Equating Tests Under The Nominal Response Model

Frank B. Baker

University of Wisconsin

Under item response theory, test equating involves finding the coefficients of a linear transformation of the metric of one test to that of another. A procedure for finding these equating coefficients when the items in the two tests are nominally scored was developed. A quadratic loss function based on the differences between response category probabilities in the two tests is employed. The gradients of this loss function needed by the iterative multivariate search procedure used to obtain the equating coefficients were derived for the nominal response case. Examples of both horizontal and vertical equating are provided. The empirical results indicated that tests scored under a nominal response model can be placed on a common metric in both horizontal and vertical equatings. *Index terms: characteristic curve, equating, item response theory, nominal response model, quadratic loss function.*

Placing two or more sets of test results on a common metric is required in a variety of situations of both practical and theoretical interest. Under item response theory (IRT), test equating involves finding the coefficients of a linear transformation of the metric of one test to that of another (Lord, 1980). The basic metric transformation is given by

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

where

A is the slope,

K is the intercept,

θ_i is the examinee's trait level parameter in the metric of the current test, and

θ_i^* is θ_i expressed in the target test metric.

There is no standard for labeling the tests involved in an equating. Thus, the term "target test metric" is used here to identify the metric into which the current test results will be transformed. Using the coefficients A and K, the values of the current test's item difficulty (b) and discrimination (a) parameters can be transformed to the target test metric by

$$b_j^* = Ab_j + K \tag{2}$$

and

$$a_j^* = a_j/A , \tag{3}$$

where $j = 1, 2, \ldots, J$ indexes the test items.

Let the probability of a correct response to a target test item be denoted by $P_j(\theta_i)$, and that for a transformed current test be $P_j^*(\theta_i)$. Then a perfect equating implies that $[P_j(\theta_i) - P_j^*(\theta_i)] = 0$ holds over the target test trait scale for the items common to the two tests, that is, the anchor items. Thus, the task is to find the values of the equating coefficients A and K that meet this criterion as closely as possible.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 3, September 1993, pp. 239-251

0146-6216/93/030239-13$1.95

A variety of equating procedures are available; however, the "characteristic curve" procedure introduced by Haebara (1980) for dichotomously scored items, and reformulated by Stocking & Lord (1983) appears to be the preferred approach (see Baker & Al-Karni, 1991). Under this approach, equating coefficients are found that minimize a quadratic loss function based on the difference between the test characteristic curves or item response functions yielded by the anchor items common to the two tests. The test characteristic curve (TCC) approach has been extended by Baker (1992) to include Samejima's (1969, 1972) graded response model. The present paper develops a similar procedure to equate tests that have been scored under Bock's (1972) nominal response model.

The Characteristic Curve Approach

The Dichotomous Response Case

Under a two-parameter IRT model

$$P_j(\theta_i) = 1/\{1 + \exp[-a_j'(\theta_i - b_j')]\} \tag{4}$$

and

$$P_j^*(\theta_i) = 1/\{1 + \exp[-a_j^*(\theta_i - b_j^*)]\} , \tag{5}$$

where a_j' and b_j' are the parameters of the anchor items from the target test calibration, and a_j^* and the b_j^* are the result of applying Equations 2 and 3 to the anchor item parameter estimates yielded by the current test calibration.

Starting from the basic difference of interest $[P_j(\theta_i) - P_j^*(\theta_i)]$, the quadratic loss function for the dichotomous response case can be defined as:

$$F = \frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{j=1}^{n}[P_j(\theta_i) - P_j^*(\theta_i)]\right\}^2 , \tag{6}$$

where $i = 1, 2, \ldots, N$ indexes the N arbitrary points employed over the θ scale, and $j = 1, 2, \ldots, n$ indexes the anchor items common to the two tests, where $n \leq j$. Because true scores are defined as sums of probabilities of correct responses, distributing the summation over items across the terms inside the brackets of Equation 6 yields:

$$T_i = \sum_{j=1}^{n} P_j(\theta_i) \tag{7}$$

and

$$T_i^* = \sum_{j=1}^{n} P_j^*(\theta_i) , \tag{8}$$

where T_i is the true score at θ_i based on the anchor items in the target test, and T_i^* is the true score at θ_i based on the anchor items in the current test after transformation of the item parameters.

