An Investigation of the Sampling Distributions of Equating Coefficients

Frank B. Baker

University of Wisconsin, Madison

Using the characteristic curve method for dichotomously scored test items, the sampling distributions of equating coefficients were examined. Simulated data for broad-range and screening tests were analyzed using three equating contexts and three anchor-item configurations in horizontal and vertical equating situations. The results indicated that the sampling distributions were bell-shaped and their standard deviations were uniformly small. There were few differences in the forms of the distributions of the obtained equating coefficients as a function of the anchor-item configurations or type of test. For the equating contexts studied, the sampling distributions of the equating coefficients appear to have acceptable characteristics, suggesting confidence in the values obtained by the characteristic curve method. *Index terms: anchor items, characteristic curve method, common metric, equating coefficients, sampling distributions, test equating.*

In order to make valid comparisons among two or more tests measuring a given attribute, the results must be expressed in a common metric. This can be accomplished using one of two fundamentally different methods—horizontal and vertical equating. Using horizontal equating, tests measuring a common trait at the same level for groups of examinees from the same population are placed on a common metric. Using vertical equating, tests measuring the same trait are administered to groups of examinees at different levels and placed on a common trait scale (Slinde & Linn, 1977). In these two situations, the item response theory (IRT) test equating problem is one of finding a linear transformation that expresses the results in a common metric. The basic transformation equation is

$$\theta_j^* = A\theta_j + K\,, \tag{1}$$

where

the slope, A, and the intercept, K, are the equating coefficients,

θ_j is person j's trait measure in the metric of the current test, and

θ_j^* is θ_j expressed in the target metric, which can be that of a second test or an arbitrary metric.

The values of the current test's item parameters, a_i (discrimination) and b_i (difficulty), can be transformed into the target metric using

$$a^* = \frac{a_i}{A} \tag{2}$$

and

$$b_i^* = Ab_i + K\,, \tag{3}$$

where i = 1, 2, 3, ..., n items in the test. Note that under Birnbaum's three-parameter IRT model (Lord & Novick, 1968, pp. 404, Equation 17.3.1) used here c_i, the pseudoguessing parameter, is not transformed because its value is independent of the θ metric.

Under IRT, the equating coefficients can be determined using the characteristic curve method (Stocking & Lord, 1983), which uses the item response functions (IRFs) (characteristic curves) of the anchor items in the two tests. Let the probability of a correct response to a target test anchor item be denoted by $P_i(\theta_j)$. Let $P_i^*(\theta_j)$ be that for the same anchor item in the current test after transformation of its item parameter estimates to the target metric. A perfect equating implies that $[P_i(\theta_j) - P_i^*(\theta_j)] = 0$ over the range of the target metric scale for all anchor items in the two tests. Thus, the fundamental task is one of finding

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 1, March 1996, pp. 45–57
© Copyright 1996 Applied Psychological Measurement Inc.
0146-6216/96/010045-13$1.90

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 values of the equating coefficients A and K that meet this criterion as closely as possible.

As originally developed by Haebara (1980) for dichotomously scored items, the characteristic curve method yielded equating coefficients that were symmetric with respect to which member of the pair of tests was designated as the target test. Stocking & Lord (1983) reformulated the approach in terms of the test characteristic curves [or test response functions (TRFs)] of the two tests. However, their procedure yields equating coefficients that are not symmetric with respect to designation of the target test. Stocking & Lord's (1983) method was used here to obtain the values of the equating coefficients. Although this procedure provides a ready means of equating tests, the sampling properties of the obtained equating coefficients have not been investigated.

In this study, the empirical sampling distributions of these coefficients were examined for two types of tests, three kinds of equating contexts, and various anchor-item configurations in both horizontal and vertical equating situations. In addition, it was of interest to investigate whether the underlying values of the equating coefficients were recovered. The characteristics of these sampling distributions were used to help determine whether the obtained equating coefficients were "well-behaved" in the various contexts examined. If the obtained sampling distributions were bell-shaped and symmetric with small variances and few or no outliers or other abnormalities, they were considered well-behaved. When this holds, such sampling distributions can be used to aid in the interpretation of the values of a pair of equating coefficients obtained in practice. To facilitate such an examination, a simulation approach was used because it affords a ready means of implementing the various contexts of interest.

The Characteristic Curve Method

For a three-parameter IRT model,

$$P_i(\theta_j) = c_i + (1-c_i)\frac{1}{1+\exp\left[-a_i'(\theta_j - b_i')\right]} \tag{4}$$

and

$$P_i{*}(\theta_j) = c_i + (1-c_i)\frac{1}{1+\exp\left[-a_i{*}(\theta_j - b_i{*})\right]}\ , \tag{5}$$

where

a_i' and b_i' are the anchor item's parameter estimates from the calibration of the target test,

$a_i{*}$ and $b_i{*}$ are the result of applying Equations 2 and 3 to the anchor item parameter estimates yielded by the calibration of the current test, and

θ_j is the trait score of person j expressed in the common metric.

Starting from the basic difference of interest, $[P_i(\theta_j) - P_i{*}(\theta_j)]$, a quadratic loss function for the dichotomous response case can be defined as

$$F = \frac{1}{N}\sum_{g=1}^{N}\left\{\sum_{i=1}^{m}\left[P_i(\theta_g) - P_i{*}(\theta_g)\right]\right\}^2\ , \tag{6}$$

where $g = 1, 2, ..., N$ indexes the N arbitrary points used over the trait scale and $i = 1, 2, ..., m$ indexes the anchor items common to the two tests.

Distributing the summation over items across the terms inside the brackets of Equation 6 yields:

$$T_j = \sum_{i=1}^{m} P_i(\theta_j) \tag{7}$$

and

$$T_j{*} = \sum_{i=1}^{m} P_i{*}(\theta_j)\ , \tag{8}$$

where T_j is the "true score" [expected number correct (ENC)] at point θ_j based on the anchor items in the target test, and $T_j{*}$ is the ENC at point θ_j based on the anchor items in the current test after transformation of the item parameter estimates using Equations 2 and 3.

