Evaluation of the Efficiency of Item Calibration

C. David Vale and Kathleen A. Gialluca
Assessment Systems Corporation

This study compared several IRT calibration procedures to determine which procedure, if any, consistently produced the most accurate item parameter estimates. A new criterion of calibration efficiency was used for evaluating the calibration procedures; this criterion considers the joint effects of individual item parameter errors as they relate to the accuracy of θ estimation. Four methods of item calibration were evaluated: (1) heuristic estimates obtained from transformations of traditional item statistics; (2) ANCILLES, a program that first fits the *c* parameter and then transforms traditional item statistics to IRT *a* and *b* parameters; (3) LOGIST, a joint maximum likelihood procedure; and (4) ASCAL, a modification of LOGIST's algorithm which applies Bayesian priors to the abilities and item parameters. These were compared with each other and with a constant item parameter baseline condition. ASCAL and LOGIST produced estimates of essentially equivalent accuracy, although ASCAL's estimates of the *c* parameters were slightly superior. The heuristic estimates and those from ANCILLES were generally poor in comparison, particularly for smaller sample sizes. *Index terms: Calibration efficiency, Item calibration, Item parameter estimation, Item response theory, Latent trait models.*

Item response theory (IRT; Lord, 1980; Lord & Novick, 1968) encompasses a family of models that describe test items by their item characteristic curves (ICCs), or item response functions. The three-parameter logistic model was developed for use with dichotomously scored multiple-choice items. The item response function in this model expresses the probability of a correct response as a function of examinee ability or trait level (θ) and three item parameters, *a*, *b*, and *c*. The item response function is an S-shaped curve going from a lower asymptote equal to the *c* parameter (the pseudo-guessing parameter) to a maximum value of 1.0. The projection of the midpoint of this curve onto the θ scale provides a value for *b*, the item difficulty parameter. The slope, or rate at which the probability increases as a function of θ, is a function of the *a* parameter (the discrimination parameter). The three-parameter logistic IRT model is

$$P_g(\theta_i) = P(U_{gi} = 1 | a_g, b_g, c_g, \theta_i) = c_g + (1 - c_g)\Psi[Da_g(\theta_i - b_g)] \quad , \tag{1}$$

where

$$U_{gi} = \begin{cases} 1 \text{ if examinee } i \text{ answered item } g \text{ correctly} \\ 0 \text{ otherwise;} \end{cases}$$

$$\Psi(x) = 1/[1 + \exp(-x)],$$

$$D = 1.7,$$

and the subscripts *i* and *g* index the examinee and the item, respectively.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 1, March 1988, pp. 53–67
© Copyright 1988 Applied Psychological Measurement Inc.
0146-6216/88/010053-15$2.00

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Item Calibration Procedures

One of the most difficult and important tasks in using IRT is estimating the item parameters, or *calibrating the items*. Good estimates of the parameters are required if IRT is to function as well in practice as the theory predicts.

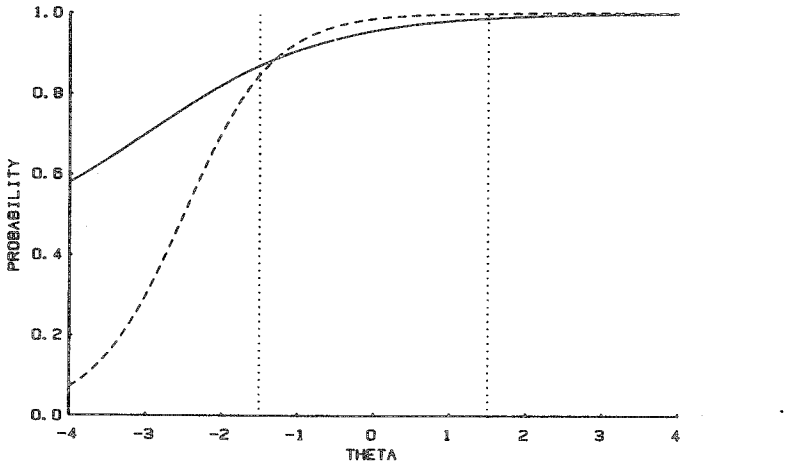
Conceptually, the process of parameter estimation amounts to fitting a curve of logistic shape through observed data points; the item parameters are the parameters describing that regression of the item score on θ. The parameters of the three-parameter IRT model are more difficult to estimate than are the parameters of some other popular IRT models (e.g., the Rasch model; Wright, 1977). The difficulty arises in the way in which the parameters interact to describe the ICC (Thissen & Wainer, 1982) and the fact that data points relevant to estimating important parameters are often scarce.

This difficulty is depicted graphically in Figure 1. When θ is a standard-score scale and the trait is normally distributed, approximately 87% of the population of examinees have θs between the two dashed vertical lines. The two items depicted in Figure 1 are very different in their characteristics. The item represented by the dashed curve is a highly discriminating easy item that cannot easily be answered correctly by guessing. The item represented by the solid curve is a poorly discriminating item that is very prone to correct responses through guessing. Despite the differences, the ICCs of these items are very similar in the θ range where most of the examinees lie. Differentiation between the two items depends on efficient use of the 6% of the examinees with θs below −1.5.

