Equating Tests Under the Graded Response Model

Frank B. Baker, University of Wisconsin

The Stocking and Lord (1983) procedure for computing equating coefficients for tests having dichotomously scored items is extended to the case of graded response items. A system of equations for obtaining the equating coefficients under Samejima's (1969, 1972) graded response model is derived. These equations are used to compute equating coefficients in two related situations. Under the first, the equating coefficients are obtained by matching, on an examinee by examinee basis, the true scores on two tests. In the second case, the equating coefficients are obtained by matching the test characteristic curves (TCCs) of the two tests. Several examples of computing equating coefficients in these two situations are provided. The TCC matching approach was much less demanding computationally and yielded equating coefficients that differed little from those obtained through the true score distribution matching approach. *Index terms: equating coefficients, graded response model, quadratic loss function, response function method, Stocking and Lord equating technique, test equating, test characteristic curves.*

Under item response theory (IRT), the equating of tests consists of finding the slope and intercept coefficients for the linear transformation of the metric of one test calibration to that of another. Due to the manner in which the identification problem is resolved in most IRT test calibration computer programs, the metric information needed for equating is contained in the item parameter estimates.

Some techniques, such as those due to Marco (1977) and Loyd and Hoover (1980), use the summary statistics of the anchor item parameter estimates yielded by the two test calibrations to obtain the equating coefficients. However, this approach is sensitive to the distributional characteristics of the item parameter estimates, and deviant estimates can distort the values of the equating coefficients.

A more sophisticated approach is one in which the equating coefficients are obtained by minimizing a quadratic loss function based on differences in "true" scores yielded by the two test calibrations. This approach was first presented by Haebara (1980) and further refined by Stocking and Lord (1983). The loss function approach is preferred over the summary statistics approach for two reasons. First, it uses the item parameter estimates for each anchor item in a test rather than their summary statistics. Therefore, the equating coefficients are based on more detailed information. Second, the minimization of a loss function produces equating coefficients that in some sense are "optimum." At the present time, the loss function technique appears to be the method of choice for obtaining equating coefficients (Baker & Ali-Karni, 1991). However, it has been developed and implemented only for the case of dichotomous item responses. The present paper extends the approach to the graded response model (Samejima 1969, 1972).

The Stocking and Lord Procedure

Lord (1980) has shown that, under IRT, the relationship between the metric of any two test calibrations is linear; thus, the basic metric transformation equation can be expressed as

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

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 1, March 1992, pp. 87-96

0146-6216/92/010087-10$1.75

where A is the slope,
K is the intercept,
θ_i is the examinee's ability parameter in the metric of the new test, and
θ_i^* is θ_i expressed in the target test metric.

The values of the new test item parameters, a_j and b_j, can be transformed to the target test metric using the same coefficients as follows:

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

and

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

Stocking and Lord's (1983) technique for obtaining the two equating coefficients is based on the quadratic loss function

$$F = \frac{1}{N}\sum_{i=1}^{N}(T_i - T_i^*)^2 \quad , \tag{4}$$

where N is the number of examinees in the common group.

Under a two-parameter item response function (IRF) model and a common group of anchor items, the true scores, T_i for the target test and T_i^* for the transformed new test, are defined as

$$T_i = \sum_{j=1}^{n} P_j(\theta_i) = \sum_{j=1}^{n} 1/\{1 + \exp[-a_j'(\theta_i - b_j')]\} \tag{5}$$

and

$$T_i^* = \sum_{j=1}^{n} P_j^*(\theta_i) = \sum_{j=1}^{n} 1/\{1 + \exp[-a_j^*(\theta_i - b_j^*)]\} \quad , \tag{6}$$

where n is the number of items common to the two tests, and a_j' and b_j' are the parameters of the anchor items from the target test calibration. Because the goal is to express the new test results in the target test metric, the θ_i appearing in both T_i and T_i^* are in the target test metric. a_j^* and b_j^* in Equation 6 are the result of applying Equations 2 and 3 to the item parameter estimates yielded by the new test calibration.

The task then is to find the equating coefficients that will minimize the quadratic loss function. 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 and Lord (1983) used an iterative multivariate search technique (Davidon, 1959; Fletcher & Powell, 1963) to find the two equating coefficients that will minimize F.