The resultant quadratic loss function is that used by Stocking & Lord (1983):

$$F = \frac{1}{N}\sum_{j=1}^{N}\left(T_j - T_j{*}\right)^2 . \tag{9}$$

Because the goal is to express two sets of test results in the same metric, the θ_j appearing in both T_j and $T_j{*}$ must be in the common metric. This metric can be established as that of the target test; alterna-

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/

tively, an arbitrarily defined metric can be used (Baker, 1992; Stocking & Lord, 1983). The majority of IRT test analysis computer programs solve the identification problem by standardizing the distribution of the examinee's θ estimates ($\hat{\theta}$). Thus, a convenient metric is one defined to have a midpoint of 0 and a unit of measurement of 1. This metric does not depend on the mean and variance of the $\hat{\theta}$s of either the current or target tests.

The task then is to find the values of the equating coefficients that will minimize the quadratic loss function of Equation 9 in such a metric. Because F is a function of A and K it will be minimized when $\partial F/\partial A = 0$ and $\partial F/\partial K = 0$, but the resulting system of simultaneous equations does not have a closed form solution. Stocking & Lord (1983) used an iterative multivariate search technique due to Davidon (1959) and Fletcher & Powell (1963) to find the values of equating coefficients that will minimize F. This process has been implemented in the EQUATE computer program (Baker, Al-Karni, & Al-Dosary, 1991) using six subroutines for implementing the multivariate search procedure obtained from *Numerical Recipes* (Press, Flannery, Teukolsky, & Vetterling, 1988). This process yields the values of the slope (A) and intercept (K) of the simple linear regression line defined by Equation 1. The obtained values are those that yield the best possible match between the TRFs of the current and target tests when both are expressed in the common metric.

Method

Instrument Specifications

An item pool of 500 items (calibrated under the three-parameter logistic model; Baker, Cohen, & Barmish, 1988) was used in the creation of the sets of item parameters for the tests used in the generation of the simulated item response data. The summary statistics of the item pool were $\mu_a = 1.18$, $\sigma_a = .41$; $\mu_b = 0$, $\sigma_b = 1.0$; and $0 \leq c \leq .35$ with $\mu_c = .23$, $\sigma_c = .04$. The items for a given test were selected from the item pool using linear programming (Theunissen, 1985). Under this procedure, a test information function (TIF) is defined over a specified θ range for the desired test. The range used was −2.5 to +2.5 with the value of the TIF calculated for points along the θ scale at .5 units. The linear programming procedure selects items whose parameters will produce a TIF that matches the desired TIF as closely as possible. The item selection process was implemented using the BPLX87 linear programming computer program (Eastern Software Products, 1985).

Broad-range test. The first type of test created from the item pool was a broad-range test that is representative of many achievement and aptitude tests. In tests of this type, the goal is to measure the trait of interest with the same degree of precision over as wide a range of the trait scale as possible. Thus, the ideal TIF would be a uniform distribution over the trait scale range of interest. The results of Baker et al. (1988) showed that if a uniform TIF was used, the linear programming technique selected items that would match the function at the extremes and exceed it over the rest of the range. To counter this effect, the broad-range TIF was set to a value of 8 over the trait scale from −1.5 to +1.5. At −2.0 and +2.0 the TIF was set to 6, and at −2.5 and +2.5 it was set to 4. This TIF was used in the linear programming computer program, which selected 40 items from the item pool. The item parameters of these 40 items will be referred to as the basal parameters of the broad-range test. The summary statistics of the 40 broad-range test items were $\mu_a = 1.654$, $\sigma_a = .331$; $\mu_b = -.277$, $\sigma_b = 1.391$; and $0 \leq c \leq .35$ with $\mu_c = .210$, $\sigma_c = .041$.

Screening test. There are many situations in which the purpose of a test is to separate examinees into two groups (e.g., college admission, professional certification, or scholarship awards). In tests of this type, a criterion score is established and the greatest precision of trait score estimation is desired at the criterion score. Thus, the TIF would be strongly peaked and have its maximum at the criterion score. In the present case, the maximum amount of information was set to 18 at the arbitrarily selected decision point of $\theta = 1.0$, with a rapid decrease in the amount of information on either side of this maximum. This TIF served as the input to the linear programming computer program that selected 40 items from the item pool. This test had $\mu_a = 1.741$, $\sigma_a =$

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/

.306; $\mu_b = .644$, $\sigma_b = .485$; and $0 \le c \le .35$ with $\mu_c = .224$, $\sigma_c = .044$. This set of item parameters will be referred to as the *basal parameters* of the screening test.

Additional tests. It also was of interest to examine the sampling distributions of the equating coefficients in a vertical equating situation. Thus, it was necessary to establish sets of item parameters that represented the administration of each of the two types of tests to groups of examinees that were positioned at a point on the trait scale different from the basal groups. To do this, advantage was taken of the manner in which the identification problem is resolved in most IRT test calibration programs. Because the $\hat{\theta}$s are standardized, any change in the location and variance of the underlying examinee θ distribution is reflected in the values of the item parameter estimates. Thus, rescaling the basal item parameter estimates using Equations 2 and 3 is equivalent to the test being administered to a group of examinees whose scores are at a different point on the θ scale and have a different variance.