Because the estimation of item parameters is so difficult, there is no single accepted method for calibrating items. At least five theoretically distinct approaches to this problem have been used. Heuristic estimates are obtained by assuming some value for the *c* parameter (e.g., the reciprocal of the number of alternatives) and then transforming the classical item statistics to the IRT *a* and *b* parameters (Jensema, 1976). Alternatively, different values for the *c* parameters can be fit to the data when the classical item statistics are transformed, as above, to IRT *a* and *b* parameters. The best set of parameters is then the one that minimizes the lack of fit to the data (Croll & Urry, 1978).

Other estimates can be obtained by iteratively estimating the item parameters and examinee θs at successive stages by the method of maximum likelihood. At each stage, θs are assumed to be known

Figure 1
Item Characteristic Curves for Two Different Items



Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

when the item parameters are estimated, and the parameters are assumed to be known when the θs are estimated (Wingersky, Barton, & Lord, 1982). Bayesian priors can also be applied to the θs and the item parameters at each stage when this iterative process is performed (Swaminathan & Gifford, 1985; Vale & Gialluca, 1985).

Finally, the parameters can be estimated using a marginal maximum likelihood procedure. This procedure (Bock & Aitkin, 1981) assumes some distribution of θ and then integrates over it, leaving only the item parameters to be estimated.

Evaluating Item Calibration Procedures

Studies investigating the effectiveness of item calibration are almost universally conducted using monte carlo simulation techniques (see, e.g., Ree, 1979; Vale & Weiss, 1975). In such a simulation, responses to items with known parameters are generated according to a statistical model. These responses are then used as though they had been given by real examinees. Item parameters can be estimated from the item responses, as can θs. A major distinction between a simulation study and one based on responses from real examinees is that in a simulation, the true θs and item parameters are known.

In simulation studies of item calibration techniques, the estimated parameters are compared to their true values. In the past, indices of parameter error have been computed separately for the *a*, *b*, and *c* parameters. Typical indices of comparison have been the average absolute or squared difference and the correlation between the true and estimated parameters.

However, as was illustrated in Figure 1, errors in the estimates of the parameters of an item can compensate for one another when observed data are fit. The degree to which θ estimation also benefits from this compensation is important if interest is centered on the estimates of examinees' θs rather than the item parameters per se. Separate evaluation cannot assess these effects, but joint evaluation can. Some researchers (e.g., Ree, 1979) have attempted to evaluate the parameters jointly by comparing the test scores and information functions produced by the estimated parameters with those produced by the true parameters. The obtained ICC could also be compared with the true ICC. Both of these techniques, however, miss the more important evaluation of how well the errant parameters allow θ to be estimated.

Previous Investigations of Item Calibration Error

Although parameter estimation is a very important aspect of IRT applications, relatively few studies have evaluated the different procedures for estimating item parameters. Moreover, even fewer of these studies compared the different procedures directly and offered guidelines and suggestions for selecting one procedure over another.

Evaluations of Individual Calibration Procedures

Lord (1975) evaluated the joint maximum likelihood procedure of the computer program LOGIST in a simulation study. For this study, item parameters for 90 verbal items of the Scholastic Aptitude Test (SAT) were estimated by LOGIST using a sample of 2,995 examinees. After correction for errors of estimate, these parameters were used as the true parameters for a monte carlo simulation in which 2,995 simulated examinees (with θs identical to those of the real examinees) "responded" to the items according to the logistic test model. These responses were then used by LOGIST to reestimate the item parameters.

Root mean squared errors (RMSEs) of estimation and the correlations between true and estimated parameters were, respectively, .130 and .920 for the *a* parameters, and .196 and .988 for the *b* parameters. For the *c* parameters, the RMSE was .070; the correlation between the true and estimated *c* parameters was not reported.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Gugel, Schmidt, and Urry (1976) reported a similar simulation study of the minimum chi-square procedure. They specified a standard normal distribution of θ and rectangular distributions of item parameters; examinee sample sizes and test lengths were systematically varied.

Of all the conditions investigated, the condition with 90 items and 2,000 examinees was most comparable to Lord's study of LOGIST. In this condition, RMSEs and correlations were, respectively, .244 and .871 for the *a* parameter, .149 and .996 for the *b* parameter, and .069 and .568 for the *c* parameter. Direct comparisons with Lord's study are not particularly meaningful, however, because the distributions of all parameters were different; this can drastically affect the comparative indices. The study noted, however, that the minimum chi-square procedure did not work well when the number of examinees was as low as 500.

Schmidt and Gugel (1976) again reported the preceding study, as well as a second study in which 100 items and sample sizes of 2,000 and 3,000 were used. In this study, little difference was apparent between the two sample sizes. The results of these two studies led Schmidt and Gugel to conclude that, as a rule of thumb, item sets should contain at least 100 items and should be administered to at least 2,000 examinees to obtain an accurate calibration.

