Some Observations on the Metric of PC-BILOG Results

Frank B. Baker
University of Wisconsin

The computer program PC-BILOG uses the estimated posterior θ distribution to establish the location and metric of the θ scale. This approach to solving the identification problem has not been examined extensively. Consequently, this study investigated the equating of PC-BILOG results to an underlying metric when a two-parameter IRT model was used. The simulation results showed that the means of the estimated item and θ parameters generally were insensitive to characteristics of the prior distribution on the item discriminations. The finding of greatest interest was that the PC-BILOG procedures preserved the variability of true θ distributions having small variances while standardizing the variability of those having large variances. However, in both cases the results could be equated to the true metric using existing techniques.
Index terms: ability metric, Bayesian estimation, BILOG, equating, item response theory, prior distributions.

Many practical applications of item response theory (IRT), such as equating, depend on the specification of a trait (θ) scale metric. One of the problems inherent in IRT, however, is that the values of the examinee and item parameters are unique only under a linear transformation. In the current IRT literature, this is referred to as the identification problem. Each computer program for estimating item and examinee parameters under IRT implements a particular approach for resolving this problem. Two of the commonly used IRT computer programs, LOGIST (Wingersky, Barton, & Lord, 1982) and PC-BILOG (Mislevy & Bock, 1982, 1987), implement different approaches to the identification problem that yield similarly-appearing numerical values for the parameter estimates, but they are based on fundamentally different logic.

The approach used in the LOGIST program is well known and was used here only as a point of reference. Because the scheme used in the PC-BILOG program is less well known, the focus of the present paper is on the characteristics of the parameter estimates it produces relative to the θ scale metric it employs; thus the goal was to examine the nature of the θ scale metric employed by PC-BILOG, and the influence several analysis options had on the results. Particular emphasis was given to transforming the obtained results to an underlying (true) metric because this process has implications for both the equating of tests and simulation studies that deal with parameter recovery.

Background

The joint maximum likelihood estimation (JMLE) procedure (Birnbaum, 1968) implemented in the LOGIST program employs two stages. In the first stage, the $\hat{\theta}$s of the examinees are assumed to be known parameters and the item parameters are estimated. In the second stage, the item statistics are assumed to be the parameters and the θ of each examinee is estimated. These $\hat{\theta}$s are standardized to a mean of 0 and a variance of 1, which become the location and scale of the θ metric. An alternating procedure between the two stages is used to achieve convergence. This approach has some interesting properties. Assuming that the item response data were generated under a two-parameter item response function (IRF) model, in a 0,1 metric, with $\bar{\theta} = .5$, $\sigma_\theta = 1.5$, $\bar{\alpha} = .75$ and $\bar{\beta} = -.5$, the LOGIST program would yield the following

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 14, No. 2, June 1990, pp. 139–150

0146-6216/90/020127-11$1.80

asymptotic values:

$$\bar{b} = (\bar{\beta} - \bar{\theta})/\sigma_{\theta} = -1.0/1.5 = -.67 \quad (1)$$

$$\bar{a} = (\sigma_{\theta}/\sigma_{\hat{\theta}})\alpha = (1.5/1.0)\ (.75) = 1.125 \quad (2)$$

$\bar{\hat{\theta}} = 0.0$ and $\sigma_{\hat{\theta}} = 1.0$.

In this case, the θ scale would be compressed through the standardization process and would result in a higher value of the mean item discrimination. If the same specifications were used, but with $\sigma_{\theta} = .6$, the results would be $\bar{b} = -1.0/.6 = -1.667$ and $\bar{a} = (.6/1.0)\ (.75) = .45$. The θ scale would be stretched by the standardization process and would result in a smaller value of the mean item discrimination. The ratio of the mean of the item discrimination estimates yielded by LOGIST to the mean of the true values would contain the information about the units of measurement of the two θ scales. As a result of this process, the difficulty of the test would become relative to a mean of 0 and be expressed in standard deviation units of the distribution of examinee $\hat{\theta}$s.

These relationships are embodied in the linear equating equations such as those due to Loyd and Hoover (1980) and can be used to transform the LOGIST results to the true metric. The basic equations are:

$$\theta_j = \frac{\bar{\alpha}_k}{\bar{\alpha}_j}(\theta_k) + \left[\bar{\beta}_j - \frac{\bar{\alpha}_k}{\bar{\alpha}_j}(\bar{\beta}_k)\right], \quad (3)$$

$$\beta_j = \frac{\bar{\alpha}_k}{\bar{\alpha}_j}(\beta_k) + \left[\bar{\theta}_j - \frac{\bar{\alpha}_k}{\bar{\alpha}_j}(\bar{\theta}_k)\right], \quad (4)$$

$$\alpha_j \sigma_j = \alpha_k \sigma_k, \quad (5)$$

and then

$$\sigma_j = \frac{\bar{\alpha}_k}{\bar{\alpha}_j}\sigma_k. \quad (6)$$

A major problem with the JMLE approach is that consistent estimates of the structural (item) parameters cannot be obtained in the presence of incidental (examinee) parameters because the latter increase with the sample size (Andersen, 1973; Wright, 1977). This problem can be overcome by using the marginal maximum likelihood/EM (MMLE/EM) procedure of Bock and Aitkin (1981) and implemented in the PC-BILOG computer program. Under this approach, the examinee's θ parameters are removed from item parameter estimation by integrating them over an assumed prior distribution of θ that is typically unit normal. The integration is accomplished by approximating a unit normal density by a histogram that is specified by quadrature points X_q and weights $A(X_q)$ (Mislevy & Stocking, 1989). Using a Bayesian approach, the posterior θ distribution is obtained by finding the number of examinees $\bar{N}(X_q)$ expected at each quadrature point, given the prior distribution of θ and the examinee's item responses. In addition, the expected number of correct responses $\bar{R}(X_q)$ to each item at each quadrature point is determined. Both of these values are based on the posterior probability of an examinee's θ belonging at each of the quadrature points.