The resultant quadratic loss function is (Stocking & Lord, 1983):

$$F = \frac{1}{N}\sum_{i=1}^{N}(T_i - T_i^*)^2 . \tag{9}$$

Because the goal is to express the current test results in the target test's metric, the θ_i appearing in both T_i (Equation 7) and T_i^* (Equation 8) must be in a common metric. Typically, this metric is that

of the target test; however, an arbitrary metric could be used. The majority of IRT test analysis computer programs solve the identification problem by standardizing the θ distribution. Thus, a convenient metric has a midpoint of 0 and a standard deviation (SD) of 1. The task then is to find the equating coefficients that will minimize the quadratic loss function of Equation 9 in the common 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 equations does not have a closed-form solution. Stocking & Lord (1983) employed 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 iterative technique requires that the derivatives (gradients) of F with respect to A and K be evaluated at each primary iteration. The gradients are given by

$$\frac{\partial F}{\partial A} = \frac{2}{N}\sum_{i=1}^{N}(T_i - T_i^*)\left\{-\left[\sum_{j=1}^{n} a_j^*\left(\frac{\theta_i - b_i^*}{A}\right)W_{ij}^*\right] + \left(\sum_{j=1}^{n} a_j^* W_{ij}^* b_j\right)\right\} \tag{10}$$

and

$$\frac{\partial F}{\partial A} = \frac{2}{N}\sum_{i=1}^{N}(T_i - T_i^*)\left(\sum_{j=1}^{n} a_j^* W_{ij}^*\right), \tag{11}$$

where

$$W_{ij}^* = P_j^*(\theta_i)Q_j^*(\theta_i) \tag{12}$$

and $Q_j^*(\theta_i) = 1 - P_j^*(\theta_i)$.

The process for obtaining the equating coefficients, based on true scores, has been implemented in the EQUATE computer program (Baker, Al-Karni, & Al-Dosary, 1991). It solves Equations 10 and 11 by using a set of six subroutines, taken from *Numerical Recipes* (Press, Flannery, Teukolsky, & Vetterling, 1986), that implement an improved version of the Davidon-Fletcher-Powell iterative multivariate search technique.

The Graded Response Case

Under Samejima's graded response model (Samejima, 1969, 1972), an item possesses m_j ordered response categories, such as in a Likert scale; the examinee can select only one of the categories and each category has a response weight associated with it. Item parameter estimation uses $m_j - 1$ boundary curves, which represent the cumulative probability of selecting response categories greater than and including the response category of interest (Samejima, 1969). The several boundary curves of a given item are characterized by a single item discrimination parameter, a_j, and by $m_j - 1$ location parameters, b_{jk}. The b_{jk} for an item are ordered, typically, from low ($k = 1$) to high ($k = m_{j-1}$). Hence, the probability of selecting response category k of target test item j is given by

$$P_{jk}(\theta_i) = \hat{P}_{j,k-1}(\theta_i) - \hat{P}_{jk}(\theta_i) \tag{13}$$

when $1 < k < m_j$,

$$P_{j1}(\theta_i) = 1 - \hat{P}_{j1}(\theta_i) \tag{14}$$

when $k = 1$, k
and

$$P_{j,m_j}(\theta_i) = \hat{P}_{j,m_j-1}(\theta_i) \tag{15}$$

when $k = m_j$,
where $\hat{P}_{jk}(\theta_i)$ are the cumulative probabilities obtained from the boundary curves.

When a test is being equated into the target test metric, $P^*_{jk}(\theta_i)$ is the probability of selecting response category k for item j after transformation of the item parameters a_j and b_{jk} using Equations 2 and 3. This probability also can be expressed in terms of transformed boundary curves. Let

$$\tilde{P}_{jk}(\theta_i) = 1/\{1 + \exp[-a^*_j(\theta_i - b^*_{jk})]\} . \tag{16}$$