To implement this, let $A = 1.2$ and $K = -.5$. The value of A reflects a group of examinees that has a greater variance than the basal group. The value of K denotes that the new group has a lower mean trait score than the basal group. The basal item parameters of the broad-range and screening tests were rescaled using Equations 2 and 3 for these values of A and K. These additional sets of item parameters defined the *rescaled broad-range test* and the *rescaled screening test.*

Anchor-Item Configurations

In many testing situations, the same test is administered to multiple groups and the results of the separate test calibrations must be placed on a common metric. Because the same test was administered to each group in such a study, all the items in the test served as the anchor items. This anchor-item configuration will be denoted the *whole test.*

An issue of interest in regard to anchor items is the number needed to effectively equate tests. A commonly accepted general rule is that 15 anchor items are adequate (Cook & Peterson, 1987; McKinley & Reckase, 1981). To investigate this, 15 of the 40 items in the broad-range test and in the screening test were designated as anchor items. In each case, the subset of items was selected such that the TIF based on their item parameter values had the same form as the TIF of the full set of 40 items, but at a lower level of information. Note that the basal item parameter values of the whole test were retained; however, the identification of those items serving in the role of anchor items was changed. This anchor-item configuration is referred to as *15/40.*

When alternative forms of a given test are administered, placing the test results on a common metric involves the use of a subset of items that are common to all forms of the test. In the present case, the 15 items selected above from each set of basal item parameters for the broad-range and screening tests were retained as the anchor items. The remaining 25 items in a given 40-item test would be unique to that particular form of the test. This anchor-item configuration is referred to as *15/25.*

These three anchor-item configurations provided a means of examining a number of issues. Because the whole test and the 15/40 configurations involved the same basal item parameter values, the trait scale metric yielded by the BILOG (Mislevy & Bock, 1986) calibration is unaffected by the anchor-item configuration. Thus, comparing the sampling distributions of the equating coefficients obtained under the whole test and 15/40 configurations should provide some insight into the effect of using a subset of anchor items. In the 15/25 configuration, the trait scale metric yielded by BILOG will be unique to the particular form of the test used. Thus, the 15 anchor items were used to transform a greater variety of obtained metrics into a common metric. It was suspected that this situation would result in sampling distributions of the equating coefficients that differed from those yielded by the other two configurations.

Data Generation Procedures

The test administration results to be simulated for each type of test were based on the anchor-item configurations. Because the whole test and the 15/40 cases shared a common set of item parameters, they also shared the same sets of generated data.

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/

Thus, the basal item parameters for a given type of test were used in the GENIRV program (Baker, 1986) to create an item response vector of length 40 for each of 1,000 examinees. The group size was based on the results of Hulin, Lissak, & Drasgow (1982) who found that in order to obtain acceptable parameter estimates using the three-parameter IRT model, at least 1,000 examinees were required. The θs of these examinees were randomly sampled from a normal distribution with mean 0 and unit variance. The resulting item response vectors served as the input to BILOG, which produced the item parameter estimates for the 40 items in a test. This generation-estimation process was repeated 1,000 times for each of the four types of test specifications (broad, rescaled broad, screening, and rescaled screening tests).

In the case of the 15/25 anchor-item configuration, the parameter values of the 15 anchor items for a given type of test were fixed and the parameter values of the 25 unique items were randomly sampled from distributions having the same distributional characteristics as the basal item parameters of the 40 items. The same generation and estimation process was repeated for the four types of tests.

Equating Contexts

Item parameter recovery study. The evaluation of the item parameter estimation capabilities of an IRT test calibration procedure is commonly accomplished using a recovery study. In these studies, item response data are generated from known values of the item and examinee parameters. These data are then analyzed using a given estimation procedure and the recovery of the underlying item parameter values is assessed. [For examples, see Hulin et al. (1982), Swaminathan & Gifford (1983), and Yen (1981)].

However, to make a proper assessment, the obtained item parameter estimates must be placed on the metric of the underlying item parameters (Yen, 1987). The equating process involved is a special case of horizontal equating, and all the items in a test constitute the anchor items. In this case, a set of BILOG item parameter estimates served as the current test results, and the basal item parameters defined the target test. These two sets of item parameter values served as input to the EQUATE computer program (Baker et al., 1991), which yielded the values of the equating coefficients. This process was repeated for each of the 1,000 sets of item parameter estimates for the broad-range test and also for the screening test.

Equating to a baseline test. A common equating situation is one in which the results of a particular administration of a test serve as a baseline test. The same test or alternative forms of the test are administered and the results of these administrations are transformed into the same metric as the baseline test results. To obtain the item parameter estimates for the baseline tests, an additional set of item response data was generated for the broad-range, rescaled broad-range, screening, and rescaled screening tests using both the whole test and the 15/25 anchor-item configurations. Then, each dataset was calibrated using BILOG, yielding the item parameter estimates for the eight baseline tests that served as the target tests in the equating process.

Pairwise equating. The simplest equating task is one in which two administrations of a test are placed on a common metric. In the case of horizontal equating, the item parameter estimates for two test administrations were randomly sampled from the set of test results for a given type of test. The second test selected was equated to the first test selected and the values of the equating coefficients were obtained. This pairwise equating was performed 1,000 times under each of the three anchor-item configurations. Because the two tests were paired at random, the mean values of the obtained equating coefficients should recover the nominal values of $A = 1.0$ and $K = 0.0$.

To implement vertical equating, a set of test results was selected from those for a given test. Then a second set of test results was selected from those for the corresponding rescaled test. The first test then was equated to the second test using EQUATE and the values of the equating coefficients were obtained. This process was repeated for all 1,000 pairs of test results. Both types of tests were equated to their corresponding rescaled tests under all three anchor-item configurations. Again, because the tests

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/

were paired at random, the mean values of the obtained equating coefficients should recover the values of $A = 1.2$ and $K = -.5$ used to rescale the basal item parameters of the two types of tests.

Results

Item Parameter Recovery

The results for the item parameter recovery studies provide a frame of reference for the subsequent results. The summary statistics of the equating coefficients of the broad-range test were $\overline{A} = .961$, $\text{SD}_A = .025$ and $\overline{K} = -.023$, $\text{SD}_K = .038$; for the screening test, $\overline{A} = .910$, $\text{SD}_A = .023$ and $\overline{K} = .010$, $\text{SD}_K = .041$ (see Table 1).