Comparisons of Different Calibration Procedures

Ree (1979) compared four calibration techniques in three different populations. The four calibration techniques were: (1) OGIVIA, minimum chi-square estimation (Urry, 1977); (2) ANCILLES, minimum chi-square estimation with ancillary correction for errors in estimation of θ (Urry, 1978); (3) LOGIST, version 4 (Wood, Wingersky, & Lord, 1976); and (4) Jensema's (1976) transformation procedure. Ree simulated three different θ distributions: rectangular, standard normal, and selected. The hypothetical items used in the simulation had parameters distributed normally in ranges typically found in real item sets. All analyses were performed on samples of 2,000 examinees and tests of 80 items.

Overall, the best estimates of the item parameters were obtained using LOGIST and a rectangular distribution of θ. LOGIST generally produced the highest correlations between estimated and true item parameters. Correlations between true number-correct scores (i.e., the sum of the ICCs) computed from true and estimated item parameters showed very little difference among methods and only a small deviation from unity. All calibration methods, except the transformations, produced information curves similar to the true information curve in the rectangular and normal θ distributions; in the selected distribution, all methods produced noticeable departures from the true information curve. Ree noted that OGIVIA required less than one-tenth the computer time required by LOGIST; however, both OGIVIA and ANCILLES deleted many more items from the estimation process than did LOGIST.

The second study comparing various calibration procedures was done by Swaminathan and Gifford (1983). They compared ANCILLES and LOGIST (version 4) in simulation at test lengths of 10, 15, 20, and 80 items and sample sizes of $N = 50$, 200, and 1,000. Three distributions of θ were used: standard normal, uniform, and negatively skewed beta.

A trend toward higher correlations between true and estimated parameters with increased test length was observed, and median correlations for LOGIST were slightly higher than those for ANCILLES. Differences between ANCILLES and LOGIST diminished with increasing sample size and test length. No substantial differences were observed among the θ distributions. Swaminathan and Gifford concluded that although LOGIST produced slightly better parameter estimates than did ANCILLES, it cost considerably more to run and the gain was probably not worth the cost. They also noted that ANCILLES deleted more items and examinees during the estimation process than did LOGIST. They further concluded that *a* and *c* parameters should not be estimated using tests containing 15 or fewer items.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Method

This study was undertaken to compare some of the existing IRT calibration procedures to determine which procedure, if any, consistently produced the most accurate item parameter estimates across different conditions of calibration. A new criterion of calibration efficiency was used for evaluating the calibration procedures; this criterion considers the joint effects of item parameter error as it relates to the accuracy of θ estimation. This efficiency criterion was used in addition to the more traditional indices of individual item parameter error.

Calibration Procedures

As a baseline, constant parameters of $a = 1.0$, $b = 0.0$, and $c = .20$ or .25 were used. The effect on calibration efficiency of using constant parameters is identical to the effect of simply summing the number of correct responses; for evaluation of information, this is also the same as using the one-parameter logistic (or Rasch) model to estimate item parameters. Thus, in terms of efficiency, the constant-parameter condition is equivalent to Rasch calibration.

Four methods of item calibration were evaluated:

1. Heuristic estimates obtained from transformations of traditional item parameters (Jensema, 1976).
2. ANCILLES-X, a version of ANCILLES (Croll & Urry, 1978).[1] The ANCILLES-X parameters were included as a low-cost calibration option, as recommended by Swaminathan and Gifford.
3. LOGIST, version 5 (Wingersky et al., 1982). LOGIST was included as the high-cost state-of-the-art calibration option.
4. ASCAL (Vale & Gialluca, 1985), a microcomputer-based calibration program whose procedures are modeled after those in LOGIST. ASCAL differs from LOGIST in that Bayesian prior distributions are added to the pseudo-likelihood functions (i.e., likelihood functions modified for omitted item responses; see Lord, 1974) for the θ estimates and the a and c parameters. No prior distribution is used for the b parameters (although the b-parameter estimates are bounded by ± 3.0). The prior distribution of θ is a standard normal distribution; symmetric beta distributions are used as the specified Bayesian priors for the a and c parameters.

Simulated Item Responses

Three sets of true parameters were obtained by applying ASCAL to real datasets. The first set was obtained from a 25-item test containing general science items. To obtain an effective test length of 50 items, each item and its parameters were included in the test twice. All items in Test 1 had four alternatives. Test 1 had a restricted range of item difficulty, typical of a conventional test but inappropriate for adaptive testing. To create a test more appropriate for adaptive testing, Test 2 was created by multiplying all of the b parameters in Test 1 by 2.0. Tests 1 and 2 were therefore identical except for the b parameters.

Test 3 was modeled after a 57-item test of shop knowledge. Each item in Test 3 had five alternatives. Because Test 3 was developed as part of an adaptive item pool, it had a slightly wider range of difficulty than did Test 1, although its difficulty range was not as wide as that of Test 2. Test 3 was also more difficult than the other tests, and its items were less discriminating. Table 1 presents the means and standard deviations of the true parameters used as models for the simulations.