Then,
for $1 < k < m_j$,

$$P^*_{jk}(\theta_i) = \tilde{P}_{j,k-1}(\theta_i) - \tilde{P}_{jk}(\theta_i) ; \tag{17}$$

when $k = 1$,

$$P^*_{j1}(\theta_i) = 1 - \tilde{P}_{j1}(\theta_i) ; \tag{18}$$

and
when $k = m_j$,

$$P^*_{jm}(\theta_i) = \tilde{P}_{j,m-1}(\theta_i) . \tag{19}$$

Using the difference of interest $[P_{jk}(\theta_i) - P^*_{jk}(\theta_i)]$, the quadratic loss function for the graded response case can be defined as

$$F = \frac{1}{N} \sum_{i=1}^{\hat{N}} \left\{ \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} [P_{jk}(\theta_i) - P^*_{jk}(\theta_i)] \right\}^2 , \tag{20}$$

where u_{jk} is the weight allocated to the response category. Typically, the numerical value of the weight is the same as the integer index of the response category. In the graded response model, the true score is defined as

$$T_i = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} P_{jk}(\theta_i) . \tag{21}$$

Distributing the summations over the response categories and over the items across the terms inside the brackets in Equation 20 yields a quadratic loss function, expressed in terms of differences in true scores, that is the same as Equation 9 for the dichotomous case.

For the purpose of taking derivatives, T_i based on the target test can be defined in terms of the boundary curves:

$$T_i = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} P_{jk}(\theta_i) = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} [\hat{P}_{j,k-1}(\theta_i) - \hat{P}_{jk}(\theta_i)] . \tag{22}$$

Once the item parameters of the current test are transformed into the target test metric, a true score is given by

$$T^*_i = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} P^*_{jk}(\theta_i) = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} [\tilde{P}_{j,k-1}(\theta_i) - \tilde{P}_{jk}(\theta_i)] . \tag{23}$$

The Davidon-Fletcher-Powell minimization technique requires that the derivatives (gradients), with respect to A and K, of the quadratic loss function of Equation 20 be evaluated at each primary iteration. Baker (1992) has shown that, on substituting the derivatives of the true scores with respect to

A and K, the following gradients for the graded response case are obtained:

$$\frac{\partial F}{\partial A} = \frac{-2}{N} \sum_{i=1}^{N} (T_i - T_i^*) \left\{ \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} [a_j^* (\tilde{w}_{jk} y_{jk} - \tilde{w}_{j,k-1} y_{j,k-1})] \right\} \tag{24}$$

and

$$\frac{\partial F}{\partial K} = \frac{-2}{N} \sum_{i=1}^{N} (T_i - T_i^*) \left\{ \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} [a_j^* (\tilde{w}_{jk} - \tilde{w}_{j,k-1})] \right\} , \tag{25}$$

where

$$\tilde{w}_{j,k-1} = \tilde{P}_{j,k-1}(\theta_i) \tilde{Q}_{j,k-1}(\theta_i) , \tag{26}$$

$$\tilde{w}_{jk} = \tilde{P}_{jk}(\theta_i) \tilde{Q}_{jk}(\theta_i) , \tag{27}$$

$$y_{j,k-1} = [b_{j,k-1} + (\theta_i - b_{j,k-1}^*) / A] , \tag{28}$$

and

$$y_{jk} = [b_{jk} + (\theta_i - b_{jk}^*) / A] . \tag{29}$$

Using these gradients based on true scores in the Davidon-Fletcher-Powell procedure, EQUATE has been extended to produce values of A and K for the graded response case.

A Procedure for Equating Nominally Scored Tests

Under the nominal response model (Bock, 1972), the multivariate logistic function is used to simultaneously model the category response functions (CRFs; also referred to as operating characteristic curves or trace lines) for the m_j response categories of an item. This function is given by

$$P_{jv}(\theta_i) = \frac{\exp[c_{jv} + d_{jv}(\theta_i)]}{\sum_{k=1}^{m_j} \exp[c_{jk} + d_{jk}(\theta_i)]} , \tag{30}$$

where $k = 1, 2, \ldots, m_j$ indexes the item's response categories,
v indexes the response category of interest,
c_{jk} is the intercept parameter for response category k of item j, and
d_{jk} is the slope parameter for response category k of item j.
Let $Z_{jk}(\theta_i) = c_{jk} + d_{jk}\theta_i$ be the multivariate logistic deviate for item response category k of item j at θ_i.

Bock (1972) resolved the identification problem by imposing the following restriction on each item separately:

$$\sum_{k=1}^{m_j} Z_{jk}(\theta_i) = 0 , \tag{31}$$

which also implies that

$$\sum_{k=1}^{m_j} c_{jk} = 0 \tag{32}$$

and

$$\sum_{k=1}^{m_j} d_{jk} = 0 \tag{33}$$

hold at the individual item level.