Stocking and Lord (1983) referred to their technique as the characteristic curve method. However, they actually presented two different approaches without clearly differentiating between them. As employed in Equation 4, the quantity $(T_i - T_i^*)$ is computed for each examinee, squared, and then summed over all examinees in the common group. Because of the summation over examinees, the process is one that minimizes the difference between the two distributions of true scores based on the common anchor items of the two tests administered to a common group of examinees. Thus, it would be employed only in horizontal equating situations. Haebara's (1980) method is similar except that he grouped the examinees on the θ scale before computing the true scores and then multiplied the squared difference by the group frequency count.

The second approach results from what Stocking and Lord (1983) presented as a scheme for simplifying the computational procedure. Rather than computing the quantity ($T_i - T_i^*$) at the target test θ level of each examinee, an arbitrary set of N values along the target test θ scale could be used. The values of the true scores T_i and T_i^* would be computed directly from the two sets of anchor item parameter estimates for the N arbitrary points along the target test θ scale. The summation in Equation 4 is now over the N arbitrary points rather than over examinees. The T_i and T_i^* obtained are the values of the two test characteristic curves (TCCs) at each selected point on the target test θ scale. In this second case, the loss function minimizes the difference between these two TCCs. It is this version that was implemented in the computer program (Stocking, 1985) for obtaining the equating coefficients. This second approach has two advantages. First, because the T_i and T_i^* are not computed for each examinee, it is much less demanding computationally than the former approach. Second, it requires only a target test θ scale and two sets of item parameter estimates for the common anchor items—it does not require a common group of examinees. As a result, it can be employed in a wider range of equating situations (e.g., vertical equating or equating a set of test results to an underlying metric) than can the former approach.

The Graded Response Case

Under Samejima's (1969, 1972) graded response model, an item possesses m_j ordered response categories, such as in a Likert scale, and the examinee selects only one of the categories. Each category has a response weight associated with it so that an examinee's true score is defined as

$$T_i = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} P_{jk}(\theta_i) \quad , \tag{7}$$

where k denotes an item response category of item j;
m_j is the number of response categories of item j—therefore, $1 \leq k \leq m_j$; and
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, and category m_j is allocated the largest weight. $P_{jk}(\theta_i)$ is the probability of selecting category k of item j for an examinee of ability θ_i. As was the case with dichotomously scored items, the examinee's true score does not depend on their vector of item response choices.

The estimation of the item parameters under the graded response model involves the use of $m_j - 1$ boundary curves representing the cumulative probability of selecting response categories greater than and including the response category of interest (Samejima, 1969). The boundary curves are characterized by an item discrimination parameter, a_j, and by the $m_j - 1$ location parameters, b_{jk}. The b_{jk} for an item are ordered, typically from low ($k = 1$) to high ($k = m_j$).

For a given item, a_j is the same over all boundary curves. As a result, the probability of selecting a given response category of a target test item is given by the following expressions:
when $1 < k < m_j$

$$P_{jk}(\theta_i) = \hat{P}_{j,k-1}(\theta_i) - \hat{P}_{jk}(\theta_i) \quad ; \tag{8}$$

and when $k = 1$

$$P_{j1}(\theta_i) = 1 - \hat{P}_{j1}(\theta_i) \quad ; \tag{9}$$

and when $k = m_j$

$$P_{jm_j}(\theta_i) = \hat{P}_{j,m_j-1}(\theta_i) \quad ; \tag{10}$$

where $\hat{P}_{jk}(\theta_i)$ are the cumulative probabilities obtained from the boundary curves. Thus, a true score based on the target test can be defined in terms of the boundary curves as follows:

$$T_i = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} P_{jk}(\theta_i) = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} [\hat{P}_{j,k-1}(\theta_i) - \hat{P}_{jk}(\theta_i)] \quad . \tag{11}$$

In the case of a new test 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} through Equations 2 and 3. This probability can also can be expressed in terms of transformed boundary curves. Let

$$\tilde{P}_{jk}(\theta_i) = 1/\{1 + \exp[-a^*_j(\theta_i - b^*_{jk})]\} \quad ; \tag{12}$$

then for $1 < k < m_j$

$$P^*_{jk}(\theta_i) = \tilde{P}_{j,k-1}(\theta_i) - \tilde{P}_{jk}(\theta_i) \quad ; \tag{13}$$

when $k = 1$

$$P^*_{j1}(\theta_i) = 1 - \tilde{P}_{j1}(\theta_i) \quad ; \tag{14}$$

and when $k = m_j$

$$P^*_{jm}(\theta_i) = \tilde{P}_{j,m-1}(\theta_i) \quad . \tag{15}$$

Once the item parameters of the new test are transformed into the target test metric through the equating coefficients A and K, an examinee's new test true score is given by

$$T_i^* = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} P^*_{jk}(\theta_i) = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} [\tilde{P}_{j,k-1}(\theta_i) - \tilde{P}_{jk}(\theta_i)] \quad , \tag{16}$$

and substituting for $\tilde{P}_{jk}(\theta_i)$ and $\tilde{P}_{j,k-1}(\theta_i)$ yields

$$T_i^* = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} \{1/\{1 + \exp[-a^*_j(\theta_i - b^*_{j,k-1})]\} - 1/\{1 + \exp[-a^*_j(\theta_i - b^*_{jk})]\}\} \quad . \tag{17}$$