The sampling distributions of A and K for the two types of tests are depicted in Figure 1. For each equating coefficient, the forms of the sampling distributions for the two types of tests were nearly the same (bell-shaped and approximately symmetric). Because the standard deviations (SDs) were essentially the same, the distributions for the two types of tests differed primarily in a shift of location. A major difference due to type of test appeared in the sampling distributions of the quadratic loss function. The distribution of F for the broad-range test was highly peaked with $\overline{F} = .043$; however, that of the screening test was nearly symmetric about a mean of .317. In addition, the SD of F for the screening test was .122, but the SD of the broad-range test was .033 (see Table 1). These results indicate that the screening test yielded poorer fit between the two TRFs than did the broad-range test.

Equating to a Baseline Test

In this equating situation, a given administration of a test served as the target test. Thus, the TRFs of both the current and target test were obtained from item parameter estimates. In the horizontal equating results, the means of A for both types of tests (.978 to 1.053) were higher than those in the item parameter recovery examples (see Table 1). The means of K showed a pattern in which the broad-range test means were greater than 0 (.001 to .041), but the screening test means (−.031 to −.117) were less than 0 (see Table 1). The screening test with a 15/25 anchor-item configuration yielded the anomalous value $\overline{K} = -.117$. In the vertical equating to a baseline test situation, the values of $\overline{A}$ for the screening test ranged from 1.093 to 1.166 and those of $\overline{K}$ ranged from −.480 to −.522 (see Table 1), reflecting the constants used to create the rescaled tests. Both the magnitude and patterns of SDs of the equating coefficients were similar across the two equating situations and types of tests.

Although the graphs for all the sampling distributions were obtained, they were similar in appearance across all three anchor-item configurations. Consequently, only those for the 15/25 configuration are provided in Figure 2. The forms of the sampling distributions of the equating coefficients again were basically bell-shaped and approximately symmetric for both types of tests and equating situations. The noticeable differences were in the locations of the distributions on the trait scale.

The summary statistics for the loss function had a consistent pattern across both equating situations and types of tests. Under horizontal equating, the broad-range test, and with a whole test anchor-item configuration, Table 1 shows that the means and SDs ($\overline{F} = .021$, $\text{SD}_F = .019$) were smaller than were observed in the item parameter recovery example ($\overline{F} = .043$, $\text{SD}_F = .033$). In the vertical equating situation, the variability of F for the screening test ($\text{SD}_F = .030$) was greater than that of the broad-range test, but was 1/4th the magnitude of that observed in the item parameter recovery data. Overall, Table 1 shows that the mean values of F were near 0 (.004 to .029) and the SDs were small (.001 to .030) indicating a good fit between the TRFs involved. The form of the sampling distribution of F for the screening test was L-shaped and closely matched that of the broad-range test.

Pairwise Equating

In this case, both the current and target tests were randomly sampled from a set of 1,000 test results. Here, pairwise equating can be viewed as a recovery study for the Stocking and Lord procedure. The nominal values $A = 1.0$, $K = 0.0$ and $A = 1.2$, $K = -.5$ should be recovered under horizontal equating and vertical equating, respectively. Under horizontal equating, the maximum bias in $\overline{A}$ was only .002

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
Mean and SD of Equating Coefficients and the Loss Function F for Three Equating Contexts, Horizontal and Vertical Equating, and Test Types

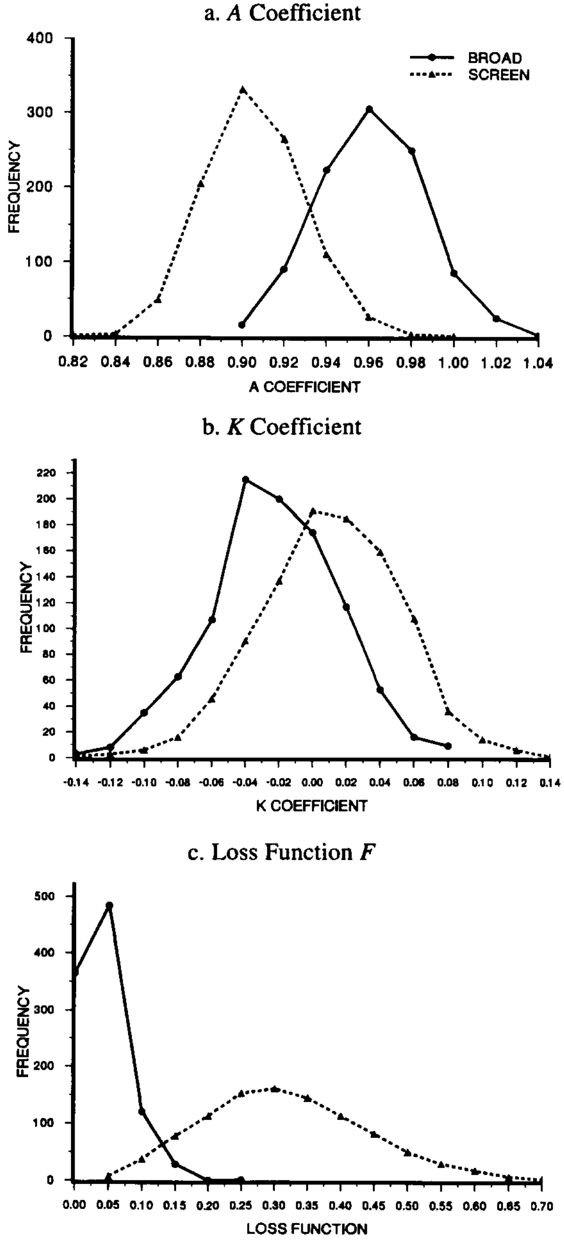
Type of Equating and Test	$\bar{A}$	SD_A	$\bar{K}$	SD_K	$\bar{F}$	SD_F
Recovery Study						
Broad–Range Test	.961	.025	–.023	.038	.043	.033
Screening Test	.910	.023	.010	.041	.317	.122
Equating to a Baseline Test						
Horizontal Equating						
Broad–Range Test						
Whole Test	1.015	.026	.001	.040	.021	.019
15/40	1.053	.037	.029	.044	.010	.008
15/25	.978	.032	.041	.046	.004	.004
Screening Test						
Whole Test	1.038	.025	–.031	.048	.026	.030
15/40	1.030	.035	–.033	.053	.015	.015
15/25	.995	.035	–.117	.054	.006	.007
Vertical Equating						
Broad–Range Test						
Whole Test	1.159	.029	–.424	.045	.029	.023
15/40	1.138	.039	–.338	.048	.004	.004
15/25	1.161	.036	–.478	.055	.007	.001
Screening Test						
Whole Test	1.166	.029	–.480	.053	.010	.012
15/40	1.144	.039	–.445	.059	.007	.009
15/25	1.093	.039	–.522	.055	.005	.005
Pairwise Equating						
Horizontal Equating						
Broad–Range Test						
Whole Test	1.001	.037	–.003	.055	.027	.027
15/40	1.002	.049	–.002	.060	.008	.008
15/25	1.001	.047	.001	.069	.008	.008
Screening Test						
Whole Test	1.002	.034	–.001	.064	.025	.030
15/40	1.000	.049	0.000	.073	.010	.014
15/25	1.000	.049	–.002	.076	.011	.015
Vertical Equating						
Broad–Range Test						
Whole Test	1.178	.044	–.530	.063	.033	.035
15/40	1.179	.063	–.523	.069	.007	.007
15/25	1.163	.059	–.519	.072	.007	.008
Screening Test						
Whole Test	1.127	.040	–.493	.066	.033	.041
15/40	1.119	.055	–.485	.078	.010	.013
15/25	1.116	.054	–.489	.079	.010	.014

