ED 431 800 TM 029 886

AUTHOR	Li, Yuan H.; Lissitz, Robert W.; Yang, Yu Nu
TITLE	Estimating IRT Equating Coefficients for Tests with Polytomously and Dichotomously Scored Items.
PUB DATE	1999-04-00
NOTE	34p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Equated Scores; *Estimation (Mathematics); *Item Response Theory; Test Format; *Test Items
IDENTIFIERS	Common Item Effect; *Dichotomous Scoring; Linking Metrics; *Polytomous Scoring

ABSTRACT

Recent years have seen growing use of tests with mixed item formats, e.g., partly containing dichotomously scored items and partly consisting of polytomously scored items. A matching two test characteristic curves method (CCM) for placing these mixed format items on the same metric is described and evaluated in this paper under a common-item linking design in which tests containing a set of common items are administered to two groups of examinees. The BIAS statistics of the recovery equating coefficients across all research conditions are usually expected to be close to zero, as happened in this simulation, and the corresponding root mean squared error (RMSE) statistics, in general, are relatively small. The shapes of the empirical sampling distributions of the estimated equating coefficients obtained from a variety of conditions are symmetric-bell shapes with small variances and few or no outliers or other abnormalities. The CCM is capable of producing accurate transformation of parameter estimates when applied to a test with mixed items as well as to only dichotomously scored or polytomously scored tests. (Contains 3 tables, 4 figures, and 35 references.) (Author/SLD)

Estimating IRT Equating Coefficients for Tests with Polytomously and Dichotomously Scored Items

Yuan H. Li

Prince George's County Public School, Maryland

Robert W. Lissitz & Yu Nu Yang

University of Maryland at College Park

Paper Presented at the Annual Meeting of the
National Council on Measurement in Education (NCME).
Montreal, Canada, April 19-23, 1999

Estimating IRT Equating Coefficients for Tests with Polytomous and Dichotomous Scored Items

Abstract

Recent years have seen growing use of a test with mixed item formats, e.g., partly containing dichotomous scored items and partly consisting of polytomous scored items. Matching two test characteristic curves method (CCM) for placing these mixed format items on the same metric is described and evaluated in this paper under a common-item linking design in which tests containing a set of common items are administered to two groups of examinees.

The BIAS statistics of the recovery equating coefficients across all research conditions are usually expected to be close to zero as happened here and the corresponding RMSE (root mean squared error) statistics, in general, are relatively small. The shapes of the empirical sampling distributions of the estimated equating coefficients obtained from a variety of conditions are symmetric-bell shapes with small variances and few or no outliers or other abnormalities. The CCM is capable of producing accurate transformation of parameter estimates when applied to a test with mixed items as well as to only dichotomous-scored or polytomous-scored tests.

Key Words: Item Linking; Item Bank; Test Equating

I. Introduction

A. Motivation

Item response theory (IRT) consists of a family of probabilistic models that hypothesize the relationship between an examinee's latent ability(ies) and a correct response to an item. Various IRT models have now been routinely used to construct a large item bank with all items on the same metric. This is possible because of the property that the metric thus calibrated is invariant under a linear transformation (Stocking & Lord, 1983). Recent years have seen growing use of a test with mixed item formats, e.g., partly containing dichotomous scored items and partly consisting of polytomous scored items (e.g., Beaton & Zwick, 1992; Maryland State Department of Education, 1997). When constructing mixed-format item banks, interesting problems present themselves. Methods for placing these mixed format items on the same metric are described and evaluated in this paper under a common-item linking design (Vale, 1986) in which tests containing a set of common items are administered to two groups of examinees.

It is important to emphasize that we are assuming in this paper that each type of item is measuring one single latent trait. Hence dimensionality differences and procedures for equating multivariate item response tests (Li & Lissitz, in press) for various types of items are not a subject of this paper. Fitting a unidimensional IRT model to the mixed-format test data may be suitable if the dichotomous- (e.g., multiple-choice items) and polytomous- scored items (e.g., free-response items) are counterparts of one another to measure an identical latent ability, unless there are effects of the item formats themselves (Thissen, Wainer and Wang, 1994). The results based on factor analysis on test data (e.g., Bennett, Rock & Wang, 1991; Thissen, et al, 1994) suggested that a one-factor solution may provide a more parsimonious fit although even a relatively small amount of local dependence among the free-response items will produce a small degree of multidimensionality. However, if the free-response items are genuinely intended to measure something different than the multiple-choice items, using the unidimensional IRT model to capture the mixed-format test data is unacceptable. Apparently, the dimensionality issue of examinee's item responses to mixed-format items is substantive and needs to be closely explored before conducting an IRT-based item linking. Assessing unidimensionality of dichotomous and polytomous data exist (Nandakumar, Yu, Li & Stout, 1998).

B. Background of IRT-based Common-item Linking

Linear Transformation Method:

In reviewing the research on common-item linking methods by a linear transformation procedure for dichotomous- (e.g., Divgi, 1985; Linn, Levine, Hastings & Wardrop, 1981; Stocking & Lord, 1983) or polytomous- (e.g., Baker, 1992; Kim & Cohen, 1995) scored items, we found that the characteristic curve method (CCM) which matches the test characteristic curves between the base and the equated tests (Stocking & Lord, 1983) is the most widely adopted approach to estimating the equating coefficients. The basic principle behind CCM is to minimize the squared difference between the two expected true scores derived directly from the two sets of common-item parameter estimates for arbitrary ability points, where the expected true score is obtained by summing the probabilities of correct responses of common items.

Several features are particularly salient when CCM is applied to estimate equating coefficients for a dichotomous- or polytomous- scored test. One is that the sampling distributions of equating coefficients produced by the CCM method either for dichotomous-scored tests (Baker, 1996) or for polytomous-scored tests (1997) "were approximately bell-shaped, had small variances, and no outliers or other abnormalities." Another is that local item dependence (LID) that will often occur on the free response items (e.g., performance assessment) is not expected to affect the accuracy of equating unless LID has significantly different impact on the parameter estimates of the same items (Yen, 1993). This is because item independence is not assumed at the process of computing the true score. Finally, the CCM equating method seems better than the other approaches if the primary concern of a test practitioner is the equivalence of the expected true scores between two groups of examinees. Accordingly, the CCM is a promising equating method to be investigated when a test is constructed with mixed-format items. Algorithms for finding CCM's equating coefficients for mixed-format common items are introduced in the section of literature review.

The results from the study conducted by Donoghue (1994) indicated that, on average, four-category polytomous items yield 2.1 to 3.1 times as much IRT information as dichotomous items. This level of additional information may be used to estimate equating coefficients by matching test information curves (MTIC) provided by a set of common items between the base and equated tests. Unfortunately, informal work with several equating examples shows that in some cases the MTIC's equating coefficient estimates are far from the true values. One likely reason behind this is that locally maximum values exist in the quadratic loss function defined by minimizing two test information curves. Thus, the MTIC method was not considered in this study.

Using the equating coefficients (e.g., produced from the CCM equating method), the linear transformation procedure put item parameter estimates calibrated from different tests on the same scale. Besides that, two alternatives are available for item linking without estimating equating coefficients. Each has been applied in some testing settings and illustrated in the following section.

Concurrent Calibration Method

One is the concurrent calibration method. The process of this linking method is, first, to combine test datasets by treating those items not taken by any particular group as "not reached items" during compilation (Hambleton, Swaminathan & Rogers, 1991). Item (or ability) parameters from different tests are then simultaneously estimated and transformed onto a common scale via a single computer run. The principle behind this is that the metric of item parameter estimates from different test forms are jointly referred to the same identical ability scale. This equating method satisfied Mislevy and Bocks' (1982) requirement that "the information needed for a proper link is found in not just the item parameter estimates and their standard errors, but in the matrix of correlations among the estimates as well (p.15)." Popular computer software packages such as BILOG (Mislevy & Bock, 1990) for dichotomous scored items and PARSCALE (Muraki & Bock, 1996) for polytomous (or mixed) scored items provide this nice feature. This linking method has been applied to the large-scale testing program of the National Assessment of Educational Progress (NAEP) (Beaton & Zwick, 1992).

Since the concurrent calibration method makes complete use of the available information and may potentially remove some equating errors produced by the inaccurate transformation functions that are used to equate the two tests, this linking method may produce a more stable equating result than other linking methods did. For polytomous-IRT item parameter linking (Kim & Cohen, 1997), this method yielded consistently, albeit only slightly, smaller root mean square differences than the CCM. On the other hand, the results from linking dichotomous-IRT parameter estimates (Kim & Cohen, 1998) showed this linking method produced larger root mean square difference than CCM for smaller number of common items, but yielded the same results for larger number of common items.