It is important to note here that Bock (1972) employed a slope-intercept parameterization for the item response categories rather than the usual difficulty-discrimination form. In order to obtain the A and K equating coefficients of Equations 1, 2, and 3, an equivalence between the two parameterizations must be established. Let

$$z_{jk}^*(\theta_i) = c_{jk}^* + d_{jk}^*\theta_i = a_{jk}^*(\theta_i - b_{jk}^*) \ . \tag{34}$$

Then by parts

$$d_{jk}^* = \frac{d_{jk}}{A} \tag{35}$$

and

$$c_{jk}^* = -a_{jk}b_{jk} - \frac{a_{jk}}{A} K \ . \tag{36}$$

But $c_{jk} = -a_{jk}b_{jk}$ and $d_{jk} = a_{jk}$. Substituting terms yields

$$c_{jk}^* = c_{jk} - \frac{d_{jk}}{A} K = c_{jk} - d_{jk}^* K \ . \tag{37}$$

Under the difficulty-discrimination parameterization, Equation 2 uses the equating coefficient K to reposition the items along the θ scale by a constant amount after the difficulties are adjusted for the change in scale. However, under the slope-intercept parameterization, the intercept cannot be shifted up and down its scale without violating the restrictions in Equations 32 and 33. As a result, the d_{jk}^*K term in Equation 37 takes into account the effect of the slope on the location parameter. Thus, in Equation 36 the value of the intercept parameter is changed, but the restrictions in Equations 31 and 32 are maintained.

One of the unique features of the nominal response model is that although θ can be estimated for an examinee, there is no intrinsic ordering to the m_j response categories; hence, the examinee does not possess a true score. This poses a major impediment to employing Stocking & Lord's (1983) TCC method for obtaining the equating coefficients. Recall that the fundamental element in the quadratic loss function for the graded response case was $[P_{jk}(\theta_i) - P_{jk}^*(\theta_i)]$. In order to define the loss function in terms of the TCCs, this term was summed first over response categories and then over items before being squared. Under the nominal response model, such a procedure is not admissible due to the lack of a true score. However, following Haebara (1980), the loss function can be redefined such that the term $[P_{jk}(\theta_i) - P_{jk}^*(\theta_i)]$ is squared before the summation over response categories and items is performed. Thus, the quadratic loss function for the nominal response case can be defined as

$$F = \frac{1}{NS} \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{m_j} [P_{jk}(\theta_i) - P_{jk}^*(\theta_i)]^2 \ , \tag{38}$$

where

$$S = \sum_{j=1}^{n} m_j \ . \tag{39}$$

The gradients with respect to A and K are

$$\frac{\partial F}{\partial A} = \frac{-2}{NS} \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{m_j} [P_{jk}(\theta_i) - P_{jk}^*(\theta_i)] \left[\frac{\partial}{\partial A} P_{jk}^*(\theta_i) \right] \tag{40}$$

and

$$\frac{\partial F}{\partial k} = \frac{-2}{NS} \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{m_j} [P_{jk}(\theta_i) - P_{jk}^*(\theta_i)] \left[\frac{\partial}{\partial K} P_{jk}^*(\theta_i) \right] . \tag{41}$$

The derivatives of $P_{jk}^*(\theta_i)$ with respect to A and K are obtained using the chain rule. For A,

$$\frac{\partial P_{jk}^*(\theta_i)}{\partial A} = P_{jk}^*(\theta_i) Q_{jk}^*(\theta_i) d_{jk}^* \left[\frac{(K - \theta_i)}{A} \right] \tag{42}$$

and for K

$$\frac{\partial P_{jk}^*(\theta_i)}{\partial K} = -d_{jk}^* P_{jk}^*(\theta_i) Q_{jk}^*(\theta_i) . \tag{43}$$

Let

$$W_{jk}^* = P_{jk}^*(\theta_i) Q_{jk}^*(\theta_i) . \tag{44}$$

Then the gradients for the nominal response case are

$$\frac{\partial F}{\partial A} = \frac{2}{NS} \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{m_j} (P_{jk} - P_{jk}^*) d_{jk}^* W_{jk}^* \left(\frac{\theta_i - K}{A} \right) \tag{45}$$

and

$$\frac{\partial F}{\partial A} = \frac{2}{NS} \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{m_j} (P_{jk} - P_{jk}^*) d_{jk}^* W_{jk}^* . \tag{46}$$

An extended version of EQUATE 2.0 (Baker, 1993) implements these gradients so that the Davidon-Fletcher-Powell procedure can be used to obtain the values of the equating coefficients.