across both types of tests. The bias in $\bar{K}$ ranged from –.003 to .001. Thus, the recovery of the underlying values of A and K was excellent. Under vertical equating, the biases of $\bar{A}$ ranged from .021 to .084 and those of $\bar{K}$ from –.030 to .015. Within both horizontal and vertical equating, the pattern of SDs was consistent across both types of tests. However, the size of the SDs was approximately .015 to .025 greater than those observed in the equating to a baseline case. Given the random pairing of the TRFs, this was to be expected.

Figure 3 shows the sampling distributions under

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/

Figure 1
Sampling Distributions of Equating Coefficients *A* and *K* and the Loss Function *F* for the Item Parameter Recovery Study



a. *A* Coefficient

b. *K* Coefficient

c. Loss Function *F*

the 15/25 anchor-item configuration. The forms of the sampling distributions again were bell-shaped and symmetric. Under horizontal equating, the distributions due to the two types of tests were nearly identical for both *A* and *K*. Under vertical equating, the forms agreed, but they differed slightly in location. The graphs of the loss function again were L-shaped but were not as peaked and had greater spread than in the baseline equating results (see Figures 2c and 2f). However, the means of all four plotted distributions of *F* were less than .011, and the SDs were less than .015. These results indicate that a good fit was achieved between the random pairs of TRFs.

Anchor-Item Configurations

Another question of interest was the impact of the anchor-item configuration on the sampling distributions of the equating coefficients. When 15 anchor items were used, SD_A was approximately .010 to .019 larger than those of the whole test configuration over all eight equating situations. The corresponding increases in SD_K ranged from .005 to .014. It was conjectured above that there would be a difference in the results between the 15/40 and 15/25 configurations, due to the differences in the trait scale metrics constructed within BILOG. The obtained metric of the 15/25 configuration was specific to the pairing of the 15 anchor items and the 25 unique items, but the 15/40 shared its metric with the whole test. With two exceptions (both involving the screening test and a baseline test), there were only slight differences in the means of the equating coefficients due to these two 15-item configurations. The SDs of the equating coefficients showed only minor differences due to these two anchor-item configurations.