A potential problem that confronts the concurrent calibration method is that this equating method may encounter problems of locating item parameter estimates when the test data are originated from groups with extreme differences with respect to the locations and variabilities of the ability distributions (see Kim & Cohen, 1997).

Fixed Precalibrated Item Parameter (FPIP) Method

Another item-linking method is to fix the precalibrated item parameters (FPIP) during the calibration process. Mislevy and Bock (1990) pointed out that "By specifying tight priors on selected item parameters, the user may hold these values essentially fixed while estimating other item parameters. This feature is useful in linking studies, where new test items are to be calibrated into an existing scale without changing parameter values for old items (pp.2-6, 2-7)." FPIP holds similar nice properties as the concurrent calibration methods have. It also takes advantage of its flexibility in being adaptable to a variety of data collection methods (see Vale, 1986). For instance, it may be employed on the CAT's (computerized adaptive testing) "on-line" item linking that is carried out while the preequating items and new non-linking items are being administered simultaneously.

However, caution should be exercised for FPIP because in most situations the context and positions of the precalibrated items cannot be fixed from one testing to the next. The critical question of whether this linking method can produce fairly robust linking results under those circumstance needs to be further investigated.

A simulation study conducted by Li, Griffith and Tam (1997) to compare the performance between the FPIP and CCM on dichotomous-IRT parameter linking showed that both methods produced similar root mean square difference for item discrimination estimates or for medium difficulty estimates. The FPIP performed slightly better on easy items; in contrast, it produced less precision on hard and medium difficulty items.

C. Statement of Research Question

The findings from Baker (1996; 1997) suggested that CCM is a sound equating method for linking dichotomous- or polytomous- item parameter estimates. The practical question of "Can the CCM produce accurate transformation of parameter estimates when applied to a test with mixed-format items?" was closely examined under several simulation conditions in this study.

The relative importance of determining the equating coefficients under CCM for each type of item depends on the magnitude of corresponding expected item true score variance. The expected true score can be analogous to the weighted raw score, where the weight is a function of an IRT model. In other words, the factor of the relative importance of each type item is very critical to determining the equating coefficients for tests with mixed-format items. Determining the impact of weighting the proportion of different item formats was one of the research questions evaluated in this study. This special issue was unable to be investigated in the study using dichotomous-scored tests (Baker, 1996) or the study using polytomous-scored tests (Baker, 1997).

It is important to stress that weighting k-category polytomous items the same as dichotomous items for computing total scores will probably not be indicative of their true relative importance within the total set of items in a constructed test form. Although this problem is irrelevant to our research questions, it may have impact on the precision of the equipercentile equating (Kolen & Brennan, 1995) results when the total raw score is used, for example.

Since the measurement errors of item parameter estimates produce inaccurate equating coefficients, modeling measurement errors of item estimates is another key feature. The approach to modeling standard errors of item estimates (refer to Thissen & Wainer, 1982) employed in Li and Lissitz's study (in press) was adopted in this study. This approach is discussed later, where the comparison with the approach used in Baker's studies (1996; 1997) is made. Besides that, several factors such as sample size, equating situations, etc. that account for the variation of the equating parameter estimates were also included in this study.

II. Literature Review

A. IRT Models for Dichotomous and for Polytomous Test Data

Three-Parameter Logistic IRT Model for Dichotomous Data

The commonly-used three-parameter logistic IRT model was used to model the dichotomous scored items in this study. Under the three-parameter logistic model, the probability, P_{ij}, of a correct response to the i item for the j examinee with ability θ_j is given by (Lord, 1980):

$$P_{ji}(\theta_j) = c_i + (1 - c_i)\frac{\exp(Da_i(\theta_j - b_i)}{1 + \exp(Da_i(\theta_j - b_i)} \qquad (1)$$

where "exp' in Equation 1 stands for the mathematical function of the natural logarithm exponential, a_i is the item discrimination, b_i is the item difficulty, c_i is the lower asymptote parameter (also known as the guessing parameter), and D (usually equal to 1.702) is a scaling factor.

Generalized Partial Credit Model

The test data from polytomous scored items was modeled by the generalized partial credit model (GPCM) (Muraki, 1992), in which the probability, P_{ijk}, of the categorical response k on item i for an individual j with ability θ is given by the familiar logistic function:

$$P_{jik}(\theta_j) = \frac{\exp\left[\sum_{v=1}^{k} Da_i(\theta_j - b_i + d_k)\right]}{\sum_{c=1}^{m} \exp\left[\sum_{v=1}^{c} Da_i(\theta_j - b_i + d_k)\right]} = \frac{\exp\left[\sum_{v=1}^{k} Da_i(\theta_j - b_{ik})\right]}{\sum_{c=1}^{m} \exp\left[\sum_{v=1}^{c} Da_i(\theta_j - b_{ik})\right]} \qquad (2)$$

where a_i is a slope parameter (or item discrimination), b_i is an item-location parameter (or item difficulty), and d_k is a category (or step) difficulty. The values of d_k within an item are not necessarily ordered sequentially. It is interpreted as the relative difficulty of category k with other category parameters within an item or the deviate from corresponding item location, b_i. Only m_i-1 item-category parameters can be identified when the number of response categories is m_i. The step difficulty of the first category (d_1)on each item is arbitrarily set to zero and the location constraint of $\sum_{k=2}^{m_i} d_k = 0$ is imposed to eliminate indeterminacy (Muraki, 1992). And, b_{ik} equals to b_i-d_k.

B. Rescaling the Metric of Ability or Item Parameter Estimates

Numerical estimates of the IRT item parameters depend upon the ability (θ) scale (Baker, 1992). Although, the ability scale is usually standardized to have a mean of zero and a standard deviation of one in "any" item response data being analyzed, the original (not standardized) ability scale is different from one test dataset to another. Thus, when the "same" set of test items is administered to two different groups and the resultant response data are calibrated separately, the two sets of item parameter estimates are usually different because they refer to different underlying ability scales. This problem can be resolved by using one of the item linking methods reviewed previously. However, the fact that IRT scales can be exchangeable only holds when the model selected fits the data (see McKinley & Mills, 1985; Reise, 1990; Tam & Li, 1997, for reviews of popular methods for evaluating model-data fit). If there is significant lack of fit, this property will no longer be valid. Besides, we should be aware that the common scale thus constructed is still arbitrary. Its interpretation must be carried out with caution.

The principle behind the IRT-based item linking is that the metric of item parameter estimates from different test forms taken by different groups are forced to refer to a "common" ability scale that is carried out by re-scaling the multiple-group ability scales. This common scale is performed by a linear transformation, given below for the case of two groups. This transformation illustrates the relationship between two groups' (base and equated) ability scales.

$$\hat{\theta}^{*}_{B} = A\hat{\theta}_{E} + B \qquad (3)$$

where the superscript * represents the transformed values from the equated scale to the base scale, A is the slope, and B is the intercept. The subscripts "B" and "E" represent the base and equated groups, respectively. In this case, the ability scale for the base group remains unchanged so that the metric of the item parameter estimates for the base group are not needed to be re-scaled. In contrast, the ability scale for the equated group is transformed into the scale defined by the base group. The probability of a correct response for any item given an individual's new-scaled ability values for the equated group remains unchanged only if the original item parameter estimates are rescaled by the following linear transformations. For the three-parameter item parameter estimates,

$$b^{*}_{B} = Ab_{E} + B \qquad (4)$$

$$a^{*}_{B} = a_{E}/A. \qquad (5)$$

The lower asymptote parameter is measured on the probability metric, no transformation ($c^{*}_{B}=c_{E}$) needs to be applied, although in reality, it is affected by sampling fluctuation (Li, et al, 1997). For the transformation of the GPCM item parameter estimates,

$$b^{*}_{B} = Ab_{E} + B \qquad (6)$$

$$d^{*}_{kB} = Ad_{kE} \qquad (7)$$

$$b^{*}_{KB} = Ab_{kE} + B \text{ (note: } b_{k} = b\text{-}d_{k}) \qquad (8)$$

$$a^{*}_{B} = a_{E}/A. \qquad (9)$$

After the above transformations, the original values of item parameter estimates that depend on an equated-group ability scale are re-referenced to the base-group ability scale.

C. Algorithms for Finding CCM's Equating Coefficients for Mixed-format Tests

Under the CCM equating method, the equating coefficients of A and B are chosen so that the average squared difference between the two expected true scores directly from the two sets of common-item parameter estimates for N arbitrary ability points is as small as possible.