It is important to note that at this point in the MMLE/EM procedure, the θ of each examinee has not been estimated, but the form of the θ distribution has been estimated. The item parameters are estimated with maximum likelihood using $\bar{N}(X_q)$ and $\bar{R}(X_q)$—that is, "the artificial data" (see Harwell, Baker, & Zwarts, 1988, for the mathematical details of the overall MMLE/EM process). Within each iteration of the item parameter estimation stage of the MMLE/EM process, the obtained discrete posterior θ distribution is used to compute adjusted quadrature weights (for the adjustment equations see Mislevy & Bock, 1985). The adjusted quadrature weights are a normalization of the discrete posterior θ distribution in which the

$$A(X_q) = \bar{N}(X_q)/N \quad (7)$$

are standardized for an a priori set of X_q, such that

$$\sum_{q=1}^{k} A(X_q)\, X_q = 0 \quad (8)$$

and

$$\sum_{q=1}^{k} A(X_q)\, X_q^2 = 1 \quad (9)$$

where N is the total number of examinees. The PC-BILOG program reports the values of the adjusted quadrature weights for the final iteration

of the item parameter estimation stage. The result is that the metric of the obtained item parameter estimates is defined by the mean and variance of the final adjusted quadrature distribution.

Under the MMLE/EM approach, there is no alternating estimation, such as that used in JMLE. Consequently, as implemented in the PC-BILOG computer program, the θ parameters are estimated in a separate stage after the item parameter estimation has been completed. Although maximum likelihood (ML), Bayesian maximum a posteriori (MAP), and Bayesian expected a posteriori (EAP) estimation of θ can be used, Mislevy and Stocking (1989) recommended EAP with a unit normal prior for the θ distribution. The primary effect of this prior is to limit extreme values of the θ estimates ($\hat{\theta}$). As a result, the variance of the EAP $\hat{\theta}$s will tend to be smaller than those of the ML $\hat{\theta}$s. Regardless of the estimation technique used, the values of the item parameter estimates obtained in the first stage are treated as "true" values. Hence the metric of the $\hat{\theta}$s is that of the final adjusted quadrature distribution through the item parameters.

Care needs to be exercised when dealing with θ within this context because there are two θ distributions of interest. In the item parameter estimation stage, the form of the posterior θ distribution is being estimated and it is reported as the final adjusted quadrature weights $A(X_q)$ at the a priori quadrature points (X_q); θ scores for individual examinees are not available. In the second stage, each examinee's θ is estimated and a distribution of $\hat{\theta}$s is obtained. Mislevy (1984) has shown that the estimated distribution of θ is not the same as the distribution of the $\hat{\theta}$.

The PC-BILOG $\hat{\theta}$s share one of the characteristics of the LOGIST estimates—namely, the variance of the distribution of $\hat{\theta}$s is given by

$$\sigma^2_{\hat{\theta}} = \sigma^2_{\theta} + \overline{SE^2_{\hat{\theta}}} \quad , \tag{10}$$

where σ^2_{θ} is the variance of the true θs, and $SE^2_{\hat{\theta}}$ is the average of the variance of the estimation errors. In both cases, as the number of items in a test and the number of examinees increases, the second variance component should decrease. Thus, with reasonable numbers of test items and examinees, $\sigma^2_{\hat{\theta}}$ should approximate the true variance.

To transform the values of the item parameter estimates yielded by PC-BILOG to the true metric, modified versions of Equations 4 and 5 must be used. Because the metric of the item parameters is defined by the mean $\bar{X}_q$ and the standard deviation σ_{x_q} of the final adjusted quadrature distribution, these quantities must be used in place of $\bar{\theta}_k$ and σ_k in these equations. In the final adjusted quadrature distribution, these values will generally be close to 0 and 1. The modified equating equations are:

$$\beta_j = \frac{\bar{\alpha}_k}{\bar{\alpha}_j}(\beta_k) + \left[\bar{\theta}_j - \frac{\bar{\alpha}_k}{\bar{\alpha}_j}(\bar{X}_q)\right] \tag{11}$$

$$\alpha_j \sigma_j = \alpha_k \sigma_{x_q} \quad . \tag{12}$$

The scale information is embedded in the ratio of the mean item discriminations. The mean item difficulty is expressed relative to the mean of the final adjusted quadrature distribution in terms of the standard deviation of the final adjusted quadrature distribution.

The metric information is solely in the item parameter estimates; consequently, PC-BILOG $\hat{\theta}$s can be transformed to the true metric using Equation 3, with the *k* subscript representing the PC-BILOG values and the *j* subscript the true metric. Equation 6 can be used to transform the standard deviation of the distribution of $\hat{\theta}$s. When transforming these values, the two-component nature of this variance should be remembered. A similar caveat applies to the LOGIST $\hat{\theta}$s.