The Davidon-Fletcher-Powell minimization technique requires that the derivatives (gradients), with respect to A and K, of the quadratic loss function of Equation 4 be evaluated at each primary iteration. Because T_i in these derivatives does not involve A and K, only $\partial T_i^*/\partial A$ and $\partial T_i^*/\partial K$ are needed. From the chain rule

$$\frac{\partial T_i^*}{\partial A} = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} \left[\frac{\partial P^*_{jk}(\theta_i)}{\partial b^*_{jk}} \cdot \frac{\partial b^*_{jk}}{\partial A} + \frac{\partial P^*_{jk}(\theta_i)}{\partial a^*_j} \cdot \frac{\partial a^*_j}{\partial A} \right] \tag{18}$$

and

$$\frac{\partial T_i^*}{\partial K} = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} \left[\frac{\partial P^*_{jk}(\theta_i)}{\partial b^*_{jk}} \cdot \frac{\partial b^*_{jk}}{\partial K} + \frac{\partial P^*_{jk}(\theta_i)}{\partial a^*_j} \cdot \frac{\partial a^*_j}{\partial K} \right] \quad . \tag{19}$$

Substituting $\tilde{P}_{j,k-1}(\theta_i) - \tilde{P}_{jk}(\theta_i)$ for $P^*_{jk}(\theta_i)$ yields

$$\frac{\partial T_i^*}{\partial A} = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} \left\{ \left[\frac{\partial \tilde{P}_{j,k-1}(\theta_i)}{\partial b^*_{j,k-1}} \cdot \frac{\partial b^*_{j,k-1}}{\partial A} + \frac{\partial \tilde{P}_{j,k-1}(\theta_i)}{\partial a^*_j} \cdot \frac{\partial a^*_j}{\partial A} \right] - \left[\frac{\partial \tilde{P}_{jk}(\theta_i)}{\partial b^*_{ja}} \cdot \frac{\partial b^*_{jk}}{\partial A} + \frac{\partial \tilde{P}_{jk}(\theta_i)}{\partial a^*_j} \cdot \frac{\partial a^*_j}{\partial A} \right] \right\} \tag{20}$$

and

$$\frac{\partial T_i^*}{\partial K} = \sum_{j=1}^{n}\sum_{k=1}^{m_j} u_{jk}\left\{\left[\frac{\partial \tilde{P}_{j,k-1}(\theta_i)}{\partial b^*_{j,k-1}} \cdot \frac{\partial b^*_{j,k-1}}{\partial K} + \frac{\partial \tilde{P}_{j,k-1}(\theta_i)}{\partial a_j^*} \cdot \frac{\partial a_j^*}{\partial K}\right] - \left[\frac{\partial \tilde{P}_{jk}(\theta_i)}{\partial b^*_{ja}} \cdot \frac{\partial b^*_{jk}}{\partial K} + \frac{\partial \tilde{P}_{jk}(\theta_i)}{\partial a_j^*} \cdot \frac{\partial a_j^*}{\partial K}\right]\right\} \quad . \tag{21}$$

Some necessary derivatives are

$$\frac{\partial b_i^*}{\partial A} = \frac{\partial}{\partial A}(Ab_{jk} + K) = b_{jk} \qquad \frac{\partial b_j^*}{\partial K} = 1 \quad , \tag{22}$$

$$\frac{\partial a_j^*}{\partial A} = \frac{\partial}{\partial A}[a_j/A] = -a_j/A^2 \qquad \frac{\partial a_j^*}{\partial K} = 0 \quad , \tag{23}$$

and

$$\frac{\partial \tilde{P}_{jk}(\theta_i)}{\partial a_i^*} = (\theta_i - b_{jk}^*)\tilde{P}_{jk}(\theta_i)\tilde{Q}_{jk}(\theta_i) \qquad \frac{\partial \tilde{P}_{jk}(\theta_i)}{\partial b_{jk}^*} = -a_j^*\tilde{P}_{jk}(\theta_i)\tilde{Q}_{jk}(\theta_i) \quad . \tag{24}$$