The equating parameter estimates are relatively stable when N is larger than 100. The function below to be minimized is (refer to Hambleton, Swaminathan & Rogers, 1991):

$$f(A,B)=\frac{1}{N}\sum_{j=1}^{N}\left[t(\hat{\theta}_{jB})-t^{*}(\hat{\theta}_{jE})\right]^{2} \tag{10}$$

where

$$t(\hat{\theta}_{jB})=\sum_{i=1}^{L}\sum_{k=1}^{m}x_{ik}P(X_i=x_{ik}|\hat{\theta})\ \text{ and} \tag{11}$$

$$t^{*}(\hat{\theta}_{jE})=\sum_{i=1}^{L}\sum_{k=1}^{m}x_{ik}P(X_i=x_{ik}|\hat{\theta}) \tag{12}$$

Where, x_{ik} is the observed score of category k in item i, ranging from 0 to 1 for a dichotomous scored item or 1 to 4 (or 0 to 3) for a four-category scored item. L is the number of common items, m is the number of category scores, and the probability of the categorical response k, P( ), can be computed either from the three-parameter model (see Equation 1) for dichotomous scored items or from the GPCM model (see Equation 2) for polytomous scored items. The expected true scores of an examinee with an ability θ_j on a set of common items in the base (B) and equated (E) tests are computed in Equations 11 and 12, respectively.

The scaling coefficients of A and B that minimize the function (Equation 10) can be derived by differentiating this function with respect to A and B and by setting the two partial derivative equations equal to zero. By fixing B to a constant B_0 and considering f(A, B_0) as a one-solution function in the case of the two-solution function f(A, B) (Equation 10), the difference approximation for the partial derivative with respect to A at A_0 can derived and written as (refer to Nakamura, 1996):

$$\frac{\partial f(A,B)}{\partial A}\approx\frac{f((A_0+\Delta A),B_0)-f(A_0,B_0)}{\Delta A}=0 \tag{13}$$

where ΔA is an interval between two consecutive points on the numerical line in the A estimate. It is usually set to a very small value such as 0.001. The same principle is applied to the partial derivative with respected to B at B_0 and is given below:

$$\frac{\partial f(A,B)}{\partial B} \approx \frac{f(A_0,(B_0+\Delta B)) - f(A_0,B_0)}{\Delta B} = 0 \tag{14}$$

where ΔB is an interval between two consecutive points on the numerical line in the B estimate. The two nonlinear equations (13 and 14) can be resolved for A and B iteratively using the Newton-Raphson procedure (Baker, 1992). Within the tth iteration, the parameter estimates of the common items in the equated test are re-scaled using the tth iterative A and B solutions while computing the expected true score (see Equation 12). The iterative equation is given below.

$$\begin{bmatrix} \hat{A} \\ \hat{B} \end{bmatrix}_{t+1} = \begin{bmatrix} \hat{A} \\ \hat{B} \end{bmatrix}_{t} - \begin{bmatrix} \frac{\partial^2 f}{\partial A^2} & \frac{\partial^2 f}{\partial A \partial B} \\ \frac{\partial^2 f}{\partial B \partial A} & \frac{\partial^2 f}{\partial B^2} \end{bmatrix}^{-1} \cdot \begin{bmatrix} \frac{\partial f}{\partial A} \\ \frac{\partial f}{\partial B} \end{bmatrix}_{t} \tag{15}$$

where the difference approximations for the second derivatives of the two-dimensional function f(A, B) at (A_0, B_0) are illustrated as (refer to Nakamura, 1996):

$$\frac{\partial^2 f(A,B)}{\partial A^2} \approx \frac{f((A_0+\Delta A),B_0) - 2f(A_0,B_0) + f((A_0-\Delta A),B_0)}{\Delta A^2} \tag{16}$$

$$\frac{\partial^2 f(A,B)}{\partial B^2} \approx \frac{f(A_0,(B_0+\Delta B)) - 2f(A_0,B_0) + f(A_0,(B_0-\Delta B))}{\Delta B^2} \tag{17}$$

$$\frac{\partial^2 f(A,B)}{\partial A \partial B} \approx \frac{f((A_0+\Delta A),(B_0+\Delta B)) - f((A_0-\Delta A),(B_0+\Delta B))}{\Delta A \Delta B} + \tag{18}$$

$$\frac{-f((A_0+\Delta A),(B_0-\Delta B)) + f((A_0-\Delta A),(B_0-\Delta B))}{\Delta A \Delta B}$$

$$\frac{\partial^2 f(B,A)}{\partial B \partial A} = \frac{\partial^2 f(A,B)}{\partial A \partial B} \tag{19}$$

The starting values of A and B are critical for solving the iterative equation (15). The starting values for A and B computed by the equations given below have resulted in successful convergence in most cases.

$$A_{\text{Start Value}} = \frac{\sum_{i=1}^{L} a_{iE}}{\sum_{i=1}^{L} a_{iB}} \tag{20}$$

$$B_{\text{Start Value}} = \left(\frac{\sum_{i=1}^{L} b_{iB}}{L} \right) - A_{StratValue} \left(\frac{\sum_{i=1}^{L} b_{iE}}{L} \right) \tag{21}$$

After obtaining the equating parameters, the ability- and item- estimates derived from the different test forms will be transformed to the same scale.

D. Modeling Errors in the IRT Parameter Estimates

Standard Error Estimates From Empirical Replication Approach

The level of precision of the equating parameter estimates produced by the CCM is dependent on the magnitudes of error in the item parameter estimates. The error in each individual parameter estimate is narrowly defined as the amount of variance around the true parameter value. The magnitude of this error in the simulation study is usually manipulated by the following procedures: (1) A test dataset is generated based on two pieces of information, a set of true item parameters and a set of known simulee's ability parameters; (2) Each test item is calibrated and linked to the metric of the true item parameter; (3) Repeat steps 1 and 2 a large number of times, which results in a large number of estimates for each individual item parameter; and (4) The standard deviation of these estimates is an estimate of the standard error of this item parameter estimate. The test length and sample size usually have impact on the magnitude of the standard error of an item parameter. The above empirical approach of modeling error in the item estimate was adopted by Baker (1996; 1997) to investigate the sampling distribution of equating coefficients produced by the CCM.

Asymptotic Standard Error Estimates From Analytic Approach

The empirical replication approach of modeling errors of item estimates is very time-consuming. In contrast, the analytic approach to estimate the asymptotic standard errors of item estimates developed by Thissen and Wainer (1982) is much easier to employ for some research topics. For example, when Li and Lissitz (in press) evaluated the precision of equating coefficients produced by the three multidimensional IRT (MIRT) equating methods developed in their study, the analytic approach was adopted to manipulate the magnitudes of error in the MIRT item estimates.

Sample size, the shape of the examinees' ability distribution and the characteristic of test items can each cause differences in the errors in the parameter estimates. A mathematical expression for this relationship has been developed by Thissen and Wainer (1982) for dichotomous item response data when the model selected fits test data and the maximum likelihood approach is used to estimate item parameters. For an item i, the likelihood of the observed dichotomous responses for N independent examinees is:

$$L_D = \prod_{j=1}^{N} P_j^{u}(1-P_j)^{1-u} \qquad (22)$$

where P can be calculated from a three-parameter model, u=1 for correct response; u=0 for incorrect response. The loglikelihood of Equation 22 is

$$\log L_D = \sum_{j=1}^{N}\left[u\log(P_j) + (1-u)\log(1-P_j)\right] \qquad (23)$$

The maximum likelihood estimates of each parameter (a_i, b_i, c_i) are located where the partial derivatives of Equation 23 are zero. For ease of expression, ξ represents the three-parameter item parameters ($\underline{a}_i$, d_i, c_i). Given a density of θ (e.g. normal distribution, N(0, 1)), for any pair of parameters ξ_s and ξ_t, the negative expected value of the second derivative of the loglikelihood function, Equation, 23, has the form (refer to Thissen, Wainer, 1982),

$$-E\left(\frac{\partial^2 \log L}{\partial\xi_s\partial\xi_t}\right) = N\int_{-\infty}^{\infty}\left[\left(\frac{1}{PQ}\right)\left(\frac{\partial P(\theta)}{\partial\xi_s}\frac{\partial P(\theta)}{\partial\xi_t}\right)\right]\Phi_j(\theta)d\theta \qquad (24)$$

where E is the expectation and Q=1-P. Equation 24 requires the derivatives of P(θ) with respect to its parameters. The numerical approximation of the integral in Equation 24 can be calculated by the Gauss-Hermite quadrature and is presented in Equation 25,

$$I_{3PL}(\xi_s,\xi_t) = N\sum_{q=1}^{q}\left\{\left(\frac{1}{PQ}\right)\left(\frac{\partial P(X)}{\partial\xi_s}\frac{\partial P(X)}{\partial\xi_t}\right)\right\}A(X_q) \qquad (25)$$

where X is a quadrature point in the ability dimension, q is the number of quadrature in the ability dimension and A(X) is the corresponding weight of the quadrature. The number of quadrature points for numerical integration are set to 39 in this study. The partial derivatives of P(X) with its parameters can be resolved using difference approximation (Nakamura, 1996) and substituted in Equation 25 to give a 3 x 3 (for the three-parameter model) information matrix corresponding to the triplet item parameters (a, b, and c). The inverse of that information matrix is the asymptotic variance-covariance matrix of the three parameters and is given in Equation 26. The square roots of the diagonal elements of the variance-covariance matrix are the asymptotic standard errors of the parameters.