In PC-BILOG, another factor playing a role in determining the metric is the prior distributions imposed on the item parameters. Under a two-parameter logistic model, the PC-BILOG default options place no prior distribution on the item difficulties and a lognormal prior distribution on each item's discrimination. It is assumed that the range of the item discriminations is $0 \leq \alpha \leq \infty$ and the transformation $\alpha = e^x$ is used. The distribution of $x = \log(\alpha)$ is normal (Kendall &

Stuart, 1967, pp. 68–69) and is given by

$$f(x) = \frac{1}{\sigma_x(2\pi)^{1/2}} \exp\left[-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x}\right] \quad -\infty \; x \leq \infty \quad . \tag{13}$$

The PC-BILOG default hyper-parameter values are $\mu_x = 0$ and $\sigma_x = .5$ in a logarithmic metric. In the 0,1 θ metric underlying the item discriminations,

$$\mu_\alpha = \exp\left(\mu_x + \frac{1}{2}\sigma_x^2\right) = 1.13 \tag{14}$$

and

$$\text{var}(\alpha) = \exp(2\mu_x + \sigma_x^2)\,[\exp(\sigma_x^2 - 1] = .364. \tag{15}$$

The value of σ_x can be roughly interpreted as the strength of the prior (Mislevy & Stocking, 1989). The smaller the value of σ_x the more concentrated the prior distribution, and the greater the prior pulls the estimates of the discrimination parameters toward its own mean. The larger the value of σ_x the more the distribution acts like a diffuse prior, and the less the pulling effect. In the PC-BILOG program, the user can set the means of the normal priors or use the Float option, which estimates the means from the item response data. This option does not apply to the standard deviations; these must be set by the user if the default value of .5 is not accepted. In the present study, the logarithmic metric was used when specifying the means and standard deviations of the prior distributions of the item discriminations.

In summary, the θ metric yielded by LOGIST depends on the mean and variance of the $\hat{\theta}$s of the examinees tested. When compared to true parameter values, there is an inverse relationship between the numerical value of the estimated item discrimination and the size of the θ variance. The item difficulty estimates are expressed relative to the mean of the distribution of the $\hat{\theta}$s in units of the group's standard deviation. Due to the use of standardized $\hat{\theta}$, the net effect is that the numerical values of the LOGIST results are group dependent.

In the PC-BILOG program, the frame of reference for the item and $\hat{\theta}$ is the final adjusted quadrature distribution that has been standardized to mean zero and unit variance. The item parameter estimates are expressed in terms of the metric of this distribution. In the second stage, the metric information is conveyed to the $\hat{\theta}$s through the item parameter estimates. As a result, the metric of the $\hat{\theta}$s is also that of the final adjusted quadrature distribution, which depends on both the group's item response data and the prior θ distribution. Thus the metric of the PC-BILOG results are also group dependent, but not in exactly the same sense as are the LOGIST results. It will be shown below that the variance of the true θ distribution plays an important role in determining the numerical values of the PC-BILOG parameter estimates. The interest here was not in comparing LOGIST and PC-BILOG results; instead, it was in examining some of the factors that affect the numerical values of the PC-BILOG results.

Simulation Study

Method

The approach taken was similar to the many simulation studies used to evaluate the parameter recovery characteristics of IRT estimation procedures (e.g., Qualls & Ansley, 1985; Yen, 1987). Three sets of item response data were generated under a two-parameter logistic IRF model in a 0,1 metric. Each set was based on 45-item tests and groups of 500 simulees. Each dataset was analyzed under several options with the PC-BILOG program, and the obtained parameter estimates equated back to the true metric. In contrast to previous item parameter recovery studies, the present interest was on the pattern of the equated results as a function of various analysis options, rather than on the root mean squares of the differences between the estimates and the true parameters. These patterns in turn provided information about the properties of the θ scale metric employed by the MMLE/EM approach that is implemented in PC-BILOG.

The defining values of the generating parameters for the three datasets are given in Table 1. Dataset 1 represented a highly discriminating test

Table 1
Generating Parameters for
45-Item Tests and Groups of 500 Examinees

Dataset	$\bar{\theta}$	σ_θ	α_{min}	α_{max}	β	σ_b
1	0.0	1.0	1.0	2.0	0.0	.8
2	-.5	1.5	.5	1.5	.5	.8
3	.5	.75	.3	.7	-.5	.8

with difficulty matched to the mean θ of a group of examinees whose θ distribution had a unit variance. Dataset 2 represented a difficult test with moderate discrimination administered to a group of examinees having a variance of 1.5. Dataset 3 represented an easy test with low discrimination given to a group having θ variance = .75. In all cases, the true item difficulty and θ parameters were normally distributed and the item discrimination parameters were uniformly distributed. No unusually deviant values of any parameters were observed in the datasets. The summary statistics of the randomly generated item parameters are given as $\bar{\beta}$ and $\bar{\alpha}$ for each dataset in Table 2. The values of the discrimination parameters were reported in a logistic metric (i.e., the 1.7 multiplier was not used).

In the PC-BILOG analysis of the datasets, the following specifications for the prior distribution of item discrimination were employed for each item in a test:

Run A: No prior distribution.

Run B: Default prior $\mu = 0, \sigma = .5$; no Float option.