Examples of the Equating Procedure

A computer program generated a set of item parameters for 16 nominally scored items each having four response categories. Within a given item, the slope and intercept parameters were each sampled from a unit normal distribution in a manner that met the constraints imposed by Equations 31–33. In a typical situation, the anchor items are embedded within a larger number of test items but these other items are not involved in the equating process. For convenience, all 16 items defined here were considered anchor items. 1,000 θ parameters were obtained from a unit normal distribution ranging from -2.8 to +2.8. The resulting item and θ parameters are referred to as the *underlying* parameters and were used to generate datasets in all examples.

Horizontal Equating

In horizontal equating, a common test is administered to two groups of examinees selected at random from the same population of examinees. The equating adjusts for differences due to sampling

variation between the two groups of examinees.

Using the parameters described above, two sets of item response data were generated with the GENIRV computer program (Baker, 1986). Different seeds were used to initialize the random number generator for the two datasets, denoted as Group 1 and Group 2. Each of the datasets was analyzed using MULTILOG. Due to the constraints imposed by Equations 31–33, the information matrix used in the maximum likelihood estimation of the item parameters was not of full rank. Thus, the item parameters c_{jk} and d_{jk} could not be estimated directly.

To resolve this problem, Bock (1972) recommended reparameterizing the items using the "unconstrained" (Thissen, 1991) item parameters. Because MULTILOG places only the unconstrained item parameter estimates in its output files, EQUATE converts them back to the usual slope and intercept estimates. This was accomplished by multiplying the obtained unconstrained estimates by the terms in the deviation contrast matrix as shown in the MULTILOG (Version 6.0) manual (Thissen, 1991, section 3, pp. 38, 39).

The item parameter estimates yielded by Group 2 were equated into the metric of the Group 1 estimates using EQUATE 2.0. In this horizontal equating, the values of the equating coefficients should have been approximately $A = 1$ and $K = 0$. The values obtained were $A = 1.0638$ and $K = -.0889$; these values of K were close to the expected values. The value of the quadratic loss function was $F = .0029$ indicating that, after equating, a very small discrepancy existed between the two sets of item response CRFs.

Parameter recovery studies form an important part of the IRT literature; however, to conduct these studies properly the obtained parameter estimates must be expressed in the metric of the parameters used to generate the item response data. Thus, such studies involve a special case of horizontal equating. To illustrate the parameter recovery situation, each of the two sets of MULTILOG item parameter estimates obtained above was equated back to the metric of the underlying parameters.

The observed equating coefficients should have been approximately $A = 1$ and $K = 0$. Group 1 yielded $A = 1.0204$ and $K = .0562$; Group 2 yielded $A = 1.0623$ and $K = -.0153$; the values of the quadratic loss functions were .0014 and .0017, respectively. In both groups, the values of A and K were very close to the desired values, and the values of the quadratic loss function were very small, indicating that the CRFs of the two sets of item response categories were well matched by the equating procedure.

Vertical Equating

Situation 1. A situation in which two administrations of a test yielded different item parameter estimates was considered. The task was to equate one set of test results into the metric of the other so that a common metric could be used to compare the test results.

To create a new set of item parameters, the numerical values of the underlying slope and intercept parameters of the 16 items created above were transformed using Equations 2 and 3 and values of $A = .5$ and $K = .5$. The new set of item parameters and the normally distributed underlying θ parameters of the horizontal equating examples were used in GENIRV to generate nominally scored item responses for 1,000 examinees to the 16 four-category items.

This dataset then was analyzed using MULTILOG. The metric of these item parameter estimates was considered to be the target test metric. Then, the item parameter estimates yielded by MULTILOG for Group 1 of the horizontal equating example were equated into this target test metric. The obtained values of the equating coefficients were $A = .5315$ and $K = .6455$. The value of the quadratic loss function was $F = .0038$, which indicated a small difference in the CRFs after equating. The value of A was close to its nominal value, but the value of K was somewhat larger than its nominal value.