---

[1] ANCILLES-X was prepared by J. B. Sympson of the Navy Personnel Research and Development Center.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 1
Means and Standard Deviations of True Item Parameters for Tests 1, 2, and 3

	Test 1		Test 2		Test 3	
Parameter	Mean	S.D.	Mean	S.D.	Mean	S.D.
a	1.492	.410	1.492	.410	1.041	.417
b	-.090	.910	-.179	1.820	.675	1.069
c	.230	.082	.230	.082	.198	.028

Item response data were generated for 2,000 examinees for each of the three tests. Examinee θ levels were sampled from a standard normal distribution. The first 500 and the first 1,000 examinees were drawn from each of these samples to form smaller samples. For reasons of economy, LOGIST was run only for $N = 2{,}000$; the other calibration programs were run for all three sample sizes.

Evaluative Criteria

Three different criteria were used to evaluate the accuracy of the item calibration programs. The RMSE, which is the square root of the average squared difference between the true and estimated parameters, was computed for each of the three parameters (*a*, *b*, and *c*) in each of the three tests. Similarly, the Pearson product-moment correlation coefficient between the estimated and the true parameters was computed for each parameter in each test. The third criterion was a calibration efficiency criterion.

Recall that the aim of this study was to evaluate the effect of item calibration error on the θ estimates. Any differences between the true and estimated θ reflect two sources of error: (1) calibration error, the error due to fallible estimates of the item parameters, and (2) measurement error, the error resulting from administering a finite-length test. Consequently, any direct evaluation of error in θ estimates (e.g., RMSE or correlations between true and estimated θ) confounds these two sources of error. The calibration efficiency criterion represents an attempt to isolate the effects of item calibration error by determining the effect on θ of administering the test infinitely many times.

The efficiency criterion is a relative-information criterion suggested by Vale, Maurelli, Gialluca, Weiss, and Ree (1981). Their procedure computes the amount of psychometric information (Birnbaum, 1968) that would be extracted from the items if they were scored using the estimated or errant parameters; the relative efficiency of the estimated parameters (i.e., the ratio of the information in the estimated parameters to the information in the true parameters) can be determined for each θ level by comparing the errant information with the true information.

θ-Γ *transformation*. The θ metric can be defined as the criterion metric along which the true parameters are anchored and along which the response probabilities are accurately described by the IRT model incorporating the θ level and the item parameters. A second metric, called Γ, is produced by scoring item responses using parameters other than the true parameters of the θ metric (i.e., by using the errant or estimated parameters). Conceptually, the Γ level corresponding to a given θ level could be determined from administering a test scored using the errant parameters an infinite number of times and scoring all the responses simultaneously. Thus, asymptotically there is a mathematical, as opposed to a statistical, relationship between Γ and θ.

It is, of course, impossible to administer an infinite-length test or to repeat a finite-length test an infinite number of times. However, the θ-Γ transformation can be determined by more practical means.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

The maximum likelihood estimate of θ, which is asymptotically unbiased, can be obtained by finding the root in θ of the following likelihood equation given by Birnbaum (1968, p. 459):

$$\sum_{g=1}^{n} a_g \Psi[Da_g(\hat{\theta} - b_g)] - \sum_{g=1}^{n} \frac{w_g(\hat{\theta}) u_g}{D} = 0 \quad , \tag{2}$$

where

$$w_g(\hat{\theta}) = Da_g \Psi[Da_g(\hat{\theta} - b_g) - \ln(c_g)] \tag{3}$$

are *locally best* weights as defined by Birnbaum (1968, pp. 442–444), $\hat{\theta}$ is an estimate of θ, and all other terms are as defined earlier, except there is no subscript indexing each examinee. If each item were repeated r times, Equation 2 could be written as

$$\sum_{g=1}^{n} \sum_{h=1}^{r} a_g \Psi[Da_g(\hat{\theta} - b_g)] - \sum_{g=1}^{n} \sum_{h=1}^{r} \frac{w_g(\hat{\theta}) u_{gh}}{D} = 0 \tag{4}$$

or

$$r \sum_{g=1}^{n} a_g \Psi[Da_g(\hat{\theta} - b_g)] - \sum_{g=1}^{n} \frac{w_g(\hat{\theta})}{D} \sum_{h=1}^{r} u_{gh} = 0 \tag{5}$$

or

$$\sum_{g=1}^{n} a_g \Psi[Da_g(\hat{\theta} - b_g)] - \sum_{g=1}^{n} \frac{w_g(\hat{\theta})}{D} P_g = 0 \quad , \tag{6}$$

where P_g is the observed proportion of correct responses to item g in r repetitions. Equation 6 is simply Equation 2 with P_g in place of u_g.

If the three-parameter model holds and true parameters are available, P_g can be computed using Equation 1. When this P_g is substituted into Equation 6 along with true item parameters, the root of the equation is found at $\hat{\theta} = \theta$.