Run C: Default prior $\mu = 0, \sigma = .5$; with Float option.

Run D: Prior $\mu = 0, \sigma = .75$; no Float option.

Run E: Prior $\mu = 0, \sigma = .75$; with Float option.

Run F: Prior $\mu = 0, \sigma = .25$; no Float option.

Run G: Prior $\mu = 0, \sigma = .25$; with Float option.

The no-prior-distribution analysis was used to obtain a baseline MMLE/EM solution. The .5, .75, and .25 values of the standard deviation (in a logarithmic metric) were used to provide several levels of "strength" of the prior distribution. In particular, the .25 value is "strong" and should pull the item discrimination estimates toward the mean value of the prior. When the Float option was not used, the user-specified value of 0 (in a logarithmic metric) for the mean of the prior was employed. In Dataset 1, this value was well below the true mean value of α. In Dataset 2, it was matched to the mean of α. In Dataset 3, it was well above the true mean value of α.

Of interest here was the degree to which the prior pulled the means of the estimated item discriminations toward the user-specified mean of the prior. When the Float option is used, the mean of the prior distribution is estimated from the item response data, is matched to that of the estimated discriminations, and there should be no pulling effect on their mean. Because the ratio of the mean of the item discrimination estimates to the mean of the true item discrimination parameters contains the information about the θ scale's unit of measurement, the Float option is an important factor in the equating process.

Each of the three datasets was analyzed with the PC-BILOG program under all seven of the above specifications on the discrimination priors. The means, $\bar{b}$ and $\bar{a}$ of the obtained item parameter estimates are reported under the "BILOG" heading in Table 2. These values were equated into the true θ metric and are reported as $\bar{b}$ and $\bar{a}$ under the heading "Equated BILOG." Results for θ are reported under similar headings. Table 2 also reports the means and standard deviations of the final adjusted quadrature distributions.

Results

The equating equations employed here depend on the means of the parameter estimates, so that the results could be sensitive to nonlinearities in the relationship between the estimates and the true parameters. To investigate this, the estimates

Table 2
Mean and Standard Deviation for Generated, Estimated, and Equated Item and θ Parameters

Dataset and Run	Prior α	Prior σ_α	Float Option Used	Equated BILOG $\bar{b}$	Equated BILOG $\bar{a}$	BILOG $\bar{b}$	BILOG $\bar{a}$	Equated BILOG $\bar{\hat{\theta}}$	Equated BILOG $\sigma_{\hat{\theta}}$	BILOG $\bar{\hat{\theta}}$	BILOG $\sigma_{\hat{\theta}}$	Quadrature Distribution Mean	Quadrature Distribution SD
Dataset 1 ($\bar{\beta}$ = .057, $\bar{\alpha}$ = 1.537; $\bar{\theta}$ = .068, σ_θ = 1.006)													
Run 1A	-	-	N	.069	1.485	.009	1.376	.080	.965	.035	1.078	.003	1.093
Run 1B	0	.50	N	.070	1.494	.006	1.325	.081	.954	.034	1.107	.004	1.136
Run 1C	0	.50	Y	.073	1.510	.010	1.374	.110	.964	.036	1.079	.004	1.106
Run 1D	0	.75	N	.072	1.503	.008	1.352	.081	.960	.035	1.092	.004	1.119
Run 1E	0	.75	Y	.074	1.511	.010	1.375	.080	.965	.035	1.078	.004	1.105
Run 1F	0	.25	N	.057	1.464	-.011	1.229	.086	.931	.025	1.164	.003	1.120
Run 1G	0	.25	Y	.066	1.510	.001	1.372	.084	.964	.031	1.078	.004	1.107
Dataset 2 ($\bar{\beta}$ = .649, $\bar{\alpha}$ = .959; $\bar{\theta}$ = -.570, σ_θ = 1.517)													
Run 2A	-	-	N	.686	.918	.971	1.246	-.562	1.393	.039	1.072	.004	1.117
Run 2B	0	.50	N	.585	1.035	.791	1.409	-.573	1.440	-.041	.980	.005	1.115
Run 2C	0	.50	Y	.691	.911	.973	1.241	-.553	1.387	.045	1.071	.005	1.113
Run 2D	0	.75	N	.670	.929	.949	1.261	-.558	1.425	.031	1.062	.006	1.118
Run 2E	0	.75	Y	.685	.914	.973	1.243	-.557	1.390	.043	1.072	.005	1.115
Run 2F	0	.25	N	.678	.876	1.037	1.162	-.528	1.394	.065	1.110	.007	1.114
Run 2G	0	.25	Y	.708	.758	.991	1.237	-.540	1.378	.069	1.069	.000	.929
Dataset 3 ($\bar{\beta}$ = -.467, $\bar{\alpha}$ = .469; $\bar{\theta}$ = .510, σ_θ = .711)													
Run 3A	-	-	N	-.641	.505	-1.500	.360	.684	.593	.000	.773	.000	1.000
Run 3B	0	.50	N	-.456	.631	-.873	.519	.500	.769	-.001	.695	.000	.866
Run 3C	0	.50	Y	-.474	.520	-1.243	.371	.517	.608	.000	.768	.000	.998
Run 3D	0	.75	N	-.734	.585	-1.034	.447	.519	.702	.000	.734	.000	.930
Run 3E	0	.75	Y	-.512	.516	-1.304	.367	.554	.604	.000	.771	.000	.999
Run 3F	0	.25	N	-.460	.725	-.632	.718	.484	.889	-.011	.581	-.001	.718
Run 3G	0	.25	Y	-.427	.515	-1.197	.367	.469	.595	.000	.760	.000	1.000