$$\mathrm{VarCov}_{3PL} = \begin{bmatrix} I_{3PL}(a,a) & I_{3PL}(a,b) & I_{3PL}(a,c) \\ I_{3PL}(b,a) & I_{3PL}(b,b) & I_{3PL}(b,c) \\ I_{3PL}(c,a) & I_{3PL}(c,b) & I_{3PL}(c,c) \end{bmatrix}^{-1} \qquad (26)$$

For an item i, the likelihood of the observed polytomous responses for N independent examinees is:

$$L_P = \prod_{j=1}^{N}\prod_{k=1}^{m} P_{jk}^{uk}(1-P_{jk})^{1-uk} \qquad (27)$$

where P_k can be calculated from a GPCM model, u=1 for the categorical response k; u=0 for responses other than category k. The loglikelihood of Equation 27 is

$$\log L_P = \sum_{j=1}^{N}\sum_{k=1}^{m} \left[u_k \log(P_{jk}) + (1-u_k)\log(1-P_{jk})\right] \qquad (28)$$

Similar principles used in the three-parameter model can be applied for the GPCM model to derive the information function when any pair of GPCM item parameter estimates is given. That is:

$$I_{GPCM}(\xi_s,\xi_t) = N\sum_{q=1}^{q}\left\{\sum_{k=1}^{m}\left(\frac{1}{P_k Q_k}\right)\left(\frac{\partial P_k(X)}{\partial \xi_s}\frac{\partial P_k(X)}{\partial \xi_t}\right)\right\}A(X_q) \qquad (29)$$

The inverse of that information matrix is the asymptotic variance-covariance matrix of the four parameters (a_i, b_{i2}, b_{i3},b_{i4}) and is given in Equation 30 for the case of a four-category GPCM model. The square roots of the diagonal elements of the variance-covariance matrix are the asymptotic standard errors of the parameters.

$$\mathrm{VarCov}_{\mathrm{GPCM}} = \begin{bmatrix} I_{\mathrm{GPCM}}(a,a\ ) & I_{\mathrm{GPCM}}(a,b_2) & I_{\mathrm{GPCM}}(a,b_3) & I_{\mathrm{GPCM}}(a,b_4) \\ I_{\mathrm{GPCM}}(b_2,a\ ) & I_{\mathrm{GPCM}}(b_2,b_2) & I_{\mathrm{GPCM}}(b_2,b_3) & I_{\mathrm{GPCM}}(b_2,b_4) \\ I_{\mathrm{GPCM}}(b_3,a\ ) & I_{\mathrm{GPCM}}(b_3,b_2) & I_{\mathrm{GPCM}}(b_3,b_3) & I_{\mathrm{GPCM}}(b_3,b_4) \\ I_{\mathrm{GPCM}}(b_4,a\ ) & I_{\mathrm{GPCM}}(b_4,b_2) & I_{\mathrm{GPCM}}(b_4,b_3) & I_{\mathrm{GPCM}}(b_4,b_4) \end{bmatrix}^{-1} \qquad (30)$$

III. Methodology

A. Key Variables

Several variables are critical factors in accounting for the variation of the equating coefficients. They are:

1. Proportion of Polytomous Items to Dichotomous Items

One factor is related to the proportion of dichotomous items to polytomous items. Under the circumstance of 5 four-category polytomous items to be included into the common items, three conditions of ten, fifteen and twenty dichotomous items have been created to reflect that they can provide information relatively less, equivalent, or more than 5 four-category polytomous items, respectively. These conditions were selected because of the results of the study by Donoghue (1994).

2. Sample Sizes and Standard Error

The item parameters of the simulated test were taken from the Reading/Writing test administered to grade four students at a school district. Each simulated item parameter from a set of parameters for an item was computed from the summation of the corresponding true item parameter and the expected measurement error, generated as a random value from multivariate normal distribution, MVN(**0**, **V**), where **V** is the asymptotic variance-covariance matrix corresponding to this set of item parameters (refer to Li, 1997). Matrix **V** is computed by Equation 26 or 30 for the three-parameter or GPCM model, respectively. As indicated previously, the magnitude of entries in the **V** matrix is varied by different sample sizes (e.g., N=1000, 2000 and 3000).

3. Equating Situations

Two study situations were explored. One is a parameter recovery study, where error of the parameter estimates only exists in the equated test; the other is an equating study of a real dataset, where the error of the parameter estimates exists in both the equated test and the base test.

4. Horizontal and Vertical Linking

Two types of item linking were explored. One is horizontal equating where tests measuring a common trait are administered to groups of examinees at similar ability levels and placed on a common scale. The other is vertical equating where tests measuring the same trait are administered at different ability levels (sometimes different developmental stages, e.g., 3rd grade and 6th grade) for groups of examinees and placed on a common scale.

Referring to Equation 3, we can see that, by setting B=0 and A=1, the shape of examinee's abilities from base and equated groups are the same. This condition is one of horizontal linking. In contrast, by setting B=0.5 and A=1.2, the equated group has lower mean trait score than the base group. Also, the corresponding ability variance for the equated group is lower than the base group. The value of 0.5 was chosen because groups that differ by 1/2 a S.D. are different enough to consider the problem as vertical equating and the value of 1.2 used as variance multiplier would be a reasonable level of difference in ability variances between groups (note, 1.2 was used in Baker's studies (1996, 1997). Therefore, the latter set of transformation coefficients is considered one of vertical linking.

B. Data Analysis and Evaluation of Result

Finally, there are 36 different combinations of conditions (Three proportions of the two types of items X Three sample sizes X Two study situations X Two linking situations). One thousand replications were conducted for each combination. The accuracy of the CCM's equating coefficients was assessed by using the BIAS (average differences between the true parameters and the corresponding estimates) and the RMSE (root mean square errors).

The approach of modeling standard errors of common-item parameter estimates will have impact on recovering equating coefficients. And the level of the precision on recovering equating coefficients will then affect the accuracy of BIAS or RMSE statistic for each of equating coefficients. Since the assumptions (e.g., data-model fit, the distribution of abilities is known, see Thissen & Wainer, 1982) made for estimating asymptotic standard errors of item parameter estimates adopted in this study is unlikely to be exactly true in practice, the asymptotic standard error estimates from analytic approach represent lower limits for actual standard errors (Thissen & Wainer, 1982). Accordingly, caution in interpreting the BIAS or RMSE indices produced in this study must be maintained since they can be larger if they are estimated from real test data.

Regression models regressing the Log[RMSE] (a log transformation of RMSE) of the transformation-parameter estimates on the factors of the study has been conducted to examine whether one simulation factor has significant impact on the precision of an equating coefficient when other simulation factors are held constant (Harwell, Stone, Hsu & Kirisci, 1996).It should be noted that a log transformation for the outcome variable RMSE was conducted in order to better satisfy (approximate) the normality assumption. In addition, the main reason for choosing a multiple regression method rather than an analysis of variance (ANOVA) as an inferential approach is that one of the simulation factors such as sample size is quantitative.

Since two equating situations were used in this study, an indicator variable (dummy variable) was coded for representing the equating situation (Horizontal coded as 0; Vertical coded as 1). Similarly, two study situations (Parameter Recovery coded as 0; Equating coded as 1) were separately coded as dummy variables. Meanwhile, since there were three types of the proportion of mixed-format items, two dummy variables (called as DM01 and DM02) were coded for representing the three levels, where the 5+10 was coded "0 0" as a reference. The characteristics of the empirical sampling distributions of the estimated equating coefficients obtained from various combinations of research conditions are used to help determine whether these coefficient estimates are well-behaved. If the obtained sampling distributions are symmetric-bell shaped, with small variances and few or no outliers or other abnormalities, they are considered well-behaved (Baker, 1996).

