2$, or $.1^2$ in a lognormal metric). The number of examinee conditions was selected to represent a range from a small ($n = 75$) to a moderately large sample ($n = 1{,}000$). The $n = 250$ condition was included because this value often appears when Bayesian estimation methods are employed (e.g., Mislevy, 1986). The number of items conditions were believed to represent very short (i.e., 15) and moderately short tests.

In the no-prior-variance case, parameter estimation took place in the absence of prior information for α_j. In this condition, item parameter estimation (assuming that an appropriate prior is imposed for abilities and that the IRT model is correct) is equivalent to the marginalized maximum likelihood solution of Bock and Aitkin (1981). The remaining prior variances ($.75^2$, $.5^2$, $.25^2$, and $.1^2$) represent increasingly strong (i.e., informative) prior distributions (Mislevy & Stocking, 1989). In BILOG, $.5^2$ is the default prior variance for α_j.

The means and variances of the selected prior distributions for a_j are given in the metric of the a_j in Table 1. The transformation equations are given in Brown and Aitchison (1957, p. 8). The μ_a and σ_a^2 are also given in Table 1, assuming that $\mu_\alpha = 0$. Note that the BILOG default values of $\mu_\alpha = 0$ and $\sigma_\alpha^2 = .5^2$ are equivalent to $\mu_a = 1.13$ and $\sigma_a^2 = .36$.

All combinations of conditions were examined, resulting in $5 \times 2 \times 5 = 50$ BILOG computer runs. The following process was repeated for each BILOG run: (1) item response data corresponding to a two-parameter logistic model were generated for a specified number of examinees and items using the GENIRV (Baker, 1982) computer program; (2) a_j and b_j were sampled randomly for each item and

Table 1
Mean and Variance of the Prior Distributions of the Discrimination Parameters Used in the Simulation Study ($\mu_\alpha = 0$)

	Prior Distribution			
Statistic	1	2	3	4
σ^2_α	$.75^2$	$.5^2$	$.25^2$	$.1^2$
μ_a	1.32	1.13	1.03	1.01
σ^2_a	1.32	.36	.07	.01
σ_a	1.15	.6	.26	.1

θ value for each examinee; (3) θ and b_j were sampled from a range of −3 to +3 from a normal distribution with mean = 0 and SD = 1; and (4) a_j were sampled randomly from a uniform distribution from a range of .6 to 1.9 (in the logistic metric). This range was selected because estimated discrimination parameters for real data often fall within these values.

Additional computer runs were made to determine if sampling a_j from a uniform distribution with a range of .1 to 3 produced approximately the same pattern of results as those for the .6 to 1.9 range. Several computer runs were also made sampling θ and b_j from uniform distributions to ensure that sampling from a particular distribution in GENIRV did not predispose the results. No systematic differences in the accuracy of estimation were observed when sampling θ and b_j from normal versus uniform distributions.

A trial-and-error process led to the following method of generating item response data in GENIRV. For a given number of items, the same random seed was used for the various numbers of examinees and prior variance conditions. This resulted in the same a_j and b_j being sampled and serving as the true parameters for each of these conditions. Based on these data, BILOG was used to recover the same item parameters for 15 items across all numbers of examinees and prior variance conditions. Because GENIRV used a single seed to generate item parameters and responses, these parameters and responses for the 15-item case served as the first 15 items for the 25-item case. Thus, the 15- and 25-item cases are directly comparable. The means and SDs of the a_j and b_j sampled in GENIRV are given in Table 2.

The above decision was based on the following reasoning. First, it was desired to hold the effect constant for all but one of the conditions in order to examine the effect of the condition that was varied (e.g., $\sigma^2_\alpha = .5^2$ versus $.25^2$). Second, initial computer runs using different seeds for small datasets led to tests that seemed, at best, "randomly parallel." For example, tests with very different item parameters resulted when item response data were generated by using (1) a particular seed for the 15 items, $n = 100$ examinees, and no-prior-variance case; and (2) a different seed for the 15 items, $n = 100$ examinees, $\sigma^2_\alpha = .75^2$ case. The tendency for one seed to produce a very different pattern of large and/or small item parameters than other seeds made it difficult to compare reasonably the ability of BILOG to recover item parameters across conditions.