a, *b*, and $\hat{\theta}$ from all 21 analyses were plotted against the corresponding true parameters. In Datasets 1 and 2, the plots for all analyses exhibited linear relationships with scatter about the line that was typical for IRT results. In all 14 cases, the correlations between the estimates and the true parameters were greater than .91, .97, and .95 for α, β, and θ, respectively. In Figures 1a, 1b, and 1c, the results from Run 1F were used to illustrate a typical case, even though it exhibited the greatest scatter about the relationship lines of any of the 14 runs. On the basis of these plots, it appears that the use of the means in the Loyd and Hoover equations is consistent with the data. The results for Dataset 3 are presented below.

When the summary statistics for the PC-BILOG results were equated to the true metric, they were generally close to those of the true parameters (see Table 2). In Dataset 1, the unit variance of the true θ distribution matched that of the initial quadrature distribution and the average item discrimination was high. In this situation, the effect of the three different values of the standard deviation of the item discrimination priors on the equated means of item difficulty and discrimination was slight. Not using the Float option set the mean of the discrimination prior at a value of $\alpha = 1.13$, and had the effect of pulling the mean of the equated item discriminations below the true value of 1.537 by about .03 to .07. When the Float option was used the mean item discrimination was underestimated by about .03. Thus this option had little effect on the equated mean discrimination estimates.

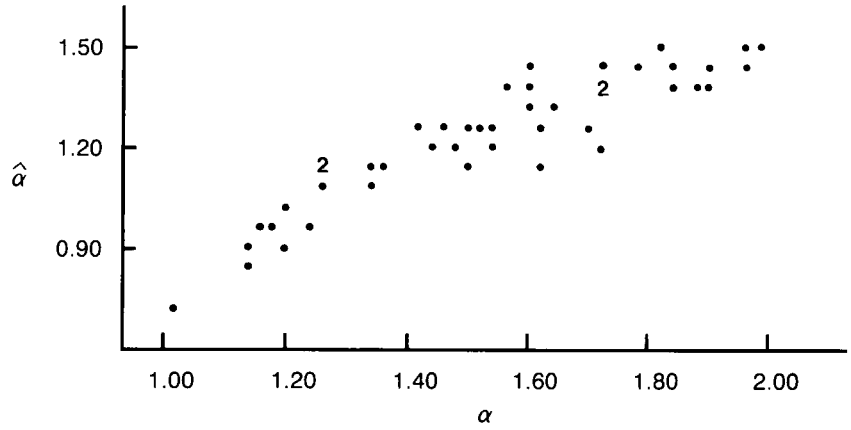
The mean equated item difficulties were only slightly affected by the characteristics of the prior discrimination distributions because the obtained values were little different from those observed when no prior was employed. The equated means of the $\hat{\theta}$s were generally slightly larger than the true mean of .068, both when priors were and were not imposed on the item discriminations. The equated standard deviations of the $\hat{\theta}$s were generally a slight underestimate of the true value of 1, which remains consistent with the use of a Bayesian EAP to estimate θ.

In Dataset 2 the standard deviation of the true θ distribution was larger than that of the initial quadrature distribution, and the item discrimination was moderate. In this dataset the effect of the three different values of the standard deviations of the item discrimination priors on the equated mean item difficulty and discrimination also was slight. Not using the Float option set the mean of the discrimination prior at a value of $\alpha = 1.13$, which roughly matches the true mean discrimination of 1. Using the Float option in this dataset should have also matched the mean item discrimination; however, using the Float option had the effect of lowering the equated mean discriminations. In Run 2F ($\sigma_\alpha = .25$ and no Float option) the equated mean discrimination was .876, which underestimated the true value of .959. In Run 2G ($\sigma_\alpha = .25$ and Float option), however, the equated mean discrimination (.758) was lower than when the Float option was not used, and it underestimated the true value by .2. The means of the equated $\hat{\theta}$s in this dataset were not affected by the characteristics of the discrimination prior distributions. The obtained values were little different from those obtained when no prior was employed, and they were close to the true value of –.570. As anticipated, the equated standard deviations of the $\hat{\theta}$s were consistently a slight underestimate of the true value.

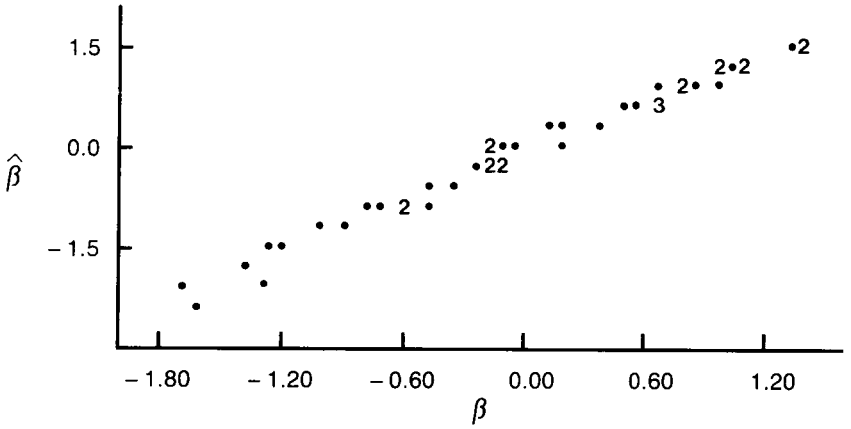
In Dataset 3 the standard deviation of the true θ distribution was smaller than that of the initial quadrature distribution, and the mean item discrimination was low. In this situation the effect of the three different values of the standard deviations of the item discrimination priors on the equated mean item difficulty and discrimination was inconsistent. Four of the equated mean item difficulties were more negative than the true value. The more diffuse the discrimination prior, the more negative the mean equated item difficulty. All the equated mean discriminations were overestimates of the true value. This reflects the fact that the mean of the discrimination prior had a value of $\alpha = 1.13$, which was larger than the true value of .469. Thus it would