IV. Research Results and Discussions

A. Equating Parameter Recovery

The descriptive statistics of BIAS and RMSE for each of the equating coefficients under the Recovery Study situation are reported on the top half of Table 1. Similarly, the results for the Pairwise Equating situation are presented on the lower half of Table 1. The descriptive statistics are sequentially listed under the simulation factors of equating situation (Horizontal or Vertical), sample size and the proportion of polytomous items to dichotomous items. For example, the value of -0.0027, reported in the first data row and second data column of Table 1, represents the BIAS statistic of the estimate-A under the combination of parameter recovery study, horizontal linking, sample size=1000 and 5 polytomous-scored items with 10 dichotomous-scored items.

As seen in Table 1, the BIAS values produced by the CCM equating method across all combinations of conditions are close to zero as expected. The magnitudes of RMSE produced from the CCM equating method across all simulation conditions become smaller as the sample size increases. The histogram graphs for the recovery equating parameter estimates A and B for the three increasing levels of sample sizes (N=1000, 2000 and 3000) were depicted for better interpretation, for instance, under the combination of conditions, recovery study, vertical linking and TL=5+20. The histogram graphs clearly indicate that the larger the sample size, the lower the variances of A and B estimates when the rest of the conditions were held constant. The shapes of the empirical sampling distributions of the estimated equating coefficients obtained from these conditions are generally symmetric-bell shapes with small variances and few or no outliers or other abnormalities. Similar results were found for the rest of the sampling distributions of equating coefficients.

These results obtained from descriptive statistics and graphical analysis imply that the CCM method for finding equating parameters is an empirically unbiased and effective estimator when it is applied to a test with mixed-format items, as well as to dichotomous-scored test (Baker, 1996) and polytomous-scored test (Baker, 1997).

[Insert Table 1 about here]

[Insert Figures 1 and 2 about here]

B. The effect of the Proportion of Dichotomous Items to Polytomous Items on the Precision of Equating parameter estimates

The effect of the proportion of dichotomous items to polytomous items on the precision of equating parameter estimates was examined. The descriptive statistics of BIAS and RMSE listed in Table 1 did not clearly indicate that the higher number of dichotomous items, the less BIAS and RMSE statistics of equating coefficients was produced when the number of poltytomous items was held constant (5 four-category items). The histogram graphs for the recovery equating parameter estimates A and B for the three types of mixed-format items (5+10, 5+15 and 5+20), for instance, under the combination of conditions, recovery study, horizontal linking and sample size=1000, were prepared to demonstrate this phenomena. When a four-category item is equivalent to three dichotomous items in terms of computing the total raw score, the set of common items (5+10) can be considered as 25 dichotomous test items that are usually sufficient to produce stable equating coefficients (see Baker, 1996). This is one tentative reason why the set of common items, either 5+15 or 5+20, did not apparently produce less errors of equating coefficient estimates. Further analyses presented below will provide more defensible reasons.

[Insert Figures 3 and 4 about here]

Regression models regressing the Log[RMSE] of the transformation-parameter estimates on the factors of the study have been conducted to examine whether the factor of the proportion of dichotomous items to polytomous items has significant impact on the precision of an equating coefficient with other simulation factors held constant. The standardized regression coefficients of the two dummy variables of this factor were statistically significant from zero, indicating that this factor has significant impact on the precision of an equating coefficient. As the mixed-format items of 5+10 was coded the " 0 0", as indicated previously, the positive regression coefficients of dummy variables showed that the mixed-format items of 5+15 and 5+20 produced more RMSE of equating coefficients of A and B. This is an unexpected finding.

When the number of the polytomous-scored items was held constant, why did the test with higher number of dichotomous items not produce lower BIAS and RMSE statistics of equating coefficients? The search for the reasons goes back to examine the method of modeling errors in the IRT parameter estimates. As illustrated previously, Equation 26 was used to generate the standard error estimates of the three-parameter item parameter estimates. Using this equation, the average standard errors of item-discrimination, difficulty and guessing parameters for the set of 15 or 20 dichotomous-scored items were larger than those from the set of 10 items (see Table 3). The unintended design effect of the larger standard errors obtained from the two long test lengths rather than the short test length might cause the two long tests to produce more true-score (see Equation 11 or 12) variance. This effect will subsequently cause the two long tests to produce less precision of the CCM-based recovery equating coefficients. This unintended design effect seemed to play a more substantive role in producing the variance of the recovery equating coefficients than the factor of the number of common items. The question of determining the impact of weighting the proportion of different item formats on the precision of equating coefficients has been evaluated but not been clearly resolved in this study.

C. The effect of the Simulation Factors on the Precision of Equating parameter estimates

Separate regression analyses were performed to predict Log[RMSE] in each of the transformation-parameter estimates (A and B) by fitting those simulation factors. The adjusted R^2 and standardized regression coefficients for each predictor in the regression model are presented in Table 2.

Regarding the regression model of the Log[RMSE] of A or B estimate, all the standardized regression coefficients were statistically significant from zero. These results indicate that each simulation factor chosen in this study has significant impact on the precision of an equating coefficient with other simulation factors held constant.

The results of adjusted R^2s for each of the Log[RMSE] models, ranging from 0.983 to 0.986, reported on the right side of Table 2, suggest that this set of simulation factors was very sensitive to variations in each of the transformation-parameter estimates. Thus, the accuracy of estimating each of the transformation-parameter estimates appeared to depend heavily on these simulation factors.

Several specific findings were highlighted in Table 2. The pairewise equating study produced more RMSE of equating coefficients and the larger sample size produced less RMSE of equating coefficients. The vertical equating study could produce more RMSE of equating coefficient A and less RMSE of equating coefficient B. The latter result was unusual partly because the true value of B in this study was set to zero which is usually relatively hard to estimate so that its variance could be relatively large.

V. Summary and Conclusion

The general issue of whether the CCM produce accurate transformation of parameter estimates when applied to a test with mixed items was closely examined under several simulation conditions in this study. The BIAS statistics of the recovery equating coefficients across all research conditions are usually expected to be close to zero, as happened here, and the corresponding RMSE statistic, in general, is relatively small. Caution in interpreting the BIAS or RMSE indices produced in this study must be maintained since they can be larger if they are estimated from real test data.

The shapes of the empirical sampling distributions of the estimated equating coefficients obtained from these conditions are symmetric-bell shapes with small variances and few or no outliers or other abnormalities.

The relative importance of determining the equating coefficients under CCM for each type of item partly depends on the proportion of dichotomous items to polytomous items. This issue was explored by comparing three types of mixed-format items as described in the section of methodology. Under the circumstance of 5 four-category polytomous items to be included into the common items, the higher number of dichotomous items did not produce lower BIAS and RMSE statistics of equating coefficients. This finding was unexpected. The unintended design effect that the 15 or 20 dichotomous-scored items used in this study had larger average measurement errors than the 10 dichotomous-scored items may cause this to occur. We are wondering whether this unintended design effect will also be generated by the empirical-replication approach to modeling measurement errors of the same set of item parameters used in this study. On the whole, the question of determining the impact of weighting the proportion of different item formats on the precision of equating coefficients was not clearly resolved in this study. This question needs clarification by manipulating different research conditions such as using the optimal design of tests with mixed-format items (e.g., Berger, 1998), choosing different sets of item parameters or composing different types of the proportion of mixed-format items.

Although the impact of the item-parameter characteristics on producing equating coefficients is not a subject of this paper, the unexpected result may remind test practitioners that the item-parameter characteristics of the set of common items used for item linking might play a more significant role in producing the precision of equating coefficients than the factor of the number of common items.

Regarding the regression model of the Log[RMSE] of A or B estimate, all the standardized regression coefficients were statistically significant from zero. The adjusted R^2s for the two regression models were very high (.983 and .986). These results suggest that this set of simulation factors was very sensitive to variations in each of the transformation-parameter estimates and the accuracy of estimating each of transformation-parameter estimates appeared to depend heavily on these simulation factors.

One of the more interesting unanswered questions has to do with modeling measurement errors from empirical replications used in Baker's studies (1996 & 1997). Can we obtain similar results under the same conditions? Whether these findings can be generalized to equating two real test data sets is an important question to be investigated at some future time.

Figure Headings

Figure 1. Comparison of Sampling Distributions of the Equating Coefficient A for Three Types of Sample Size under a Specific Combination of Research Conditions

Figure 2. Comparison of Sampling Distributions of the Equating Coefficient B for Three Types of Sample Size under a Specific Combination of Research Conditions

Figure 3. Comparison of Sampling Distributions of the Equating Coefficient A for Three Types of Mixed-format Items under a Specific Combination of Research Conditions

Figure 4. Comparison of Sampling Distributions of the Equating Coefficient B for Three Types of Mixed-format Items under a Specific Combination of Research Conditions

References

Baker, F. G. (1992). Equating tests under the grade response model. Applied Psychological Measurement, 16, 87-96.