Table 2
Mean and Standard Deviation of the Parameters Sampled in GENIRV (Logistic Metric)

No. of Items	μ_a	σ_a	μ_b	σ_b
15	1.3	.22	−.044	.97
25	1.29	.36	−.142	1.17

The item response data generated using GENIRV were submitted to BILOG for analysis. The MINITAB (1989) statistical package was used to transform the estimated a_j and b_j from BILOG to the same metric used in GENIRV, based on the equating equations in Baker (1990). These estimates were compared to the corresponding true item parameters in GENIRV. A variety of statistics were used to assess the accuracy of parameter estimation. Consistent with previous parameter recovery studies, the correlations of the estimated and true parameters, and the RMSDs, were computed, where

$$\text{RMSD} = \left[\sum_{j}^{J} \frac{(\hat{a}_j - a_j)^2}{J}\right]^{1/2}, \tag{5}$$

and $\hat{a}_j$ is an estimated discrimination parameter.

Smaller values of RMSD indicate less estimation error. RMSD′ was also computed after the 10% most extreme squared deviations were trimmed.

Results

The results of assessing BILOG's ability to recover true a_j and b_j are presented in Tables 3 and 4. Selected results for a_j are presented graphically in Figures 1 and 2. (The correlations between the estimated and true parameters were relatively uninformative; thus, the discussion below relies on the RMSD and RMSD′ indices.) For the 15-item test, smaller prior variances led to more accurate estimation for $n = 75$, 100, and 150 examinees. For $n > 150$, smaller prior variances did not noticeably improve the accuracy of item parameter estimation (Figure 1a). There was little difference in the accuracy of parameter estimation for $\sigma_\alpha^2 = .25^2$ and $.1^2$ across the number of examinees conditions. However, both $\sigma_\alpha^2 = .25^2$ and $\sigma_\alpha^2 = .1^2$ resulted in less estimation error than the default $\sigma_\alpha^2 = .5^2$ value for $n < 250$. The same results were observed for RMSD′ (Figure 1b).

The accuracy of parameter estimation for the 25-item test (Figure 2) was very similar to the results obtained for the 15-item test (Figure 1). The various prior variances had little effect on the accuracy of estimation once $n > 100$. In fact, the RMSD values for the smallest prior variance were slightly larger than those for the remaining prior variance conditions for larger numbers of examinees; this pattern persisted for the RMSD′ values. This may result from the fact that such a small prior variance causes the slopes comprising the prior distribution to be very close to the mean of the distribution. This produces, in effect, a one-parameter IRT model with varying b_j and a common slope.[1] Parameter estimation of slopes in this model would not be as sensitive to increasing numbers of examinees as it would be in a model in which slope variability in the prior distribution was greater.

One aspect of Tables 3 and 4 deserves further comment. The estimation error for a_j and b_j was typically greater for 25 items than for 15 items, especially when $n < 250$. This suggests that for smaller samples there is a test length by prior variance interaction in terms of estimation error. For example, the $n = 75$ condition in Table 3 shows less estimation error for a_j for a 15- versus a 25-item test for all the prior variance conditions (including no prior variance). Once n exceeds 250, the estimation error for the two test lengths tends to be reasonably similar. One explanation for these results is that they are attributable to sampling error.

Additional computer runs were made to assess the credibility of the sampling error explanation. Because the RMSD values favoring the 15- versus the 25-item case were most distinct for the smallest number of examinees, five replications were performed for each of the no-prior, $.5^2$, and $.1^2$ prior variance conditions for $n = 75$. This process was repeated for the 15- and 25-item cases. Item parameters were randomly sampled for each replication (i.e., item parameters for the 15- and 25-item

[1]Thanks to Walter Way of Educational Testing Service for his comments on this issue.