The two sets of item parameters used in this example provided an excellent means of evaluating the accuracy of the equating procedure and its implementation. The values of the underlying item parameters were equated into the metric of the new set of item parameters. As anticipated, EQUATE 2.0 yielded exactly $A = .5$ and $K = .5$, and the value of the quadratic loss function was $F = 0.0$. Thus, when the two sets of slopes and intercepts were free of estimation errors, EQUATE 2.0 yielded the proper values of the equating coefficients.

Situation 2. A common set of items was administered to two groups of 1,000 examinees whose θ distributions differed. The θ parameters for Group 1 reproduced a normal distribution having a mean of –.5 and a unit SD. Those for Group 2 reproduced a normal distribution with a mean of +.5 and a SD of 1.5. The two groups represented examinees of low and high trait level, with a difference in mean trait level of 1.0. They also differed with respect to their variability (with a ratio of SDs of 1.5 to 1).

To generate the item response data, the set of underlying slope and intercept parameters for the 16 four-category items used in the previous example were paired with each of these θ distributions. GENIRV was used to generate the nominally scored item response data for each group, which then was analyzed using MULTILOG. MULTILOG employs a θ scale metric based on a θ distribution with mean = 0 and SD = 1. Thus, the underlying means and variances of the two groups were not preserved in the MULTILOG θ estimates. The item parameter estimates of Group 1 were considered to be the target test metric, and the Group 2 results were equated into this metric. The obtained values of the equating coefficients were $A = 1.4446$ and $K = .9331$. The value of the quadratic loss function was $F = .0011$.

After equating, the Group 2 trait level estimates had mean = .9083 and SD = 1.2924. The mean and SD of the target test θ estimates (Group 1) were –.0088 and .8593, respectively. In the target test metric, the transformed values of the summary statistics of the θ distributions should reflect the difference in mean θs and the relative variability of the two groups. The observed difference in means was .9171 and the ratio of the SDs was 1.504, both of which are consistent with the underlying values used to generate the item response data. It should be noted that the value of A (1.4446) was approximately the ratio of the SDs, and the value of K (.9331) was close to the difference in the means of the two groups.

Discussion

From a computational point of view, the basic equating technique used in IRT is to find the equating coefficients A and K that minimize a quadratic loss function. In the case of dichotomous and graded response items, the loss function is based on $(T_i - T_i^*)$. However, as shown in Equations 6 and 21, these true scores can be expressed in terms of $[P_{jk}(\theta_i) - P_{jk}^*(\theta_i)]$ for the anchor items in the two tests.

Because a true score does not exist for nominal response items, the loss function for this case is based on squaring this latter expression before summing over response categories and items. From the illustrative examples, it appears that this approach works quite well. In both the horizontal and vertical equating situations, the values of A and K based on these estimates reflected the characteristics of the defining parameters.

When the underlying slope and intercept parameters of the two nominally scored tests were used as input for EQUATE, it returned the exact values of A and K and a quadratic loss function value of 0. Because the exact values of the equating coefficients were reproduced when parameters were used, the difference between the obtained values of A and K and their nominal values appear to be a function of sampling variation in the item responses and hence in the item parameter estimates yielded by MULTILOG.

Initially, the examples provided here used sample sizes of 300. The obtained values of A differed

by approximately .2 and those for K by approximately .18 from the values yielded by samples of 1,000. Thus, it appears that, under the nominal response model, rather large sample sizes are needed to obtain item parameter estimates to be used in the equating process. The interaction among sample size, the estimation techniques employed in MULTILOG, and the equating coefficients yielded by EQUATE are in need of further investigation.

A dichotomously scored item can be formulated as an item with two response categories with the item characteristic curves (ICCs) of the two categories being CRFs. Consequently, the application of the equating procedure for the nominal response case described above to the dichotomous case is direct. In the case of items scored under the graded response model, the CRFs are similar in appearance to those under a nominal response model and represent the probability of selecting each response category as a function of trait level. This suggests that quadratic loss functions defined in terms of squared differences in CRFs can be defined for all three item scoring models. This alternative function is outlined in the Appendix.