Then

$$\frac{\partial T_i^*}{\partial A} = \sum_{j=1}^{n}\sum_{k=1}^{m_j} u_{jk}\{\{-a_j^*\tilde{P}_{j,k-1}(\theta_i)\tilde{Q}_{j,k-1}(\theta_i)[b_{j,k-1} - (\theta_i - b^*_{j,k-1})/A]\} - \{-a_j^*\tilde{P}_{jk}(\theta_i)\tilde{Q}_{jk}(\theta_i)[(b_{jk} - (\theta_i - b_{jk}^*)/A]\}\}. \tag{25}$$

Let

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

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

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

and

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

then

$$\frac{\partial T_i^*}{\partial A} = \sum_{j=1}^{n}\sum_{k=1}^{m_j} u_{jk}[-a_j^*(w_{jk}z_{jk} - w_{j,k-1}z_{j,k-1})] \quad . \tag{30}$$

For the intercept coefficient K

$$\begin{aligned}\frac{\partial T_i^*}{\partial K} &= \sum_{j=1}^{n}\sum_{k=1}^{m_j} u_{jk}\{-a^*\tilde{P}_{j,k-1}(\theta_i)\tilde{Q}_{j,k-1} - [-a_j^*\tilde{P}_{jk}(\theta_i)\tilde{Q}_{jk}(\theta_i)]\} \\ &= \sum_{j=1}^{n}\sum_{k=1}^{m_j} u_{jk}(a_j^*)\,[\tilde{P}_{jk}(\theta_i)\tilde{Q}_{jk}(\theta_i) - \tilde{P}_{j,k-1}(\theta_i)\tilde{Q}_{j,k-1}(\theta_i)] \\ &= \sum_{j=1}^{n}\sum_{k=1}^{m_j} u_{jk}(a_j^*)\,(w_{jk} - w_{j,k-1}) \quad .\end{aligned} \tag{31}$$

The gradients are given by

$$\frac{\partial F}{\partial A} = \frac{-2}{N}\sum_{i=1}^{N}(T_i - T_i^*)\frac{\partial T_i^*}{\partial A} \tag{32}$$

and

$$\frac{\partial F}{\partial K} = \frac{-2}{N}\sum_{i=1}^{N}(T_i - T_i^*)\frac{\partial T_i^*}{\partial K} \quad . \tag{33}$$

Substituting for the true score derivatives with respect to A and K yields

$$\frac{\partial F}{\partial A} = \frac{-2}{N}\sum_{i=1}^{N}(T_i - T_i^*)\left\{\sum_{j=1}^{n}\sum_{k=1}^{m_j} u_{jk}[a_j^*(w_{jk}z_{jk} - w_{j,k-1}z_{j,k-1})]\right\} \tag{34}$$

and

$$\frac{\partial F}{\partial K} = \frac{-2}{N}\sum_{i=1}^{N}(T_i - T_i^*)\left\{\sum_{j=1}^{n}\sum_{k=1}^{m_j} u_{jk}[a_j^*(w_{jk} - w_{j,k-1})]\right\} \quad . \tag{35}$$

The equations presented above can be used when the equating coefficients are based on matching the examinee's true scores or on matching TCCs. In the former, the terms are computed on an examinee by examinee basis, and the summation is over the N examinees. In the latter, the terms are computed only for the N arbitrary points on the target test θ scale.

Because the two cases differ only in the number and definition of the points on the θ scale employed, both of the above procedures for estimating the equating coefficients were implemented in a single computer program written in Professional FORTRAN for the IBM PC/AT. In the program implementation, two simplifying assumptions were made. First, all anchor items had the same number of response categories. Second, the response category weights were integers where $1 \leq u_{jk} \leq m_j$.

The Davidon-Fletcher-Powell technique was implemented using a set of six subroutines taken from *Numerical Recipes* (Press, Flannery, Teukolsky, & Vetterling, 1986). Each primary iteration of the process requires the evaluation of F, $\partial F/\partial A$, and $\partial F/\partial K$, and each secondary iteration requires that F be evaluated. There is approximately a four-to-one ratio of secondary iterations to primary iterations. As a result, the procedure requires a moderate amount of computing time on a personal computer. The iterative process terminates when a convergence criterion based on successive values of F is met.