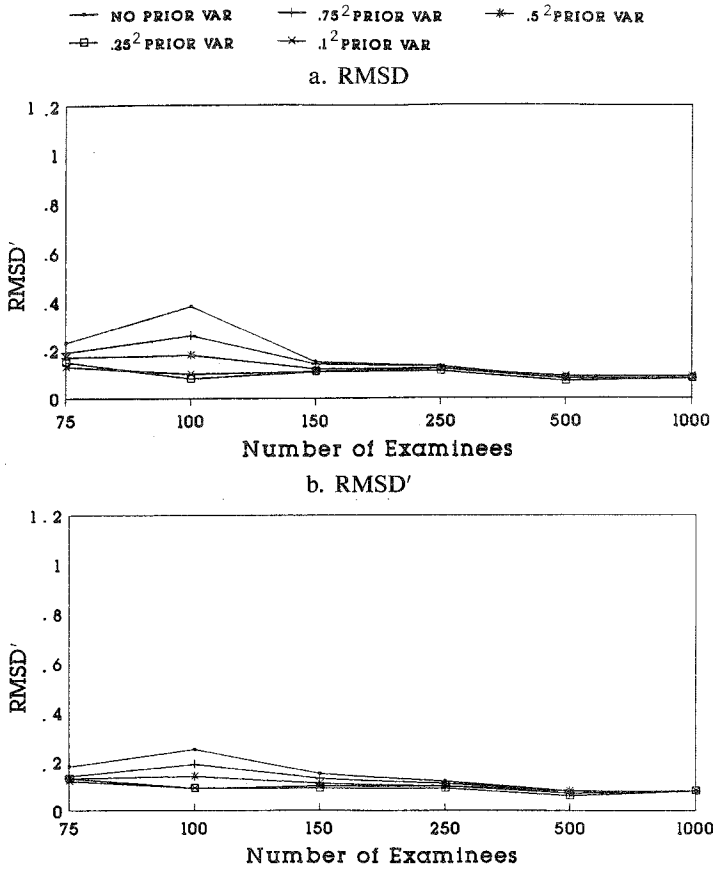
Table 3
Correlation Between Estimated and True Discrimination Parameters (r), RMSD, and RMSD′, for No Prior and Four Levels of Prior Variance, for 15- and 25-Item Tests and N = 75 to 1,000 Examinees (For the 15-Item, No-Prior-Distribution Condition, Two Different Seed Values Were Used to Generate Item Response Data)

No. of Items and Examinees	No Prior						$.75^2$			$.5^2$			$.25^2$			$.1^2$		
	r		RMSD		RMSD′		r	RMSD	RMSD′	r	RMSD	RMSD′	r	RMSD	RMSD′	r	RMSD	RMSD′
	1	2	1	2	1	2												
15 Items																		
N = 75	.38	.49	.23	.70	.18	.65	.41	.19	.14	.41	.17	.13	.41	.15	.13	.41	.13	.12
N = 100	.76	.73	.38	.20	.25	.18	.78	.26	.19	.79	.18	.14	.79	.08	.08	.76	.10	.09
N = 150	.61	.70	.15	.20	.15	.19	.62	.14	.13	.62	.12	.11	.60	.10	.09	.57	.11	.10
N = 250	.73	.91	.13	.07	.12	.06	.73	.13	.11	.73	.12	.10	.73	.11	.09	.74	.12	.10
N = 500	.82	.88	.08	.09	.08	.08	.82	.08	.08	.82	.08	.07	.83	.07	.06	.84	.09	.08
N = 1,000	.86	.99	.08	.06	.08	.05	.86	.08	.08	.86	.08	.08	.83	.08	.08	.86	.09	.08
25 Items																		
N = 75	.04		.42		.31		.06	.35	.27	.08	.30	.25	.14	.23	.21	.21	.21	.21
N = 100	.58		.24		.23		.60	.20	.18	.59	.19	.18	.60	.19	.17	.61	.21	.20
N = 150	.67		.20		.19		.68	.18	.15	.69	.17	.15	.71	.17	.16	.71	.20	.20
N = 250	.87		.14		.13		.87	.14	.12	.87	.14	.12	.88	.15	.13	.89	.18	.19
N = 500	.88		.12		.12		.88	.12	.11	.88	.11	.10	.89	.10	.10	.89	.14	.14
N = 1,000	.96		.06		.05		.97	.06	.05	.96	.06	.05	.97	.06	.05	.97	.10	.10