Appendix:
Obtaining the Equating Coefficients Using an Alternative Quadratic Loss Function

The quadratic loss functions for all three item scoring procedures (the dichotomous case, the graded response case, and the nominal response case) were initially expressed in terms of differences in probabilities of selecting response categories. However, in both the dichotomous and graded response cases, summations of these differences were performed within the loss function to obtain true scores. Thus, in these two cases, the quadratic loss function was based on squared differences in true scores. Yet in Haebara's (1980) original formulation and in the nominal response case, the quadratic loss function was based on squaring $[P_{jk}(\theta_i) - P_{jk}^*(\theta_i)]$. This suggests that the quadratic loss function used in the nominal response case can encompass all three scoring procedures. The basic quadratic loss function is

$$F = \frac{1}{NS} \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{m_j} \{[P_{jk}(\theta_i) - P_{jk}^*(\theta_j)]\}^2 , \tag{47}$$

where

$$S = \sum_{j=1}^{n} m_j . \tag{48}$$

The Dichotomous Response Case

Because the dichotomous response case can be formulated as an item with two response categories—the incorrect response indexed by $k = 1$ and the correct response indexed by $k = 2$—the quadratic loss function would be

$$F = \frac{1}{NS} \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{2} [P_{jk}(\theta_i) - P_{jk}^*(\theta_i)]^2 , \tag{49}$$

where $S = 2n$.

The expressions for the gradients of this quadratic loss function with respect to A and K are needed. To be consistent with the formulation of the nominal response model, the slope-intercept parameterization will be employed here. Then

$$\frac{\partial F}{\partial A} = \frac{2}{NS}\sum_{i=1}^{N}\sum_{j=1}^{n}\left\{[P_{j1}(\theta_i) - P_{j1}^*(\theta_i)]d_{j1}^*P_{j1}^*(\theta_i)Q_{j1}^*(\theta_i)\left(\frac{\theta_i - K}{A}\right) + [P_{j2}(\theta_i) - P_{j2}^*(\theta_i)]d_{j2}^*P_{j2}^*(\theta_i)Q_{j2}^*(\theta_i)\left(\frac{\theta_i - K}{a}\right)\right\} \quad (50)$$

and

$$\frac{\partial F}{\partial K} = \frac{2}{NS}\sum_{i=1}^{N}\sum_{j=1}^{n}\{[P_{j1}(\theta_i) - P_{j1}^*(\theta_i)]d_{j1}^*P_{j1}^*(\theta_i)Q_{j1}^*(\theta_i) - [P_{j2}(\theta_i) - P_{j2}^*(\theta_i)]d_{j2}^*P_{j2}^*(\theta_i)Q_{j2}^*(\theta_i)\}\ , \quad (51)$$

and expressed in terms of the correct response

$$\frac{\partial F}{\partial A} = \frac{2}{NS}\sum_{i=1}^{N}\sum_{j=1}^{n}\{[P_{j2}(\theta_i) - P_{j2}^*(\theta_i)] - [Q_{j2}(\theta_i) - Q_{j2}^*(\theta_i)]\}d_{j2}^*P_{j2}^*(\theta_i)Q_{j2}^*(\theta_i)\left(\frac{\theta_i - K}{A}\right) \quad (52)$$

and

$$\frac{\partial F}{\partial K} = \frac{2}{NS}\sum_{i=1}^{N}\sum_{j=1}^{n}\{[P_{j2}(\theta_i) - P_{j2}^*(\theta_i)] - [Q_{j2}(\theta_i) - Q_{j2}^*(\theta_i)]\}d_{j2}^*P_{j2}^*(\theta_i)Q_{j2}^*(\theta_i)\ , \quad (53)$$

which after simplification yields

$$\frac{\partial F}{\partial A} = \frac{2}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n}[P_{j2}(\theta_i) - P_{j2}^*(\theta_i)]d_{j2}^*W_{j2}^*\left(\frac{\theta_i - K}{A}\right) \quad (54)$$

and

$$\frac{\partial F}{\partial K} = \frac{2}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n}[P_{j2}(\theta_i) - P_{j2}^*(\theta_i)]d_{j2}^*W_{j2}^*\ , \quad (55)$$

where

$$W_{j2}^* = P_{j2}^*(\theta_i)Q_{j2}^*(\theta_i)\ . \quad (56)$$

Thus, the gradients can be computed using only the correct response category. In the gradients, the probability of selecting the correct response category could be formulated using the multivariate logistic function. However, it is more convenient to employ the usual logistic ogive to model the ICC.