Examples of Computing Graded Response Equating Coefficients

Horizontal Equating Under the True Score Approach

Two simulated datasets based on the same anchor items and a common group of 300 examinees were used to illustrate horizontal equating under the true score approach. Both sets were based on a test of 30 items with four response categories. All 30 items were used as the anchor items in the equating. The discrimination parameters of the test were generated from a uniform distribution ranging from 1.34 to 2.65 in a logistic IRF metric. The three difficulty parameters for the boundary curves of an item were generated from a normal distribution (mean = 0, variance = 1). Each set of three ordered boundary curve difficulty parameters of an item was randomly paired with a single discrimination index. The θ levels of the 300 simulated examinees were sampled from a unit normal distribution over the range -2.8 to $+2.8$.

The GENIRV computer program (Baker, 1986) was used to generate the vectors of examinee response category choices for the target test. Then, using the same item and θ parameter specifications and a new random number generator seed, a new set of examinee item response choice vectors was generated. Each of the two datasets was then analyzed by the MULTILOG computer program (Thissen, 1988) yielding the estimates of the item and θ parameters. The second set of test results was to be expressed in the metric of the first test calibration. The examinee θ score estimates from the

MULTILOG analysis of the target test data were used as the N θ levels employed in the equating process. Because the equating task was that of horizontal equating, the theoretical values of the equating coefficients should be $A = 1.0$ and $K = 0$. The coefficient values obtained, after four primary iterations of the equating computer program, were $A = .9934$ and $K = -.0703$. These values are in good agreement with the theoretical values. The value of the quadratic loss function at convergence, $F = .0285$, suggests that the minimum was nearly attained.

Using these coefficients, the item and θ parameter estimates of the new test were transformed to the target test metric. The means of the MULTILOG item and θ parameter estimates of the target test and those of the new test after transformation are reported in Table 1. Because the two datasets differed only with respect to the random number seed used to generate the examinees' item response category choice vectors, there should be good agreement of the two sets of test results. This was the case, as shown by the fact that the means of the two sets of results differ only in the third decimal place.

Table 1
Summary Statistics of Target and Transformed Parameters For Three Equatings

Parameter	Horizontal Equating: True Score Matching	Horizontal Equating: TCC Matching	Nonhorizontal Equating
Coefficients			
A	.9934	.9986	1.0083
K	-.0703	-.0715	.5432
F	.0285	.0359	.0756
Means for Target Test			
$\bar{a}$	1.992	1.992	1.969
$\bar{b}_1$	-1.559	-1.559	-.977
$\bar{b}_2$	-.221	-.221	.364
$\bar{b}_3$	1.465	1.465	2.284
$\bar{\theta}$	-.024	-.024	.032
Means for Transformed Test			
$\bar{a}$	1.899	1.899	1.871
$\bar{b}_1$	-1.560	-1.568	-.968
$\bar{b}_2$	-.195	-.197	.365
$\bar{b}_3$	1.470	1.477	2.107
$\bar{\theta}$	-.024	-.040	.575

Horizontal Equating Under the Test Characteristic Curve Approach

These same two sets of test data were also used to illustrate computation of horizontal equating coefficients using the TCC approach. The two sets of item parameter estimates and 21 points equally spaced from -4 to +4 on the target test θ scale were entered into the computer program for estimating the equating coefficients. The obtained values were $A = .9986$ and $K = -.0715$. The values agree with both the anticipated values and with those yielded by the true score procedure. The value of the quadratic loss function at convergence, $F = .0359$, was trivially larger than that obtained from the true score procedure. In terms of the values of the obtained equating coefficients, there was little difference in the outcomes of the two approaches. The means of the target test and the transformed test item and θ parameter estimates are also reported in Table 1. Again, the differences in mean values are in the third decimal places. In addition, the differences between the mean values of the trans-

formed parameter estimates obtained from the two analysis procedures were very small.

Nonhorizontal Equating

To illustrate a nonhorizontal equating situation, the item parameters of the 30-item test used in the preceding examples were transformed through Equations 2 and 3 using the values $A = .9$ and $K = .5$. These values were then used in GENIRV (Baker, 1986) to generate a target test dataset based on 300 examinees whose θ parameters were normally distributed (mean = 0, variance = 1). The generated examinee item response vectors were then analyzed by MULTILOG to obtain the item parameter estimates for the target test. The means are reported in Table 1.