In a simulation, it is possible to control when the true parameters are available and when estimates must be used. Let P_g be construed as the probability with which an examinee will respond to item g with a correct answer. The probability of an examinee's response being correct is governed by his or her true θ and the true item parameters. Thus, P_g should be computed using θ and a_g, b_g, and c_g. When θ is estimated (i.e., in a real-world environment), the estimated parameters must be used. If the parameters are in error, the resulting estimate ($\hat{\theta}$) will not converge on θ but rather on Γ. The value of Γ corresponding to a given θ can be determined by substituting the true P_g into Equation 6 and finding the root using the errant parameters. Thus, using θ to denote the true trait level; Γ to denote the asymptotic value obtainable with errant parameters; a_g, b_g, and c_g to denote the true parameters; and $\hat{a}_g$, $\hat{b}_g$, and $\hat{c}_g$ to denote the estimated parameters, Equation 6 can be rewritten as

$$\sum_{g=1}^{n} \hat{a}_g \Psi[D\hat{a}_g(\Gamma - \hat{b}_g)] - \sum_{g=1}^{n} \left\{ \frac{\hat{w}_g(\Gamma)}{D} [c_g + (1 - c_g)] \Psi[Da_g(\theta - b_g)] \right\} = 0 \quad , \tag{7}$$

where

$$\hat{w}_g(\Gamma) = D\hat{a}_g \Psi[D\hat{a}_g(\Gamma - \hat{b}_g) - \ln(\hat{c}_g)] \quad . \tag{8}$$

If the errors of calibration are 0 or the estimated parameters are consistent with the true parameters, the transformation of θ to Γ will be linear. When this is not the case, as in almost all real calibration situations, the transformation will be nonlinear. This transformation from θ to Γ completely describes the asymptotic effect of item parameter error on θ estimation; that is, it describes the effect of item parameter error on θ estimation for tests of infinite length.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Efficiency. The information at θ for a specific test score (or scoring function), X, can be expressed as the ratio of the squared derivative of the expected value of the scoring function to the variance of the scoring function at θ:

$$I(\theta) = I(\theta;X) = \frac{\left[\frac{\partial}{\partial\theta} \mathrm{E}(X|\theta)\right]^2}{\sigma^2_{X|\theta}} \tag{9}$$

(Birnbaum, 1968, Equation 20.1.1). When the score is a linear combination of 0-1 item responses (with each item response weighted w_g), the components of the information equation can be written as

$$\frac{\partial}{\partial\theta} \mathrm{E}(X|\theta) = \sum_{g=1}^{n} \frac{\partial}{\partial\theta} w_g \mathrm{E}(U_g|\theta) = \sum_{g=1}^{n} \frac{\partial}{\partial\theta} w_g P_g(\theta) = \sum_{g=1}^{n} w_g P'_g(\theta) \tag{10}$$

and

$$\sigma^2_{X|\theta} = \sum_{g=1}^{n} w_g^2 P_g(\theta)\,[1-P_g(\theta)] \tag{11}$$

where $w_g = w_g(\theta)$ is defined as in Equation 3, and

$$P'_g(\theta) = (1-c_g)Da_g\Psi[Da_g(\theta-b_g)] \quad . \tag{12}$$

Note that the weight-based formulation of information is generally appropriate only if the scoring function is independent of θ. Because the weights applied here are functions of θ, this is not strictly true. But if θ is fixed at any chosen value, then the scoring function is independent of θ, and in particular it is the locally best scoring function at that point (see Birnbaum, 1968, p. 456). In other words, the fact that the optimal weights are a function of θ is incidental to their usability in a scoring formula. The scoring formula can be applied across the range of θ; however, only at the one point on the θ continuum will the weights be optimal. Thus, the weight w_g is not considered a function of θ while taking the derivative in Equation 10. However, the optimal weights to apply at any point can still be determined from θ.

The information available from scoring response vectors using errant parameters can be viewed as equivalent to the information available in a linear combination of item responses using those weights determined to be locally best at Γ, using the estimates of the item parameters. Thus, substituting Equations 1 and 12 into Equation 9, and using the errant weights from Equation 8, the information available from the errant parameters is given by

$$I(\theta;\hat{a},\hat{b},\hat{c}) = \frac{\left[\sum_{g=1}^{n} \hat{w}_g(\Gamma)P'_g(\theta)\right]^2}{\sum_{g=1}^{n} \hat{w}_g^2(\Gamma)P_g(\theta)[1-P_g(\theta)]} \quad . \tag{13}$$

This is similar to Birnbaum's (1968) Equation 20.2.2. Equation 13 represents the information contained in the errant parameters as a function of θ. To produce a single-quantity estimate of the information available, this function may be weighted by a standard normal density function and integrated over θ (i.e., numerically integrated).

To provide a relative efficiency index, the information thus obtained may be compared to information available from the true parameters. This information may be computed in the same manner using true parameters throughout, or it may be computed using any of the formulas provided by Birnbaum (1968).

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Efficiencies for this study were computed in the manner described above. Efficiency, as reported herein, refers to the ratio of errant to true information.

Results

RMSE

The RMSEs between the true and estimated item parameters are presented in Table 2 for each of the calibration procedures. According to this criterion, the *a* parameters were generally less well estimated than were either the *b* or the *c* parameters, particularly for Tests 1 and 2. Test 2, with its exaggerated range of true *b* parameters, yielded parameter estimates with larger RMSEs than any other test; this was true for all three item parameters. With few exceptions, the constant parameters typically yielded the largest RMSEs.