Figure 1
BILOG Parameter Estimates and True Values for Dataset 1F

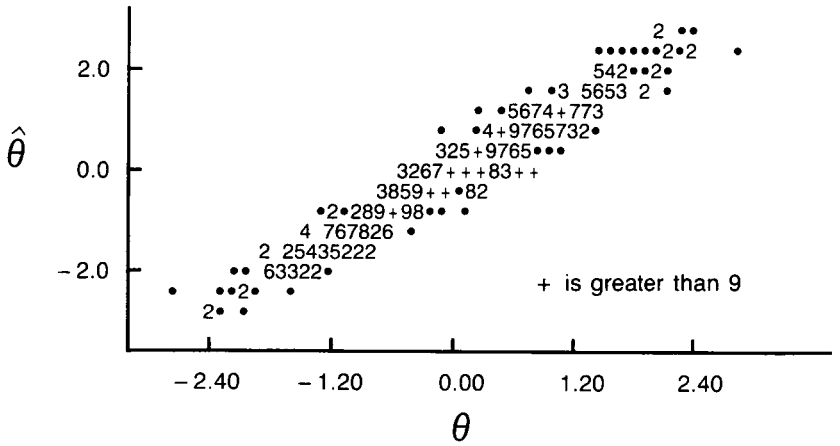
(a) $\hat{a}$ and a (r = .913)



(b) $\hat{b}$ and b (r = .994)



(c) $\hat{\theta}$ and θ (r = .967)



pull the discrimination estimates in an upward direction.

When the standard deviation of the discrimination prior distribution was .5 or .75, use of the Float option appeared to reduce the mean equated discrimination by about .10. In Run 3F (σ_x = .25 and no Float option) the equated mean discrimination had a value of .725, and Run 3G (σ_x = .25 and Float option) yielded a mean of .515. Here the use of the Float option, in conjunction with a "strong" prior, resulted in a mean discrimination that was .21 closer to the true value. When no prior distributions were imposed on the item discriminations, the mean value of the $\hat{\theta}$s (.684) overestimated the true value of .510. When item discrimination priors were used, the equated means of the $\hat{\theta}$s were close to the true values. The values of the equated standard deviations of the $\hat{\theta}$s were smaller than the true values in five of the seven cases.

The most interesting overall result was the manner in which PC-BILOG handled the distribution of $\hat{\theta}$s relative to the true θ distribution. When the variance of the true θ distribution was larger than that of the prior distribution of θ, PC-BILOG yielded a distribution of $\hat{\theta}$s having mean 0 and variance 1. When the variance of the true θ distribution was smaller than that of the prior θ distribution, quite a different phenomenon was noted. PC-BILOG set the mean of $\hat{\theta}$ to 0, but its variance approximated the variance of the true distribution.

In Dataset 3 all but one of the seven PC-BILOG analyses yielded standard deviations of the $\hat{\theta}$s that were close approximations of the true value of .711. This is in sharp contrast to LOGIST where distributions with variances larger and smaller than 1 were all standardized to that value; comparing Run 2E and Run 3E is instructive in this regard. In Run 2E the true θ distribution had σ_θ = 1.517, and the mean item discrimination was $\bar{\alpha}$ = .959. The PC-BILOG results were $S_{\hat{\theta}}$ = 1.072 and $\bar{a}$ = 1.2429, where an increase in discrimination reflects the compression of the θ scale. In Run 3E the true θ distribution had σ_θ = .711, and the mean item discrimination was $\bar{\alpha}$ = .469. The PC-BILOG results were $S_{\hat{\theta}}$ = .771, and $\bar{a}$ = .367, both of which approximate the true values, thus indicating that the true θ distribution was essentially retained.

Run 2E had $\bar{\beta} - \bar{\theta}$ = .649 - (-.570) = 1.219, and PC-BILOG yielded $b - \bar{\hat{\theta}}$ = .973 - .043 = .93. When the latter value was equated to the true metric it became 1.177, which is a good approximation of the true value. Run 3E was $\bar{\beta} - \bar{\theta}$ = -.467 - .510 = -.977, and PC-BILOG yielded $\bar{b} - \bar{\hat{\theta}}$ = -1.304. When it was equated to the true metric it became 1.066, which is a reasonable estimate of the true value. In spite of the differential manner in which the variability of the true θ distribution was handled, the metric of the PC-BILOG parameter estimates was embedded in the item parameter estimates and was appropriate.