Table 4
Correlation Between Estimated and True Location Parameters (r), RMSD, and RMSD′, for No Prior and Four Levels of Prior Variance, for 15- and 25-Item Tests and N = 75 to 1,000 Examinees (For the 15-Item, No-Prior-Distribution Condition, Two Different Seed Values Were Used to Generate Item Response Data)

No. of Items and Examinees	No Prior						$.75^2$			$.5^2$			$.25^2$			$.1^2$		
	r		RMSD		RMSD′		r	RMSD	RMSD′	r	RMSD	RMSD′	r	RMSD	RMSD′	r	RMSD	RMSD′
	1	2	1	2	1	2												
15 Items																		
N =	.89	.19	.64	3.74	.39	1.41	.94	.40	.33	.95	.34	.30	.97	.27	.26	.97	.25	.24
N = 100	.95	.96	.36	.29	.31	.25	.97	.26	.25	.98	.21	.19	.98	.18	.18	.98	.19	.18
N = 150	.98	.97	.22	.20	.21	.18	.98	.21	.20	.98	.21	.20	.98	.21	.20	.98	.22	.21
N = 250	.98	.98	.19	.14	.14	.12	.99	.15	.14	.99	.15	.13	.99	.12	.10	.99	.10	.07
N = 500	.99	.99	.11	.12	.11	.09	.99	.11	.10	.99	.11	.10	.99	.10	.10	.99	.11	.10
N = 1,000	.99	.99	.08	.10	.04	.20	.99	.08	.06	.99	.08	.06	.99	.07	.06	.99	.06	.06
25 Items																		
N = 75	.92		.99		.67		.96	.52	.41	.96	.43	.32	.96	.36	.26	.96	.36	.26
N = 100	.91		.75		.53		.96	.41	.36	.96	.36	.34	.97	.29	.26	.97	.35	.32
N = 150	.94		.52		.25		.96	.34	.24	.97	.29	.24	.98	.24	.22	.98	.25	.22
N = 250	.98		.22		.21		.98	.23	.21	.98	.24	.22	.98	.26	.22	.97	.30	.23
N = 500	.99		.21		.16		.99	.20	.18	.99	.19	.15	.99	.17	.16	.99	.19	.15
N = 1,000	.99		.11		.09		.99	.12	.09	.99	.11	.09	.99	.13	.09	.99	.17	.09

Figure 1
RMSD and RMSD′ for Discrimination Parameters (15 Items)