ASCAL typically produced better estimates of the *a* parameter (according to this criterion) than did the other calibration procedures; the only exception to this finding occurred for Test 3 and the largest sample size, where the RMSE was .150 for LOGIST and .161 for ASCAL. The differences between ASCAL and the other calibration procedures were greater for $N = 500$ and 1,000 than for $N = 2{,}000$. ANCILLES-X was consistently worse than either ASCAL or LOGIST; for Tests 1 and 2 and $N = 500$, *a* parameter estimates from ANCILLES-X were also worse than the heuristic estimates and the constants. Typically, the RMSEs from the heuristic estimates were intermediate between the values for the constants and those for ANCILLES-X.

ASCAL consistently produced estimates of the *b* parameters with a lower RMSE than did the other calibration procedures; again, differences between ASCAL and the other procedures were more marked at the smaller sample sizes. Of the remaining procedures (and at the largest sample size, where data from LOGIST were available), ANCILLES-X was best for Test 1, LOGIST was best for Test 2, and both LOGIST and the heuristic estimates were best for Test 3. The heuristic estimates outperformed ANCILLES-X at the smaller sample sizes for Tests 1 and 2 and in all cases for Test 3.

For Tests 1 and 2, ASCAL produced *c* estimates with a lower RMSE than did the other procedures; these differences among the calibration procedures were larger for Test 1 than they were for Test 2. For these two tests, the constant *c* values (which are identical to the heuristic *c* estimates and equal to the reciprocal of the number of alternatives) were at least as good as the estimates produced by ANCILLES-X and LOGIST. None of the calibration procedures produced *c* estimates for Test 3 that had RMSE less than that of the constant values; of the remaining procedures, ASCAL was best.

Correlations

Table 3 presents the Pearson product-moment correlations between the true and estimated item parameters for the calibration procedures. Correlations could not be computed for any of the constant parameters because they did not vary from item to item; for the same reason, correlations could not be computed for the heuristic estimates of the *c* parameters.

The correlations between the true and estimated parameters were higher for *b* than they were for either *a* or *c*; this was true for all three tests. The *a* parameters were best estimated (according to this criterion) in Test 3; Test 2, with its wide range of *b* values, had poorly estimated *a* parameters. ASCAL produced consistently higher *a* parameter correlations than did ANCILLES-X and the heuristic estimates at all sample sizes for Test 1; correlations for the heuristic estimates of the *a* parameters were uniformly low (.055 to .219).

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 2
Root Mean Squared Error Between True and Estimated Parameters for Tests 1, 2, and 3 and Examinee Sample Sizes of 500, 1000, and 2000

Test, Sample Size, & Parameter	Constant Parameters	Item Calibration Procedure			
		Heuristic Estimates	ANCILLES-X	LOGIST	ASCAL
Test 1					
500					
a	.641	.550	.782	--	.286
b	.914	.187	.209	--	.150
c	.084	.084	.099	--	.058
1000					
a	.641	.569	.372	--	.197
b	.914	.172	.132	--	.088
c	.084	.084	.087	--	.050
2000					
a	.641	.579	.428	.231	.132
b	.914	.194	.152	.182	.088
c	.084	.084	.084	.080	.049
Test 2					
500					
a	.641	.632	.839	--	.357
b	1.828	.403	.741	--	.290
c	.084	.084	.121	--	.079
1000					
a	.641	.624	.569	--	.383
b	1.828	.313	.331	--	.204
c	.084	.084	.106	--	.080
2000					
a	.641	.674	.540	.497	.386
b	1.828	.307	.287	.250	.189
c	.084	.084	.091	.090	.079
Test 3					
500					
a	.419	.349	.290	--	.252
b	1.264	.235	.396	--	.189
c	.028	.028	.106	--	.039
1000					
a	.419	.319	.262	--	.171
b	1.264	.216	.380	--	.169
c	.028	.028	.114	--	.043
2000					
a	.419	.318	.208	.150	.161
b	1.264	.184	.253	.183	.176
c	.028	.028	.096	.066	.055

For Test 2, the differences among the calibration procedures were less marked, except for the heuristic estimates, which were again uniformly poor (.035 to .173). ASCAL outperformed ANCILLES-X at the smallest sample size and at the largest sample size; ANCILLES-X was slightly better for $N = 1{,}000$. LOGIST produced *a* parameters that were essentially as good as ASCAL's for the largest sample size. Differences among the calibration procedures were even smaller for Test 3, and for this test the heuristic estimates

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 3
Product-Moment Correlations Between True and Estimated Parameters for Tests 1, 2, and 3 and Examinee Sample Sizes of 500, 1000, and 2000

Test, Sample Size, & Parameter	Item Calibration Procedure			
	Heuristic Estimates	ANCILLES-X	LOGIST	ASCAL
Test 1				
500				
a	.219	.587	--	.757
b	.984	.985	--	.992
c	--	.747	--	.732
1000				
a	.136	.805	--	.881
b	.984	.993	--	.996
c	--	.761	--	.809
2000				
a	.055	.747	.848	.952
b	.978	.992	.986	.996
c	--	.806	.612	.828
Test 2				
500				
a	.173	.337	--	.570
b	.982	.951	--	.988
c	--	.654	--	.577
1000				
a	.159	.570	--	.560
b	.991	.987	--	.994
c	--	.720	--	.612
2000				
a	.035	.563	.596	.598
b	.990	.989	.991	.995
c	--	.725	.498	.620
Test 3				
500				
a	.757	.818	--	.810
b	.977	.965	--	.990
c	--	.216	--	.546
1000				
a	.814	.874	--	.914
b	.981	.975	--	.992
c	--	.185	--	.530
2000				
a	.834	.908	.934	.926
b	.985	.987	.990	.993
c	--	.271	.377	.463

did not perform as poorly as they did on the other tests. ANCILLES-X was best at the smallest sample size; ASCAL was best for $N = 1{,}000$. At the largest sample size, LOGIST outperformed the other procedures.