The original MULTILOG test results were then equated to the metric of these results by the TCC approach using 21 points equally spaced from −4 to +4 on the θ scale. The obtained values of the equating coefficients were $A = 1.0083$ and $K = .5432$. These agree reasonably well with the underlying values, although the A coefficient is approximately .1 larger. The obtained value of the loss function F was .0756, which indicates that the minimum was approximated. In the present equating situation, the two sets of underlying item parameters differed primarily in terms of the locations of the items. Thus, when the two sets of common items were equated, the change in location was reflected in the mean abilities of the two groups of examinees. The summary statistics in Table 1 for the nonhorizontal equating reflect this relationship. There was very good agreement between the mean values of the item parameter estimates for the target test and the transformed test results. As expected, the mean θ of the examinees of the transformed test (.5751) was approximately .5 above the mean (.0317) of the target test group.

Discussion

The definition of the true score for the case of graded response items can be considered an extension of the true score definition for the case of dichotomously scored items. In the dichotomous case, the true score is the sum of the probabilities of correct response over the n items at a given θ level. Because the true score is an expected value, it does not depend on the examinee's vector of responses to the items.

In the graded response case, each item response category has a weight (u_{jk}) associated with it. Thus, the product of the probability of selecting a response category, at a given θ level, and its weight are summed over the m_j categories to obtain the true score for an item. The true score, at a given θ level, is the sum of these item true scores over the n items. Again, the true score does not depend on the examinee's vector of item response category choices because it is an expected value.

The first of the two Stocking and Lord (1983) procedures for computing the equating coefficients finds the values of A and K that minimize the difference between the true score distribution based on the anchor items in the target test and the true score distribution of the new test after transformation of the anchor item parameters. Horizontal equating of two tests containing anchor items administered to a common group of examinees is the only feasible type of equating in this situation.

In the second Stocking and Lord (1983) procedure, the equating coefficients are obtained by minimizing the difference between the TCC of the target test based on the anchor items and that of the transformed test anchor items. The advantage of this second approach is that it also can equate tests in which there are two groups of examinees that differ with respect to ability. Thus, nonhorizontal equating can be accomplished. As shown in the examples above, when both approaches are applied in the horizontal equating situation, similar values of the equating coefficients are obtained. In practice, when the values of the equating coefficients based on the common anchor items have been determined, all the item parameters of the new test and the corresponding examinee θ parameters can

be transformed into the target test metric.

In the three examples shown above, the obtained values of the equating coefficients were very close to the theoretical values. In addition, the loss function value at convergence of the iterative estimation process was very close to 0. The minor differences in the reported means of the item and θ parameter estimates were due to sampling variation in the examinees' item response choice vectors, which is reflected in the MULTILOG parameter estimates. Thus, the present paper has extended the Stocking and Lord characteristic curve method of estimating equating coefficients to the graded response model. Future research should examine the effect of such sampling variation on the obtained equating coefficients in a variety of conditions.

A FORTRAN program was written to implement the computation of the equating coefficients for graded response tests. Due to the iterative nature of the process, computing times were quite long. For example, under the true score distribution matching approach, the datasets required approximately two hours of computer time on an IBM PC/AT to perform four primary iterations. The amount of time required reflects the need to compute the probability of choice for all response categories for each anchor item for each examinee in both tests. When the TCC matching approach was employed with 21 points along the θ scale, the same test data required approximately 13 minutes for five primary iterations. Thus, the latter approach is much less demanding computationally. Because there was little difference in the obtained equating coefficients, the TCC matching approach is the method of choice.

With some datasets, a peculiarity of the Davidon-Fletcher-Powell gradient search method was observed that partially accounts for the long computer runs under the true score approach. The gradients typically decreased rapidly from a large initial value to a small value, and the process usually converged in three to four primary iterations. In some cases, after the gradients became quite small (e.g., 10^{-6}), the process required a total of five to ten primary iterations to achieve convergence. In the final three or four primary iterations, the loss function F was essentially 0, yet convergence was not readily achieved. This phenomenon appears to be a characteristic of the convergence criterion employed within the *Numerical Recipes* subroutines (Press et al., 1986) that depends on the relative values of the quadratic loss function in two successive primary iterations rather than on an absolute difference.

References

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

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

Davidon, W. C. (1959). *Variable metric method for minimization* (Research and Development Rep. No. ANL-5990, revised edition). Argonne IL: Argonne National Laboratory, U.S. Atomic Energy Commission.

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

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

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 2, 139–160.

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. (1985). *Transforming B's using a least squares technique: The TBLT system.* [Computer

program manual]. Unpublished manuscript, Educational Testing Service, Princeton NJ.

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

Thissen, D. (1988). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Mooresville: Scientific Software.

Author's Address

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