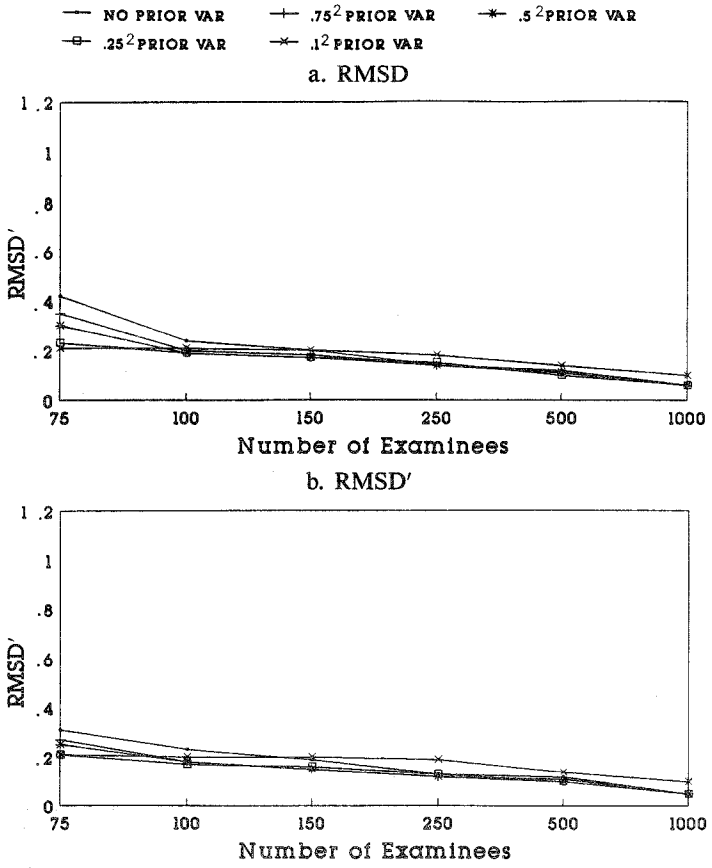


cases differed). The average of the five RMSD values, the SD of the five RMSDs, and the average of the five SDs of the randomly sampled a_j and b_j values in GENIRV are reported in Table 5 for each condition.

The results for a_j in Table 5 suggest that the apparent prior variance by test length interaction is attributable to sampling error. In particular, the variability in the a_j in Table 5 is very similar for both the 15- and 25-item cases; it is also similar to the variability in the a_j for 25 items in Table 2. However, the variability in the a_j in Table 2 for 15 items is noticeably less than that reported in Table 5. This suggests that the a_j that were initially randomly sampled in GENIRV for 15 items were less variable than would be expected.

The smaller-than-expected variability in these true a_j values may partially account for the lower RMSD and RMSD′ values in Table 3 for 15 items. Under these conditions, the a_j values would tend to be relatively close in value, which would make very large or very small values unlikely and thus improve estimation. In the case of b_j, the initially sampled values in GENIRV (Table 2) for the 25-item case are more variable when compared to the values in Table 5. This might account to some extent for the smaller RMSD and RMSD′ values for 25 items in Table 3, as compared to those in Table 5. No distinguishable pattern emerged for the 15-item case. Table 5 generally shows less estimation

Figure 2
RMSD and RMSD′ for Discrimination Parameters (25 Items)



error for $n = 75$ than do Tables 3 or 4.

Figure 3 presents results on the accuracy of parameter recovery in BILOG when the sampled range of a_j was .1 to 3. This provided a check on whether the pattern of results for the a_j parameter would change in a systematic way if a wider range of values were sampled in GENIRV. Comparison of Figure 1 with Figure 3 shows that, except for the large error in the no-prior-variance case for $n = 75$ examinees, approximately the same pattern of RMSD and RMSD′ values were obtained for the .6 to 1.9 and the .1 to 3 ranges.

When $n > 150$, the size of the prior variance made little difference in the accuracy of parameter estimation for both the 15- and 25-item tests as measured by the RMSD and RMSD′ (except for $\sigma_\alpha^2 = .1^2$). These results are probably related to the role of the likelihood function and the prior distributions in estimating item parameters. Other things being equal, as the number of examinees increases, the contribution of the likelihood to estimating item parameters (based on the observed data) increases relative to the contribution of the prior distributions. For smaller numbers of examinees (e.g., $n = 75$, 100, or 150), prior variances noticeably affected the accuracy of parameter estimation for both 15- and 25-item tests.