Baker, F. B. (1996). An investigation of the sampling distributions of equating coefficients. Applied Psychological Measurement, 20, 45-57.

Baker, F. B. (1997). Empirical sampling distributions of equating coefficients for graded and nominal response instruments. Applied Psychological Measurement, 21, 157-172.

Beaton, A. E. & Zwick, R. (1992). Overview of the national assessment of educational progress. Journal of Educational Statistics, 17, 95-109.

Bennett, R. E., Rock, D. A., Wang, M. (1991). Equivalence of free-response and multiple-choice items. Journal of Educational Measurement, 28, 77-92.

Berger, M. P. F. (1998). Optimal design of tests with dichotomous and polytomous items. Applied Psychological Measurement, 22, 248-258.

David, T. Wainer, H. & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. Journal of Educational Measurement, 31, 113-123.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. Applied Psychological Measurement, 9, pp. 413-415.

Donoghue, J. R. (1992). An empirical examination of the IRT information of polytomous scored reading items under the generalized partial credit model. Journal of Educational Measurement, 31, 295-311.

Hambleton, R. K. & Swaminathan, H. & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park: CA. SAGE Publications, Inc.

Harwell, M. R., Stone, C. A., Hsu, T. , & Kirisci, L. (1996). Monte carlo studies in item response theory. Applied Psychological Measurement, 20, 101-125.

Kim, S. & Cohen, A. S. (1995). A minimum χ^2 method for equating tests under the graded response model. Applied Psychological Measurement, 19, 167-176.

Kim, S. & Cohen, A. S. (1997, March). A comparison of linking and concurrent calibration under the graded response model. Paper presented at the annual meeting of the American Educational Research Association, IL: Chicago.

Kim, S. & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. Applied Psychological Measurement, 22, 131-143.

Kolen, M. J. & Brennan, R. L. (1995). Test equating: Methods and practices. New York: Springer-Verlag.

Li, Y. H. (1997). An evaluation of multidimensional IRT equating methods by assessing the accuracy of transforming parameters onto a target test metric. Applied Psychological Measurement. Unpublished Doctoral Dissertation. University of Maryland at College Park.

Li, Y. H., Griffith, W. D. & Tam, H. P. (1997, June). Equating multiple tests via an IRT linking design: Utilizing a single set of Anchor items with fixed common item parameters during the_calibration process. Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Li, Y. H. & Lissitz, R. W. (in press). An evaluation of multidimensional IRT equating methods by assessing the accuracy of transforming parameters onto a target test metric. Applied Psychological Measurement.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. New Jersey: Lawrence Erlbaum Associates, Inc.

Maryland State Department of Education (1997). Technical report: 1997 Maryland School Performance Assessment Program. Baltimore: Author.

McKinley, R. L. & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 9, 49-57.

Mislevy, R. J. & Bock R. D. (1982, July). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In: Item Response Theory and Computerized Adaptive Testing Conference Proceedings (Wayzata, MN).

Mislevy, R. J. & Bock, R. D. (1990). BILOG-3: Item analysis and test scoring with binary logistic models. Mooresvilk: Scientific Software.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm.

Applied Psychological Measurement, 16, 159-176.

Muraki, E. & Bock, R. D. (1996). PARSCALE (Version 3.): IRT based test scoring and item analysis for graded open-ended exercises and performance tasks. Mooresvilk: Scientific Software.

Nakamura, S. (1996). Numerical analysis and graphic visualization with MATLAB. Upper Saddle River, NJ: Prentice-Hall, Inc.

Nandakumar, R., Yu, F., Li, H. & Stout, W. (1998). Assessing unidimensionality of polytomous data. Applied Psychological Measurement, 22, 99-115.

Tam, H. P. & Li, Y. H. (1997, March). Is the use of the difference likelihood ratio chi-square statistic for comparing nested IRT models justifiable? Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.

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

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. Applied Psychological Measurement, 14, 127-137.

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

Vale, C. D. (1986). Linking item parameters onto a common scale. Applied Psychological Measurement, 10, 333-344.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187-213.

Table 1
BIAS and RMSE Statistics of Equating Coefficients for Two Equating Contexts (Recovery Study and Pairwise Equating), Two Equating Situations (Horizontal and Vertical), Three Sample Sizes and Three Types of Mixed-format Items (Replications=1000)

		True$_A$	BIAS$_A$	RMSE$_A$	True$_B$	BIAS$_B$	RMSE$_B$
Recovery Study							
Horizontal							
N=1000	TL=5+10	1.0000	-.0027	.0306	.0000	-.0014	.0449
	TL=5+15	1.0000	-.0076	.0344	.0000	.0025	.0521
	TL=5+20	1.0000	-.0084	.0371	.0000	.0005	.0511
N=2000	TL=5+10	1.0000	-.0001	.0212	.0000	-.0011	.0309
	TL=5+15	1.0000	-.0059	.0245	.0000	.0023	.0367
	TL=5+20	1.0000	-.0061	.0254	.0000	.0018	.0348
N=3000	TL+5+10	1.0000	-.0015	.0166	.0000	.0003	.0245
	TL=5+15	1.0000	-.0017	.0206	.0000	-.0008	.0318
	TL=5+20	1.0000	-.0039	.0222	.0000	.0002	.0302
Vertical							
N=1000	TL=5+10	1.2000	-.0036	.0352	.5000	-.0025	.0305
	TL=5+15	1.2000	-.0081	.0412	.5000	-.0006	.0373
	TL=5+20	1.2000	-.0122	.0418	.5000	-.0006	.0374
N=2000	TL=5+10	1.2000	-.0027	.0260	.5000	-.0005	.0212
	TL=5+15	1.2000	-.0045	.0300	.5000	.0002	.0272
	TL=5+20	1.2000	-.0065	.0307	.5000	-.0007	.0257
N=3000	TL=5+10	1.2000	-.0011	.0198	.5000	-.0006	.0179
	TL=5+15	1.2000	-.0037	.0226	.5000	.0007	.0215
	TL=5+20	1.2000	-.0030	.0251	.5000	-.0010	.0210
Pairwise Equating							
Horizontal							
N=1000	TL=5+10	1.0000	-.0011	.0427	.0000	.0019	.0619
	TL=5+15	1.0000	.0012	.0512	.0000	.0005	.0746
	TL=5+20	1.0000	.0022	.0509	.0000	-.0021	.0714
N=2000	TL=5+10	1.0000	-.0010	.0302	.0000	.0005	.0449
	TL=5+15	1.0000	.0006	.0345	.0000	-.0013	.0535
	TL=5+20	1.0000	.0004	.0361	.0000	.0010	.0504
N=3000	TL=5+10	1.0000	.0008	.0240	.0000	-.0011	.0344
	TL+5+15	1.0000	.0017	.0286	.0000	-.0016	.0436
	TL=5+20	1.0000	-.0002	.0295	.0000	.0022	.0401
Vertical							
N=1000	TL=5+10	1.2000	.0042	.0524	.5000	-.0013	.0448
	TL=5+15	1.2000	.0065	.0605	.5000	-.0026	.0554
	TL=5+20	1.2000	.0048	.0608	.5000	-.0016	.0503
N=2000	TL=5+10	1.2000	.0023	.0366	.5000	-.0007	.0312
	TL=5+15	1.2000	.0042	.0426	.5000	-.0018	.0388
	TL=5+20	1.2000	.0007	.0439	.5000	.0009	.0373
N=3000	TL=5+10	1.2000	.0002	.0293	.5000	-.0002	.0254
	TL=5+15	1.2000	.0018	.0342	.5000	-.0001	.0316
	TL=5+20	1.2000	.0030	.0353	.5000	-.0012	.0294

Table 2
The Standardized Regression Coefficients and the Adjusted R^2 in Each Regression Model Using those Simulation Factors to Predict the Log[RMSE] of the Equating Parameter Estimate

	Predictors					Adjusted R^2
Outcome	Type of Mixed Item Format		Study Situation	Equating Situation	Sample Size	
Log[RMSE]						
	DM1	DM2				
A	.230*	.288*	.564*	.275*	-.723*	.983
B	.281*	.213*	.510*	-.481*	-.657*	.986
	(5+15 VS 5+10)	(5+20 VS 5+10)				

Table 3
The Average Value of Item Parameters for the Set of 10, 15 or 20 Dichotomous-scored Items and the Corresponding Average Standard Error Estimate (N=2000)