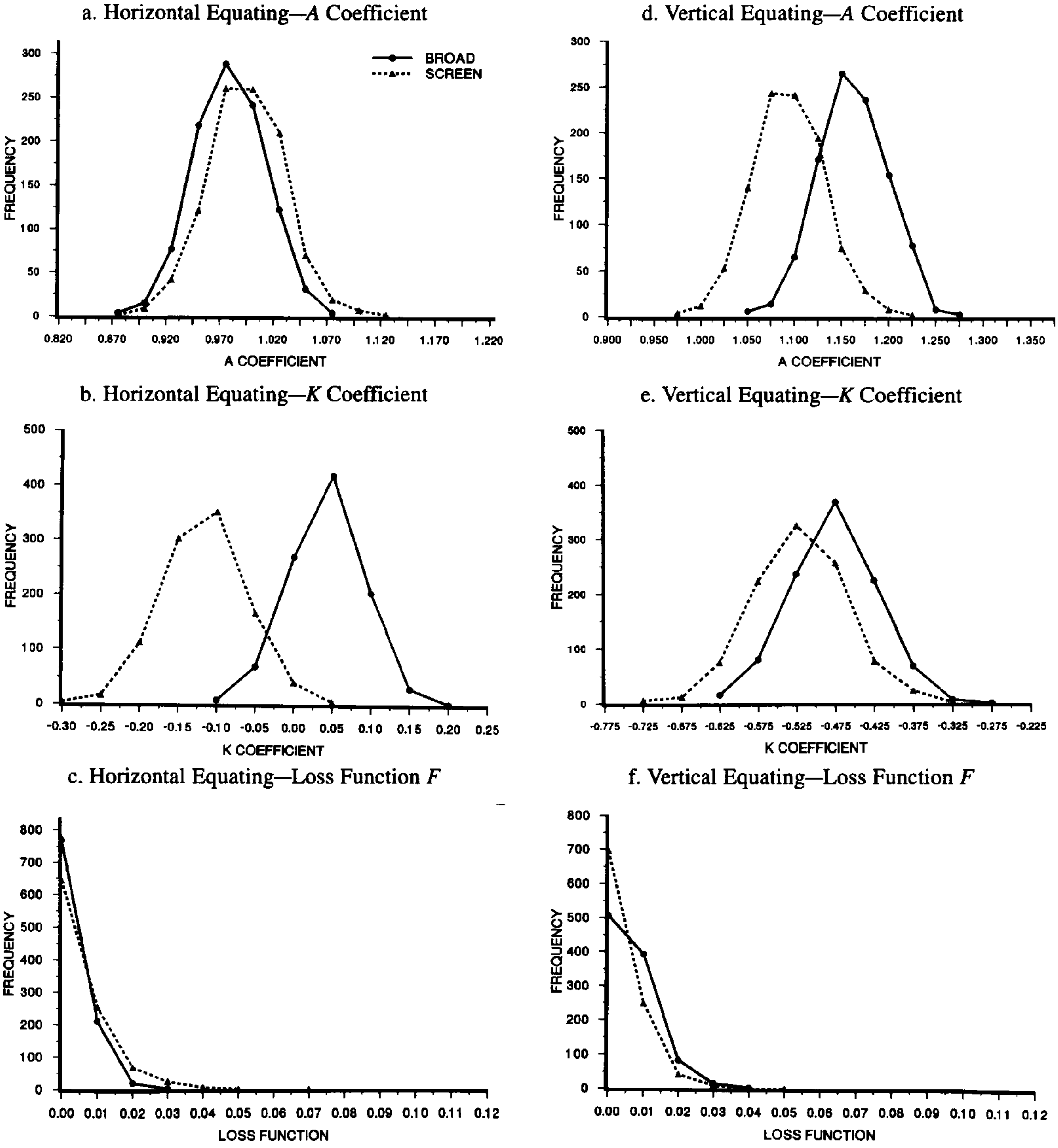
There was a distinctive pattern in the means and SDs of the loss function attributable to the anchor-item configurations. When the whole test was the anchor, $\overline{F}$ ranged from .010 to .033 and SD_F ranged from .019 to .041. When the number of anchor items was 15, $\overline{F}$ ranged from .004 to .011 and SD_F ranged from .001 to .015.

Impact of Sampling Variability of the Equating Coefficients

It was also of interest to examine the impact of the sampling variability of the obtained values of *A* and *K* on the transformation of the current test pa-

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/

Figure 2
Sampling Distributions of Equating Coefficients and the Loss Function:
Baseline Equating Using the 15/25 Anchor-Item Configuration



a. Horizontal Equating—*A* Coefficient

d. Vertical Equating—*A* Coefficient

b. Horizontal Equating—*K* Coefficient

e. Vertical Equating—*K* Coefficient

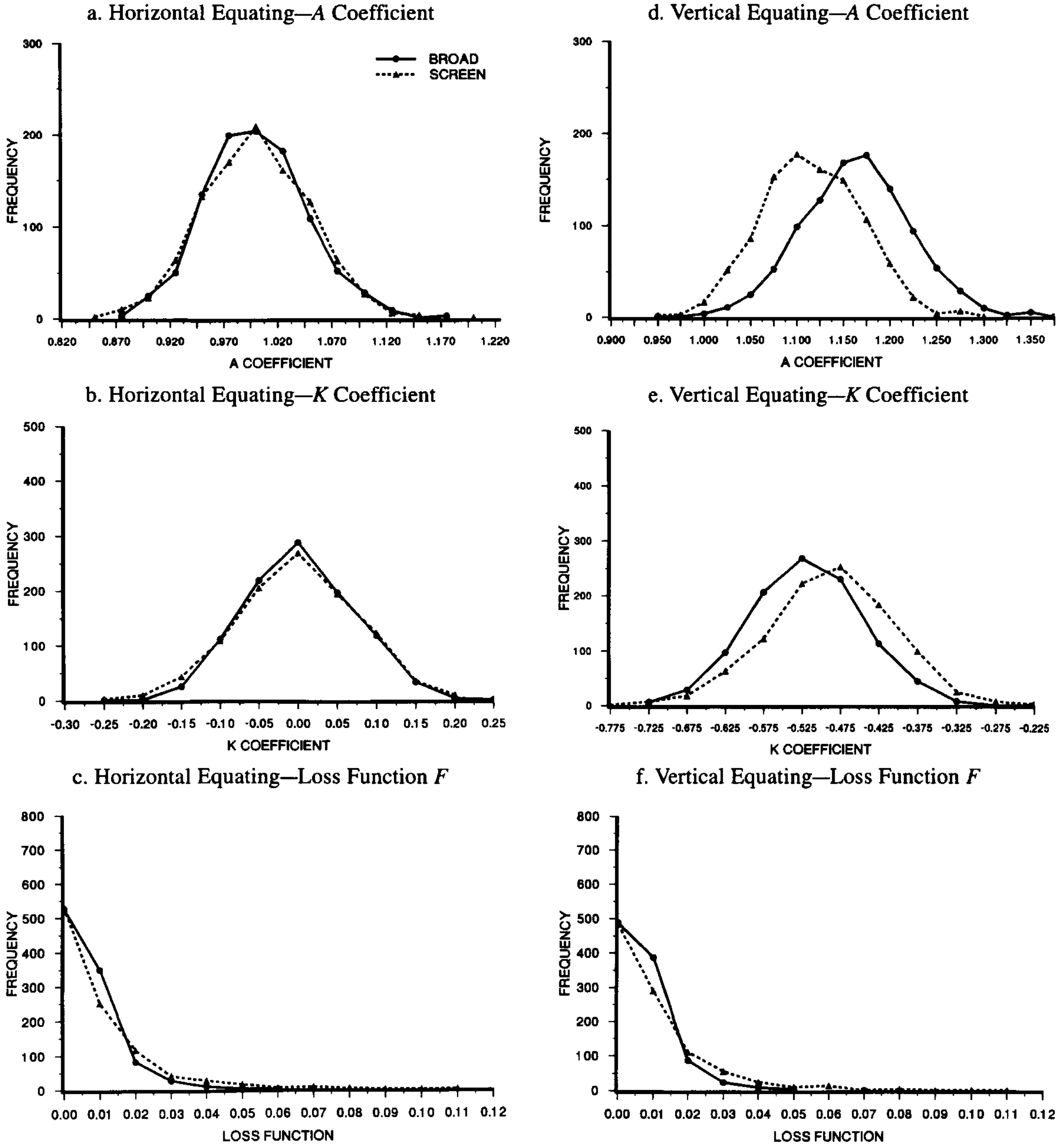
c. Horizontal Equating—Loss Function *F*