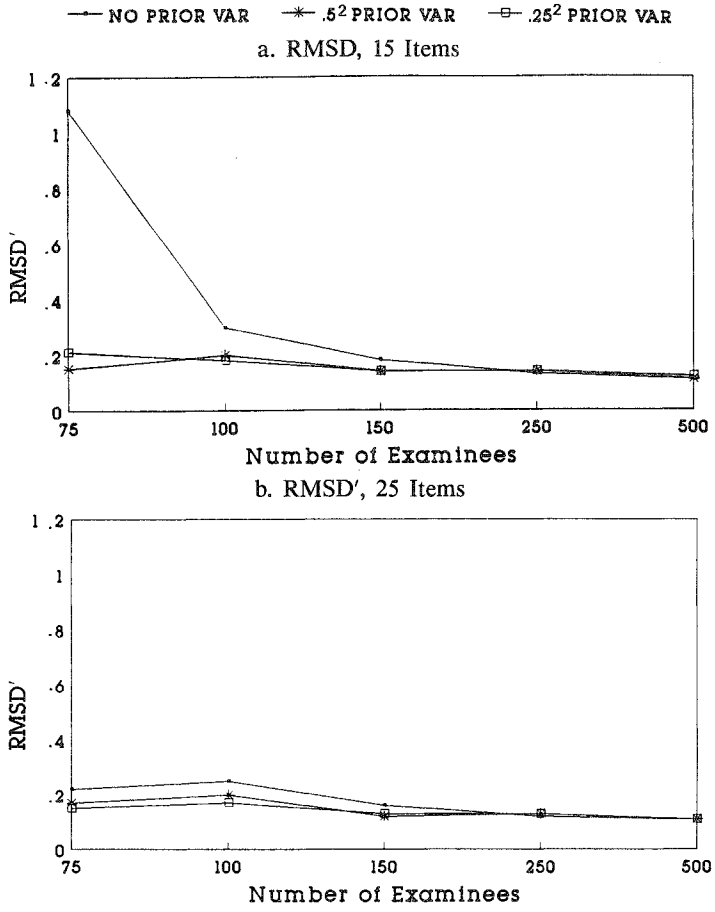
Table 5
Average of Five RMSD Values (μ_{RMSD}), SD of the Five RMSDs (σ_{RMSD}), Average of the Five SDs of the a_j (σ_a) and b_j (σ_b) Randomly Sampled in GENIRV, for Additional BILOG Computer Runs Under Three Prior Variance Conditions

	Prior Variance		
Statistic	No Prior	$.5^2$	$.1^2$
Discrimination Parameter			
15 Items			
μ_{RMSD}	.46	.22	.22
σ_{RMSD}	.2	.01	.01
σ_a	.38	.35	.42
25 Items			
μ_{RMSD}	.38	.21	.19
σ_{RMSD}	.08	.01	.01
σ_a	.37	.36	.36
Location Parameters			
15 Items			
μ_{RMSD}	1.04	.32	.34
σ_{RMSD}	.85	.03	.03
σ_b	1.11	1.09	.88
25 Items			
μ_{RMSD}	.52	.34	.28
σ_{RMSD}	.10	.03	.02
σ_b	1.03	1.00	.99

Conclusions

The results of this study suggest several conclusions. For tests of 15 and 25 items, samples of 250 examinees or more neutralize the effect of prior variances—probably because of the increasing role of the likelihood function in estimating item parameters for larger samples. For smaller numbers of examinees and a very short test (i.e., 15 items), the size of the prior variance plays a prominent role in the quality of item parameter estimation. For longer tests (i.e., 25 items), the role of prior variance is neutralized when $n > 100$. These results suggest that practitioners should not rely on the BILOG default prior variance of $.5^2$ for discrimination parameters for short tests (i.e., 15 items) and small numbers of examinees ($n < 250$). However, the literature still lacks guidelines to assist practitioners in selecting prior variances less than or greater than the BILOG default value of $.5^2$.

Figure 3
RMSD and RMSD′ for Discrimination Parameters When the Sampled Range of a_j Was .1 to 3



References

Baker, F. B. (1982). *GENIRV: A program to generate item response vectors.* Unpublished manuscript, University of Wisconsin, Laboratory of Experimental Design, Madison.

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

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

Brown, J., & Aitchison, J. A. C. (1957). *The lognormal distribution.* Cambridge, England: Cambridge University Press.

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

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

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

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, 34,* (Series B), 1–41.

MINITAB handbook. (1989). Boston MA: Duxbury Press.

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

Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG I maximum likelihood item analysis and test scoring: Logistic model.* Mooresville IN: Scientific Software, Inc.

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

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

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

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

Acknowledgments

The authors thank Frank Baker for the use of his GENIRV computer program, and two anonymous reviewers for their helpful comments.

Author's Address

Send requests for reprints or further information to Michael Harwell, 5B27 Forbes Quad, University of Pittsburgh, Pittsburgh PA 15260, U.S.A.