Item Length	Mean a	Average Standard Error	Mean b	Average Standard Error	Mean c	Average Standard Error
10	.8819	.0845	-.3445	.1347	.1172	.0641
15	.8606	.0858	-.2025	.1517	.1278	.0645
20	.8902	.0866	-.3272	.1489	.1309	.0674

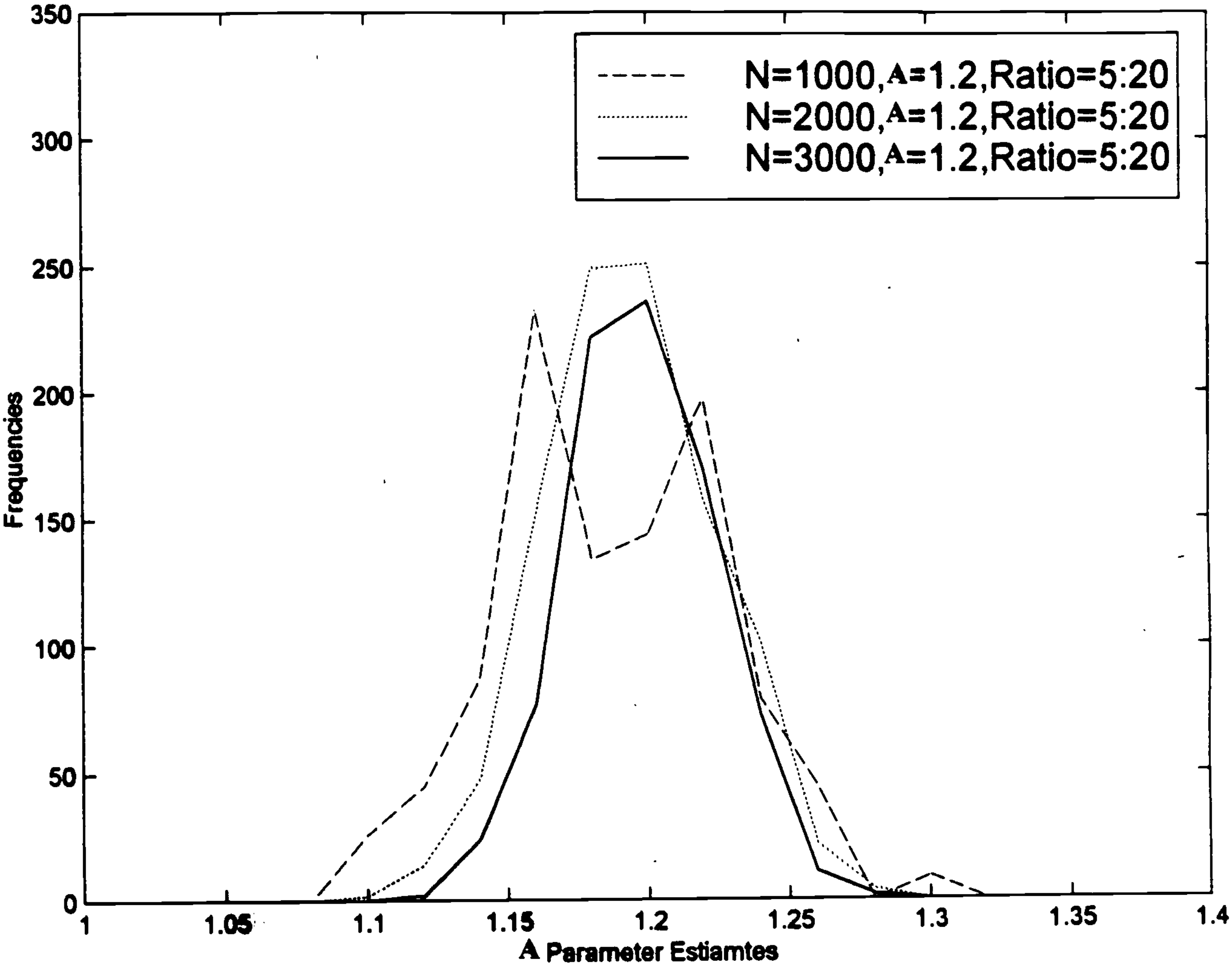

Figure 1. Comparison of Sampling Distributions of the Equating Coefficient A for Three Types of Sample Size under a Specific Combination of Research Conditions

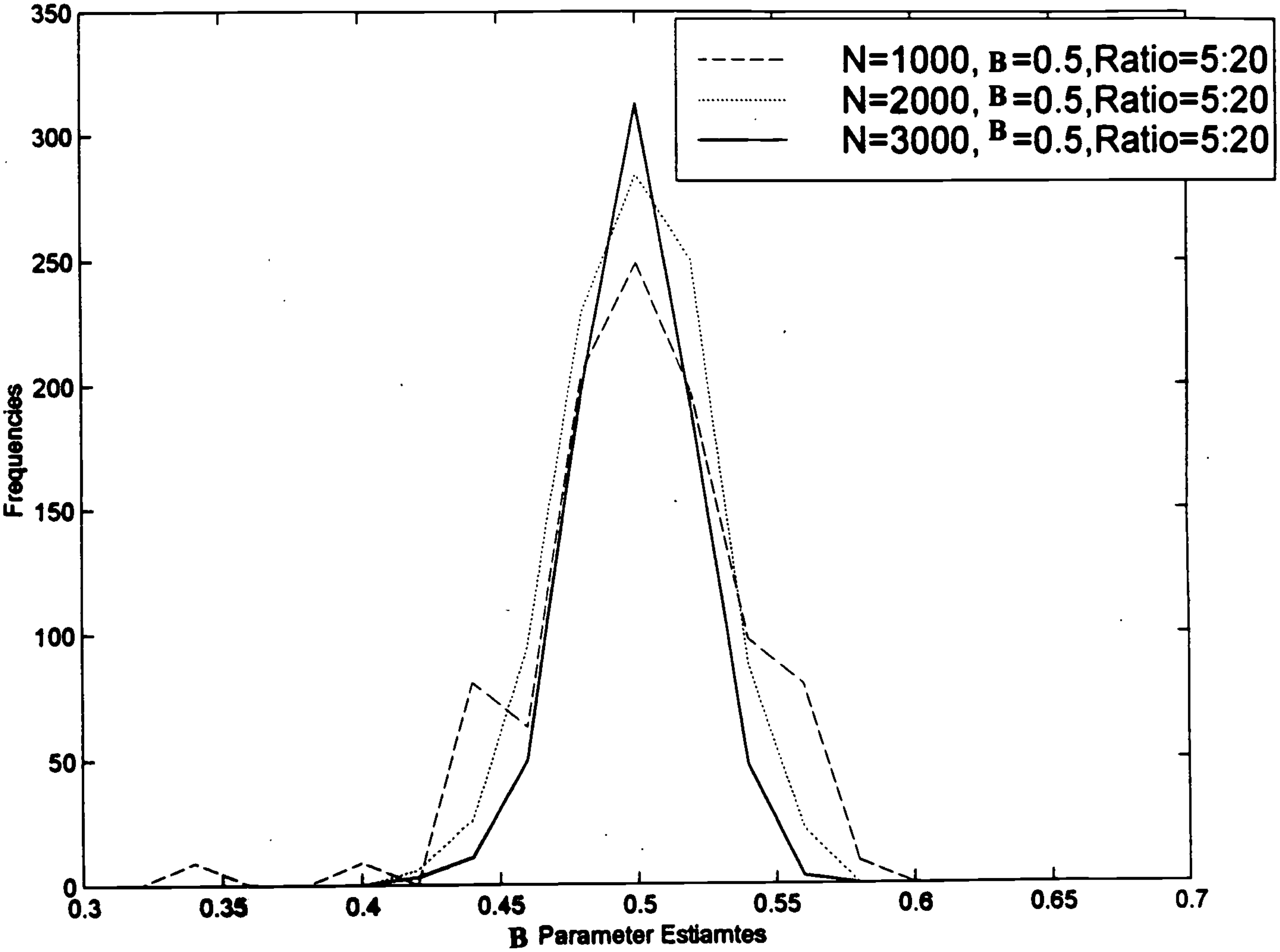

Figure 2. Comparison of Sampling Distributions of the Equating Coefficient B for Three Types of Sample Size under a Specific Combination of Research Conditions