f. Vertical Equating—Loss Function *F*

rameter estimates into the common metric. To examine this, values of the equating coefficients were determined at ±1 SD from the observed means of *A* and *K* over the 12 equating contexts within each of the horizontal and vertical equating situations. These upper and lower values were then used to transform

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/

Figure 3
Sampling Distributions of Equating Coefficients and the Loss Function:
Pairwise Equating Using the 15/25 Anchor-Item Configuration



a. Horizontal Equating—*A* Coefficient

b. Horizontal Equating—*K* Coefficient

c. Horizontal Equating—Loss Function *F*

d. Vertical Equating—*A* Coefficient

e. Vertical Equating—*K* Coefficient

f. Vertical Equating—Loss Function *F*

selected item parameter values (a_i = .5, 1.0, 1.5; b_i = −1.0, 0, 1.0) into the common metric using Equations 2 and 3. The transformed boundary values of the a_is differed from the a_i*, based on the average values of *A*, by approximately 3.8% in both the horizontal and vertical equating situations. The boundary values of

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 transformed b_is differed from the values of b_i^*, based on the average values of A and K, by 5.7% and 6.2% for the horizontal and vertical equating situations, respectively. These differences would appear to have little effect on the interpretation of the transformed item parameter estimates.

It also was useful to examine the variability of the transformed examinee θs as a function of the sampling variability of the equating coefficients. The delta method due to Kendall & Stuart (1963, pp. 231–233) yields the following result:

$$\mathrm{Var}(\theta^*) = \theta^2\mathrm{Var}(A) + \mathrm{Var}(K) + \mathrm{Cov}(A,K), \qquad (10)$$

where Var denotes variance and Cov denotes covariance. The variances of θ* were obtained at values of θ = 0, 1, and 2. Because the covariances of A and K were one or two orders of magnitude smaller than the variances, they are not reported here. However, they were used when computing the variances of the transformed θs using Equation 10. The ratio of the standard errors (SEs) of the θ* to the SEs of the $\hat{\theta}$s provides an additional means of evaluating the impact of the sampling variability of the equating coefficients. The latter were obtained from the TIFs based on the underlying item parameters for the broad-range and screening tests. Because the sampling variances of the equating coefficients were different for the whole test and 15 anchor-item configurations, the averages of the ratios for each case are reported in Table 2.

In the horizontal equating situation, the ratios yielded by the whole test ranged from .121 to .186 and those for the 15 anchor items ranged from .160 to .247. There was an increase in the whole test ratios as θ increased from θ = 0 to θ = 2. A somewhat greater increase was noted in the results for the combined 15/25 and 15/40 conditions. In the vertical equating situation, the variances of θ* were similar to those obtained under horizontal equating but the ratios were somewhat larger. Again, the ratios for the whole test (.1774 to .2274) were slightly smaller that those for the 15 anchor-item case (.2235 to .2888). The maximum values for both the whole test (.2448) and 15 anchor items (.2883) cases occurred at θ = 1 rather than at θ = 2 as was the case for horizontal equating.

Discussion

A primary interest of the present study was to determine whether the equating coefficients (A and K) were "well-behaved" under random sampling in a variety of equating contexts. Although the sampling distributions of A and K were graphed for only the 15/25 anchor-item configuration, these graphs were representative of those observed for the other two configurations. The obtained sampling distributions were bell-shaped, symmetric, and had no outliers or other abnormalities in all of the contexts studied. Only slight differences in the forms of the distributions could be attributed to the type of test used. For the contexts studied, it appears that the sampling distributions of the equating coefficients were indeed well-behaved.

Differences in the summary statistics of A and K were associated with the target tests used. The item parameter recovery study is a special case of horizontal equating in which the target test TRF is defined by the underlying item parameters. If the

Table 2
Variances of θ* and Ratio of SE(θ*)/SE($\hat{\theta}$) at Selected Trait Scale Values for Horizontal and Vertical Equating Using the Whole Test and 15 Anchor Items

Type of Equating	θ = 0		θ = 1		θ = 2	
and Test	$\sigma^2(\theta^*)$	Ratio	$\sigma^2(\theta^*)$	Ratio	$\sigma^2(\theta^*)$	Ratio
Horizontal Equating						
Whole Test	.002	.121	.003	.152	.006	.186
15 Items	.003	.160	.005	.199	.010	.247
Vertical Equating						
Whole Test	.003	.227	.004	.245	.008	.177
15 Items	.004	.247	.006	.288	.013	.224

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/

estimation procedures in BILOG recover the underlying item parameter values, $A = 1.0$ and $K = 0.0$ should be closely approximated. The obtained mean values differed somewhat from these nominal values. However, this does not imply that they were biased. The Stocking & Lord (1983) procedure simply finds the values of A and K of Equation 1 that best match the TRF of the current test to that of the target test. If BILOG does not recover the underlying item parameter values, this will be reflected in the TRFs and hence in the values of A and K.