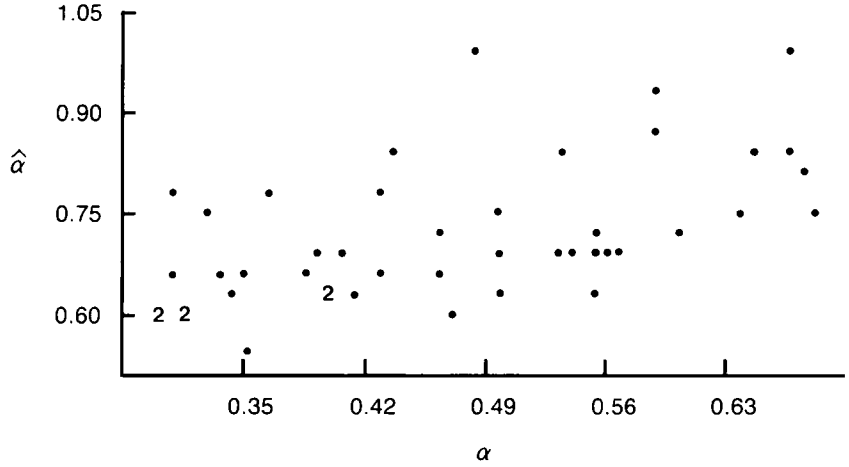
When the Float option was used in Dataset 3, the standard deviation of the final adjusted quadrature distribution was, essentially, 1. When the Float option was not used, the standard deviation of this distribution was smaller, and in the case of Run 3F it was only .718. A similar, but not quite so dramatic, result can be seen in Run 3B, in which the standard deviation was .866.

As a check on the anomalous results for Run 3F, the obtained PC-BILOG estimates were plotted against the true parameter values. These are shown in Figures 2a, 2b, and 2c, for which only the difficulty estimates show a reasonably tight clustering around the relationship line, and r = .945 was obtained. The item discrimination estimates were quite widely scattered about the relationship line, and three values could be considered outliers. The correlation between the estimates and the true parameters was r = .531. The $\hat{\theta}$s were also much more widely scattered about the relationship line than was observed in Datasets 1 and 2. This scatter was reflected in a correlation of .723. Inspection of the plots for the other six runs of Dataset 3 revealed similar results and correlations, but less scatter was observed.

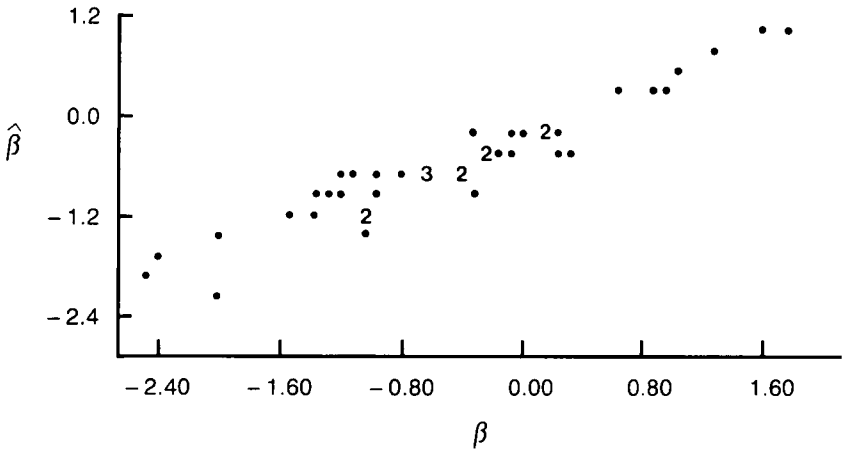
The final adjusted quadrature distribution yielded by Run 3F was very sharply peaked over a small number of quadrature points. This quad-

Figure 2
BILOG Parameter Estimates and True Values for Dataset 3F

(a) $\hat{a}$ and a (r = .531)



(b) $\hat{b}$ and b (r = .824)



(c) $\hat{\theta}$ and θ (r = .723)

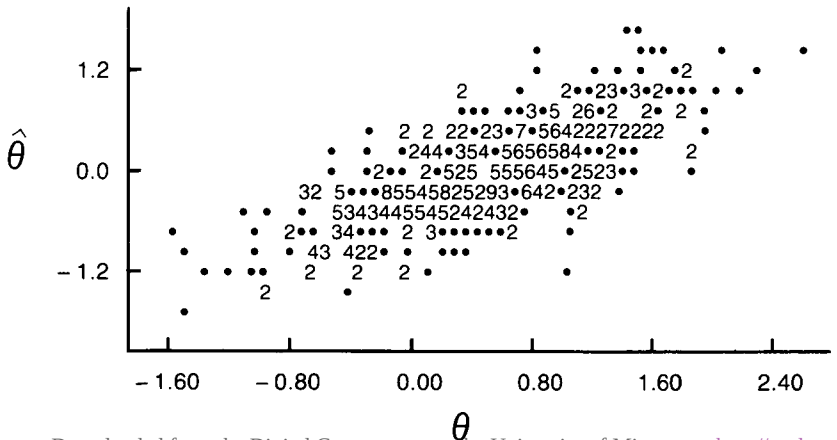
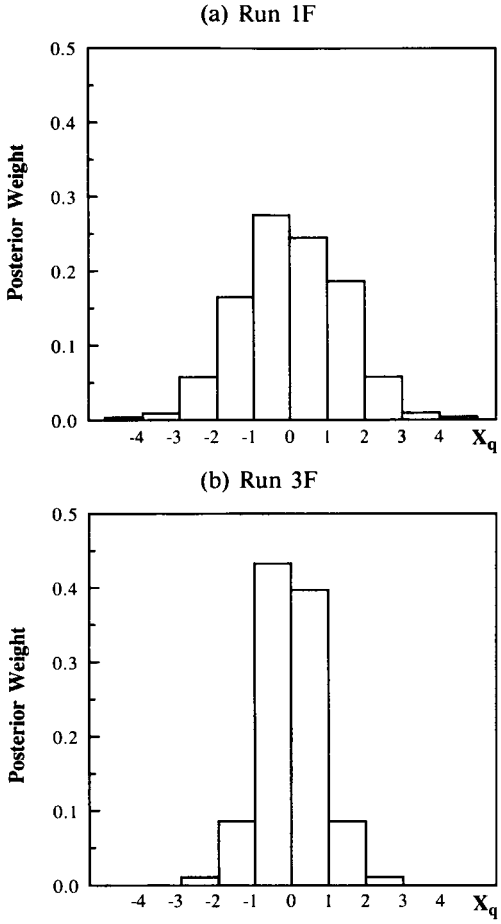