ASCAL produced *b* parameters that were more highly correlated with their true values than those produced by any of the other calibration procedures; this was true for all tests and sample sizes, although none of the differences among the calibration procedures was large. The *b* parameter correlations from

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

ANCILLES-X were generally higher than those from the heuristic estimates for Test 1, but generally lower for Tests 2 and 3. For the largest sample size, LOGIST's correlations were lower than ASCAL's and higher than those for the other procedures. The *b* parameter correlations throughout Table 3 are all well above .950.

The *c* parameter estimates were more highly correlated with their true values for Test 1 than they were for any other test. For Test 1, ASCAL was better for $N = 1{,}000$ and 2,000, and ANCILLES-X was better at the smallest sample size. For Test 2, ANCILLES-X produced the best *c* estimates at all sample sizes, and LOGIST once again produced the worst. ASCAL produced markedly better *c* estimates for Test 3 than did the other procedures; the correlations from ANCILLES-X were low for this test (.185–.271).

Efficiency

The efficiency statistics are presented in Table 4 for all item calibration procedures and sample sizes. The efficiency of item calibration increased with sample size for all calibration procedures; this was not strictly true for the other criteria just discussed. The constant parameters, of course, were the same for all sample sizes for any one test, and therefore the efficiencies did not vary with sample size. The differences among the calibration procedures were largest for Test 1 and smallest for Test 3.

ASCAL resulted in the highest efficiency of item calibration in every instance except one (Test 3 with $N = 2{,}000$, where the efficiencies for LOGIST and ASCAL were .986 and .984, respectively). For Tests 1 and 2, LOGIST was the next best calibration method. ANCILLES-X produced better parameters than the heuristic procedure for every situation, except for Test 2 at the smallest sample size. The efficiency statistics were lower for Test 2, with its exaggerated range of *b* parameters, than they were for either of the other two tests. Differences in the efficiencies between Tests 1 and 3 across calibration methods were neither large nor consistent.

Summary

Each of the three criteria used to evaluate item calibration error (i.e., RMSE, correlations, and efficiency) suggested the same conclusion: In most cases, ASCAL produced better item parameter estimates

Table 4
Efficiency Statistics for Tests 1, 2, and 3 for Examinee Sample Sizes of 500, 1000, and 2000

Test and Sample Size	Constant Parameters	Item Calibration Procedure: Heuristic Estimates	ANCILLES-X	LOGIST	ASCAL
Test 1					
500	.717	.928	.894	--	.972
1000	.717	.929	.962	--	.984
2000	.717	.933	.973	.981	.990
Test 2					
500	.499	.895	.876	--	.958
1000	.499	.922	.956	--	.968
2000	.499	.924	.965	.971	.975
Test 3					
500	.725	.920	.953	--	.964
1000	.725	.939	.967	--	.978
2000	.725	.945	.981	.986	.984

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

than did the other procedures investigated here. LOGIST invariably produced estimates of nearly equivalent quality, although its estimates of *c* were not as good. The heuristic parameter estimates and the estimates from ANCILLES-X were generally poor compared to ASCAL and LOGIST, particularly at the smaller sample sizes.

Test 2 had an atypically wide range of *b* parameters; that is, its *b* parameters were distributed more widely than would be found on conventional and many adaptive tests. This had a dramatic effect on the efficiency of item calibration: The efficiency statistics for Test 2 were lower than those for either of the other two tests. The same phenomenon was observed upon examination of the calibration errors in the individual parameters: Estimates of all three item parameters (not just the estimates of the *b*s) were notably worse for Test 2.

Discussion and Conclusions

Accuracy of Item Calibration

The results of this evaluation suggest that ASCAL produces parameter estimates that are, in general, at least as accurate as those produced by LOGIST and more accurate than those produced by ANCILLES-X and the heuristic procedures. The difference between ASCAL and the latter two procedures was most noticeable at the smaller sample sizes employed in this study (i.e., $N = 500$ and 1,000); LOGIST calibrations were available only for $N = 2{,}000$, and differences between LOGIST and ASCAL at this sample size were small. A word of caution is in order here, however. The fact that the true parameters used to generate the simulated item responses were obtained in previous calibrations using ASCAL may have biased the results in favor of ASCAL. This would be true to the extent that ASCAL has a favored range of estimates that is different from the range of estimates that LOGIST favors.