The Graded Response Case

The quadratic loss function of Equation 47 is also applicable to the graded response case, and the gradients with respect to A and K are needed. Under this model, parameters for the CRFs are not available and the response category probabilities are obtained using the boundary curves. Using the slope-intercept parameterization, the transformed boundary curve can be written as

$$\tilde{P}_{jk}(\theta_i) = \frac{1}{1 + \exp[-(c_{jk}^* + d_j^*\theta_i)]}\ , \quad (57)$$

where

$$\frac{\partial c_{jk}^*}{\partial A} = d_{jk}^*\left(\frac{K}{A}\right) \quad (58)$$

and

$$\frac{\partial d_j^*}{\partial A} = -\frac{d_j^*}{A} , \tag{59}$$

and

$$\frac{\partial \tilde{P}_{jk}(\theta_i)}{\partial c_{jk}^*} = \tilde{P}_{jk}(\theta_i)\tilde{Q}_{jk}(\theta_i) \tag{60}$$

and

$$\frac{\partial \tilde{P}_{jk}(\theta_i)}{\partial d_{jk}^*} = \theta_i \tilde{P}_{jk}(\theta_i)\tilde{Q}_{jk}(\theta_i) . \tag{61}$$

Then using the chain rule

$$\frac{\partial P_{jk}^*}{\partial A} = \frac{\partial}{\partial A}\left[\tilde{P}_{j,k-1}(\theta_i) - \tilde{P}_{jk}(\theta_i)\right] = d_j^*(\tilde{W}_{j,k-1} - \tilde{W}_{jk})\left(\frac{K - \theta_i}{A}\right) . \tag{62}$$

Similarly for K

$$\frac{\partial P_{jk}^*(\theta_i)}{\partial K} = \frac{\partial}{\partial K}\left[\tilde{P}_{j,k-1}(\theta_i) - \tilde{P}_{jk}(\theta_i)\right] = -d_j^*(\tilde{W}_{j,k-1} - \tilde{W}_{jk}) . \tag{63}$$

Substituting these derivatives in Equation 42 yields the gradients

$$\frac{\partial F}{\partial A} = \frac{2}{NS}\sum_{i=1}^{N}\sum_{j=1}^{n}\sum_{k=1}^{m_j}\{[P_{jk}(\theta_i) - P_{jk}^*(\theta_i)]\}\left[d_j^*(\tilde{W}_{j,k-1} - \tilde{W}_{jk})\left(\frac{\theta_i - K}{A}\right)\right] \tag{64}$$

and

$$\frac{\partial F}{\partial K} = \frac{2}{NS}\sum_{i=1}^{N}\sum_{j=1}^{n}\sum_{k=1}^{m_j}\{[P_{jk}(\theta_i) - P_{jk}^*(\theta_i)]\}[d_j^*(\tilde{W}_{j,k-1} - \tilde{W}_{jk})] , \tag{65}$$

where

$$S = \sum_{j=1}^{n} m_j . \tag{66}$$

These gradients are simply those for the nominal response case with W_{jk} replaced by $(\tilde{W}_{j,k-1} - \tilde{W}_{jk})$. It should be noted that item response category weights u_{jk} do not appear in either the quadratic loss function or the gradients of Equations 64 and 65.

From Equations 44, 45, 53, 54, 64, 65, and 66 it is clear that the basic structure of the gradients used in the Davidon-Fletcher-Powell technique is common to the three item scoring procedures. The three cases differ with respect to the weights W_{jk}, which depend on the item scoring model employed.

In the dichotomous and graded response cases, the TCC approach has two implementation advantages over this alternative approach. First, the use of true score differences tends to "smooth" out the errors in the differences between the two sets of item response category probabilities. Thus, it might yield somewhat better values of A and K.

Second, using differences in item response category probabilities requires that the computer program be provided with the information needed to match the corresponding anchor items in the current and target tests, whereas the true score approach is indifferent to the order of the items in the two tests. However, the use of true scores in the graded response case does require that the item response category weights be provided.

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. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17,* 20.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28,* 147–163.

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.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

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.

Fletcher, R., & Powell, M. J. D. (1963). A rapidly convergent decent 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.

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

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph,* No. 18.

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

Thissen, D. (1991). *MULTILOG*: Multiple, categorical item analysis and test scoring using item response theory. Mooresville IN: Scientific Software.

Author's Address

Send requests for reprints or further information to Frank B. Baker, Department of Educational Psychology, 859 Educational Sciences Building, 1025 W. Johnson St., University of Wisconsin, Madison WI 53706, U.S.A.