Figure 3
Final Adjusted Quadrature Distributions



rature distribution is compared with that from Run 1F in Figure 3. There is clearly a major difference in the two distributions; yet the procedures for standardizing the histogram of the posterior quadrature distribution are not documented in the PC-BILOG manual. Figure 3 suggests that BILOG was not meeting the constraints given in Equations 8 and 9, under certain conditions.

The specification of the prior distributions of the item parameters in Run 3F was such that the mean of the prior was considerably larger than the true discrimination values. In addition, it would be a "strong" prior that would have a pulling effect in an unwanted direction. It is possible that the PC-BILOG procedures attempted to cope with this by shrinking the quadrature distribution in order to raise the item discrimination values to that desired by the prior. It did accomplish that goal, in that the obtained mean item discrimination was .718. The means of the item difficulties and of the $\hat{\theta}$s were also less affected because both were reasonable values; however, the equated standard deviation of the $\hat{\theta}$s was larger than that obtained in the other six cases for Dataset 3. From these results, it is clear that specifying a prior on the item discriminations that is both wrong and strong is to be avoided. The plotted data for all seven analyses of Dataset 3 also suggested that, with low true item discriminations, the wrong prior mean also has the effect of scattering the discrimination and $\hat{\theta}$s, regardless of the strength of the prior.

There appears to be some interaction between the mean difficulty of the test relative to the mean θ and to the effect of the discrimination prior distributions. The test difficulty in Dataset 1 was matched to the mean θ of the group. The mean item discrimination was .4 above the mean of the prior discrimination distributions. Yet there was little "pulling" effect due to the mean of the prior in any of the six cases. The test in Dataset 3 was very easy for the group ($\bar{\beta} - \bar{\theta} = -.977$).

The mean item discrimination was .66 below the mean of the discrimination prior distributions. The "pulling" effect here was pronounced when the Float option was not used.

Discussion and Conclusions

The equating equations used above were from Loyd and Hoover (1980) and employed means of the item parameters to convey the metric information. The lack of effect of the item discrimination priors in some datasets may have been a reflection of the use of means. Perhaps an equating technique that uses individual item parameters, rather than means, such as that of Stocking and Lord (1983), would be more sensitive to the impact of the prior distributions on the item dis-

criminations, and hence on the slope and intercept of the equating equations.

From an equating frame of reference, the PC-BILOG parameter estimates cannot be viewed in the same manner as LOGIST results. When transforming PC-BILOG item parameter estimates, the mean and standard deviation of the final adjusted quadrature distribution must be used as specified in Equations 11 and 12. When transforming LOGIST item parameter estimates, the mean and standard deviation of the distribution of $\hat{\theta}$s are used in Equations 4 and 5. Only the $\hat{\theta}$s of the PC-BILOG and LOGIST results can be treated in the same manner. In this situation, Equation 3 can be used for both LOGIST and PC-BILOG to transform the $\hat{\theta}$s, and Equation 6 can be used to transform the standard deviation of the PC-BILOG $\hat{\theta}$s. The difference in equating procedures rests fundamentally on the fact that the posterior θ distribution is not the same as the distribution of the $\hat{\theta}$s.

References

Andersen, E. B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31–44.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–472). Reading MA: Addison-Wesley.

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

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.

Kendall, M. G., & Stuart, A. (1967). *The advanced theory of statistics: Vol. 2.* New York: Hafner.

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

Mislevy, R. J. (1984.) Estimating latent distributions. *Psychometrika, 49,* 359–381.

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items.* Chicago: International Educational Services.

Mislevy, R. J., & Bock, R. D. (1985). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 189–202). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Mislevy, R. J., & Bock, R. D. (1987). *PC-BILOG I maximum likelihood item analysis and test scoring: Logistic model.* Mooresville IN: Scientific Software.

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). *A comparison of item and ability parameter estimates derived from LOGIST and BILOG.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago IL, U.S.A.

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

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

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116.

Yen, W. (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 Frank B. Baker, Department of Educational Psychology, 1025 W. Johnson Street, University of Wisconsin, Madison WI 53706, U.S.A.