The conventional wisdom that dictates the use of the one-parameter model or conventional scoring with small datasets may not be warranted. The relative efficiency of the worst-case calibration procedure was still markedly higher than the efficiency obtained by unit-weighting the item responses (the mathematical equivalent of the Rasch model in information comparisons).

The conclusions reached in this study are generalizable to real datasets with examinee and item characteristics similar to the ones investigated here. The differences among the calibration procedures were greater at the smaller sample sizes; this was also observed by Swaminathan and Gifford (1983). Thus, there is no reason to expect the relative rankings of the calibration procedures to be different with groups of fewer than 500 examinees.

Test length was not systematically manipulated in this study. Consequently, it is not known to what degree these conclusions remain valid for other test lengths. At best, differences among the procedures would become smaller at longer test lengths as all methods improve (see Swaminathan & Gifford, 1983). At shorter test lengths, differences among the calibration procedures would probably be larger than those observed here (see the recommendation of Schmidt & Gugel, 1976, that ANCILLES be used with tests containing 100 items or more, and the recommendation of Wood et al., 1976, that LOGIST be used with tests containing at least 40 items).

The true θs in this study were assumed to be distributed standard normal; no other θ distributions were simulated. It is worth noting that ASCAL assumes a standard normal θ distribution and specifies its Bayesian prior accordingly; it may be reasonable to expect the relative efficiency of ASCAL to drop in the context of another θ distribution. However, it should be noted that ANCILLES and the heuristic estimates also assume an underlying normal distribution of θ.

The real issue here, however, is not whether it is possible to define a θ distribution for which one calibration procedure consistently yields better parameter estimates than another. What is important is how well the various procedures estimate item parameters for examinees in the population of interest. In

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

many cases, it is reasonable to assume that examinee θs are distributed in a way that can be approximated by a normal distribution.

Other Considerations

It appears that consideration of accuracy alone would suggest the use of ASCAL over the other calibration procedures. However, other issues should also be weighed when selecting an item calibration procedure. Convenience and cost would both suggest the selection of either (1) the easily programmed heuristic procedures, or (2) ASCAL, which runs simply (with no user-supplied options) and economically on a personal computer. ANCILLES-X is fairly easy to use but requires a mainframe computer. LOGIST runs on a mainframe computer and its input requirements are extensive, unless the user employs default options. Although these extensive options make LOGIST the most difficult calibration program to use, they also make it the most flexible. If this flexibility is required, LOGIST may be the program of choice.

References

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. 397–479). Reading MA: Addison-Wesley.

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

Croll, P. R., & Urry, V. W. (1978). *ANCILLES: A program for estimation of the item parameters of normal ogive and logistic mental test models—Version 78.5.* Washington DC: US Civil Service Commission, Personnel Research and Development Center.

Gugel, J. F., Schmidt, F. L., & Urry, V. W. (1976). Effectiveness of the ancillary estimation procedure. In C. L. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing* (PS-75-6). Washington DC: US Civil Service Commission, Personnel Research and Development Center.

Jensema, C. J. (1976). A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement, 36,* 705–715.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247–264.

Lord, F. M. (1975). *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters* (Research Bulletin 75-33). Princeton NJ: Educational Testing Service.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Ree, M. J. (1979). Estimating item characteristic curves. *Applied Psychological Measurement, 3,* 371–385.

Schmidt, F. L., & Gugel, J. F. (1976). The Urry item parameter estimation technique: How effective? In W. A. Gorham (Chair), *Computers and testing: Steps toward the inevitable conquest* (PS-76-1). Washington DC: US Civil Service Commission, Personnel Research and Development Center.

Swaminathan, H., & Gifford, J. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 13–30). New York: Academic Press.

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

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397–412.

Urry, V. (1977). *OGIVIA: Item parameter estimation program with normal ogive and logistic three-parameter model options.* Washington DC: US Civil Service Commission, Personnel Research and Development Center.

Urry, V. (1978). *ANCILLES: Item parameter estimation program with normal ogive and logistic three-parameter model options.* Washington DC: US Civil Service Commission, Personnel Research and Development Center.

Vale, C. D., & Gialluca, K. A. (1985). *ASCAL: A microcomputer program for estimating logistic IRT item parameters* (ONR-85-4). St. Paul: Assessment Systems Corporation.

Vale, C. D., Maurelli, V. A., Gialluca, K. A., Weiss, D. J., & Ree, M. J. (1981). *Methods for linking item parameters* (AFHRL-TR-81-10). Brooks Air Force Base TX: U.S. Air Force Human Resources Laboratory, Manpower and Personnel Division.

Vale, C. D., & Weiss, D. J. (1975). *A simulation study*

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

of stradaptive ability testing (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum RM-76-6). Princeton NJ: Educational Testing Service.

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

Acknowledgments

The authors thank J. Stephen Prestwood and James B. Sympson for their helpful suggestions in the development of ASCAL, and Frederic Lord, Malcolm Ree, Martha Stocking, and Marilyn Wingersky for their assistance in the comparisons with LOGIST.

Author's Address

Send requests for reprints or further information to C. David Vale, Assessment Systems Corporation, 2233 University Avenue, Suite 440, St. Paul MN 55114, U.S.A.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/