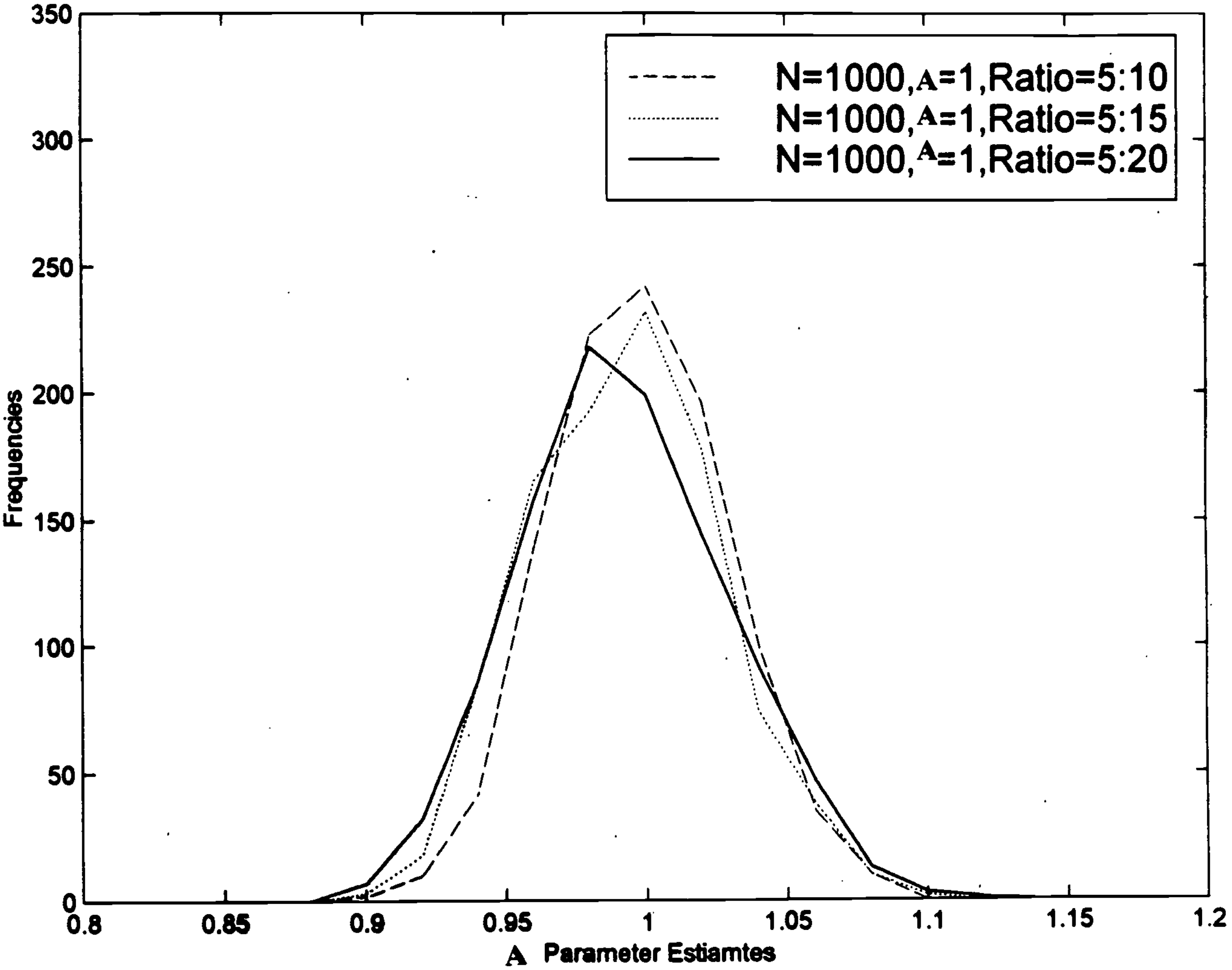

Figure 3. Comparison of Sampling Distributions of the Equating Coefficient A for Three Types of Mixed-format Items under a Specific Combination of Research Conditions

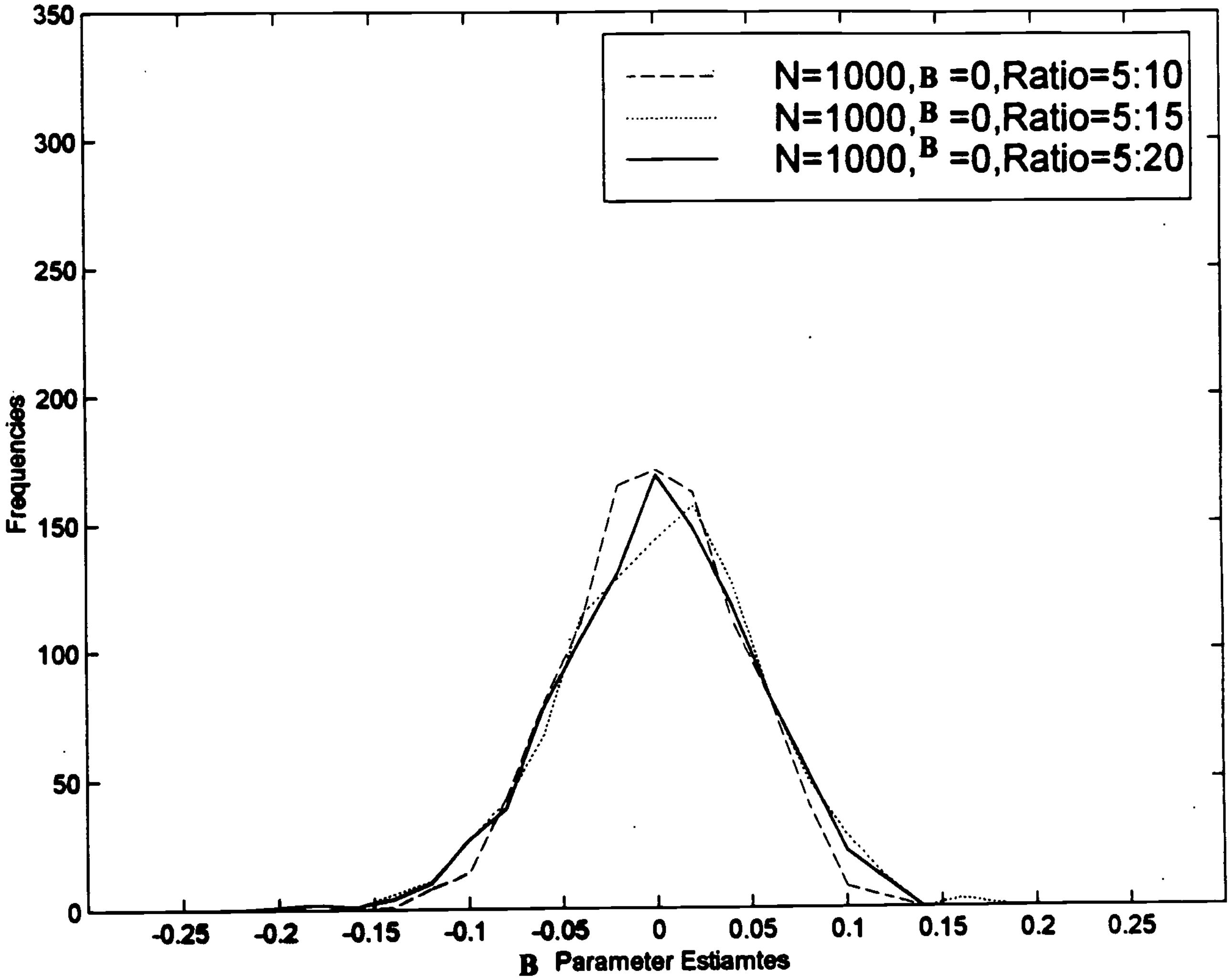

Figure 4. Comparison of Sampling Distributions of the Equating Coefficient B for Three Types of Mixed-format Items under a Specific Combination of Research Conditions

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM029886

REPRODUCTION RELEASE

(Specific Document)

NCME

I. DOCUMENT IDENTIFICATION:

Title: Estimating IRT Equating Coefficients for Tests with polytomous and Dichotomous Scored Items

Author(s): Yuan H. Li; Robert W. Lissitz & Yu Nu YANG

Corporate Source:

Publication Date: 4/99

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be p
If permission to reproduce is gr

I hereby grant to the Educational Resources Inform
as indicated above. Reproduction from the ERI
contractors requires permission from the copyright
to satisfy information needs of educators in respc

Name: Yuan H. LI
Address: Prince George's County Public Schools
Room 205
Upper Marlboro, MD. 20772
Tel: 301-952-6764
Fax: 301-952-6228
Email: jeffli@pgcps.org

Sign here,→ please

Signature:	Printed Name/Position/Title:	
Organization/Address:	Telephone:	FAX:
	E-Mail Address:	Date: 6/14/99

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com