Technically, equating to a baseline test is the same process used in the item parameter recovery study. However, the TRF of the target test is based on item parameter estimates that have estimation errors embedded in them. Surprisingly, the sampling distributions of the quadratic loss function for the screening test were dramatically better when equating to a baseline test than in the item parameter recovery case. The means and SEs under the former were much smaller than under the latter (see Table 1 and Figures 1c, 2c, and 2f). In addition, the forms of the distributions of F were distinctly different, with those under the item parameter recovery study being the deviant cases. A possible explanation is that the item parameter estimates of the baseline screening tests had characteristics that were typical of the BILOG results for the current tests. Thus, the item parameter estimates and trait scale metrics of the baseline tests were a better match to the current tests than to the underlying parameters.

In the pairwise equating case, both the current test and the target test were randomly sampled from the 1,000 sets of test results. This difference in test pairing is apparent by a somewhat tighter clustering of the equating coefficients around their means when equating to a baseline than when pairwise equating.

The anchor-item configuration had some impact on the summary statistics of the sampling distributions of A and K. There were only small differences in the means of the obtained equating coefficients as a function of whether the whole test or 15-item anchor configurations were used. Although the corresponding SEs differed by large relative amounts (15% to 45%), the absolute amounts were small (.003 to .019). The ratios of the SEs of the transformed θs to those of the $\hat{\theta}$s suggest that the sampling variability of the θ^* amounts to from 12% to 29% of the SE of estimate of the θs. A word of caution is in order here because both SEs reflect the parameter recovery characteristics of BILOG.

When equating to a baseline test, there generally were minor differences between the 15/25 and 15/40 anchor-item configuration results, despite the fact that they used different baseline tests in the same equating context. This lack of effect might be attributable to sampling the item parameters for the 25 unique items from distributions whose characteristics matched those of the basal item parameters for a given test. When constructing alternate forms of a test, an attempt is made to match the item characteristics across the multiple forms of a test. Thus, the process used here was consistent with test construction practices. The results obtained for the various anchor-item configurations tend to confirm the view that approximately 15 anchor items are adequate for equating tests under IRT.

The sampling distributions of the minimum value of the quadratic loss function F showed that in all but one case the differences between the final TRFs were very small. When the whole test anchor-item configuration was used, the mean value of F was approximately .026, but those for the 15/40 and 15/25 configurations were less than .009. These differences between the whole test and the 15-item configurations appear to be an artifact of the relative magnitudes of the ENCs. In the latter case, there is less opportunity for the TRFs to differ by large amounts and the value of F will be less. Hence, the smaller value is a function of the number of anchor items used rather than an indication of a better fit of the final TRFs. The sampling distributions of F also suggest that the multivariate search procedure for finding the values of the equating coefficients does an excellent job of minimizing the quadratic loss function of Equation 9.

The only exception to this pattern occurred in the screening test item parameter recovery study in which the distribution was bell-shaped rather than L-shaped. A partial explanation of this result lies in the nature of the item parameter estimation situation when a screening test is used. In a screening test, the test

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/

items are placed on the trait scale so as to maximize the TIF at the criterion score rather than in the center of the θ scale, as is the case for a broad-range test. Consequently, there are fewer examinees located where the item difficulties are being estimated. This results in larger bias and larger SEs of the item parameter estimates used to create the TRFs, which is then reflected in the obtained value of the loss function. Thus, the value of F reflected the parameter recovery characteristics of BILOG. This suggests that F could be used as a global measure of the recovery capabilities of an estimation procedure.

It would have been disconcerting if the forms of the distributions and the summary statistics varied widely across the many equating contexts. That they did not suggests that unusual values of the equating coefficients were not common. In addition, the comparisons of the three anchor-item configurations indicate that test equating can be conducted using a subset of 15 anchor items. The few mildly deviant results for screening tests suggest that careful selection of a baseline test would be prudent when this type of equating is used.

References

Baker, F. B. (1986). *GENIRV: Computer program for generating item responses.* Unpublished manuscript, University of Wisconsin, Madison.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87–96.

Baker, F. B., Al-Karni, A., & Al-Dosary, I. M. (1991). EQUATE: A computer program for the test characteristic curve method of IRT equating. *Applied Psychological Measurement, 15,* 78.

Baker, F. B., Cohen, A. S., & Barmish, B. R. (1988). Item characteristics of tests constructed via linear programming. *Applied Psychological Measurement, 12,* 189–199.

Cook, L. L., & Peterson, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11,* 225–244.

Davidon, W. C. (1959). *Variable metric method for minimization* (Research and Development Report ANL-5990, rev. ed.). Argonne IL: Argonne National Laboratory, U.S. Atomic Energy Commission.

Eastern Software Products, Inc. (1985). *BLP88 user's guide: Linear programming with bounded variables for the IBM PC.* Alexandria VA: Author.

Fletcher, R., & Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. *The Computer Journal, 6,* 163–168.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144–149.

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

Kendall, M. G., & Stuart, A. (1963). *The advanced theory of statistics* (Vol. 1). New York: Hafner.

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

McKinley, R. L., & Reckase, M. D. (1981). *A comparison of procedures for constructing large item pools* (Research Rep. No. 81-3). Columbia: University of Missouri, Department of Educational Psychology.

Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models.* Mooresville IN: Scientific Software Inc.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes.* Cambridge, England: Cambridge University Press.

Slinde, J. A., & Linn, R. L. (1977). Vertically equated tests: Fact or phantom? *Journal of Educational Measurement, 14,* 23–32.

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

Swaminathan, H., & Gifford, J. A. (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.

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika, 50,* 411–420.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245–262.

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

Author's Address

Send requests for reprints or further information to Frank B. Baker, Department of Educational Psychology, 1025 W. Johnson St, University of Wisconsin, Madison WI 53706, U.S.A. Email: fbaker@macc.wisc.edu.

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/