EQUATING LOGISTIC ABILITY SCALES BY A WEIGHTED LEAST SQUARES METHOD

TOMOKAZU HAEBARA¹

University of Iowa

A general procedure for equating logistic ability scales that involves the use of a loss function and an optimization process is formulated. As one of the optimum estimation procedures, a weighted least squares method is introduced and applied to an achievement scale and an intelligence scale. An advantage of the method over previously proposed methods is that highly discriminating items tend to affect equating more strongly than less discriminating items.

The purpose of this study is to develop a weighted least squares method for transforming a logistic scale in such a way that estimates of ability parameters on the transformed scale are as comparable as possible with those on another scale. In the present paper, such a scale transformation process is referred to as equating of scales.

Equating is an extremely important procedure in studies involving the logistic model and the latent trait model in general. To mention a few examples, Shiba’s stratified adaptive test of verbal ability is based on successive vertical equatings of thirteen logistic scales that jointly cover the ability levels of preschool children through adults (Shiba, 1978; Shiba, Noguchi, & Haebara, 1978; Shiba, Noguchi, & Ohama, 1980). Haebara’s (1979) study on item bias largely depends on a horizontal equating of logistic scales for two racial groups.

The equating method to be presented in this paper deals with situations in which there are two separately scaled tests that have only some items in common (vertical equating or horizontal equating) or all items in common (horizontal equating). The necessity for equating in the latter case (all items common) arises when two scales have been derived by using two different groups of examinees which may have different characteristics pertinent to the results of scaling. In the former case (only some items common), the groups of examinees may or may not be overlapped.

A conceptually simple way of equating in the situations described above is to pool all items in the two tests and all examinees in the two groups, and to reestimate examinee ability and item characteristic curve parameters simultaneously, using the enlarged set of data. However, this procedure is often prohibitively expensive, not only because it involves a large number of examinees and/or items but also because it wastes results available from the previous scaling. Hence, it is desirable to develop an alternative method for equating that utilizes the previously obtained scales.

FORMULATION

Logistic Model

In the logistic test model, the probability of an examinee with ability parameter \( \theta \) giving a correct answer to item \( g \) is mathematically stated as

\[ P(g|\theta) = \frac{1}{1 + e^{-\lambda g}} \]

where \( \lambda \) is the item discrimination parameter.
where $a_g$, $b_g$, and $c_g$ are parameters describing item $g$, and $D$ is a constant assuming the value of 1.7 (see Birnbaum, 1968; Hambleton, Swaminathan, Cook, Eigner, & Gifford, 1978). Equation (1) can be conceived of as representing the following three different logistic models: (a) Three-parameter model, where each of the three parameters is to be determined for each item; (b) Two-parameter model, where $c_g$ is zero for all items; and (c) One-parameter model, where $a_g$ is a constant across all items and $c_g$ is zero for all items. Although the ensuing discussion will focus on the three-parameter model, the method to be presented later is applicable to the other logistic models as well.

Principles of Equating

Let two logistic scales to be equated be Scale 1 and Scale 2. By the nature of the logistic model, the two scales have arbitrary units and origins. Therefore, for an ability parameter $\theta_1$ on Scale 1 to be comparable with ability parameters on Scale 2, it must be transformed to an equated value $\theta_{12}$ by

$$\theta_{12} = k\theta_1 + d,$$

(2)

where $k$ and $d$ are the equating coefficients. (Any nonlinear transformation of the scale violates the basic assumption of the logistic model by causing non-logistic item characteristic curves.)

Then a perfect equating implies that

$$P_{g,1}(\theta_1) = P_{g,2}(\theta_{12}), \quad -\infty < \theta_1 < \infty,$$

(3)

for all common items $g$ ($g=1, 2, \ldots, m$, say). Here the second subscript to $P_g$ refers to the scale it is defined on. Thus the equating problem in the present context is to find the values of $k$ and $d$ in (2) that satisfy (3) for all of the $m$ common items.

In practice, however, (3) never holds for all common items because of sampling errors and a possible lack of fit of the model. Therefore, the actual task is to estimate $k$ and $d$ so that (3) holds as nearly as possible. An objective and valid way of solving the problem is to define a criterion function that reflects the degree of departure of an equating result from the relationship stated in (3) and to find such values of $k$ and $d$ as to minimize the criterion function.

For item $g$ ($g=1, 2, \ldots, m$) and examinee $a$ ($a=1, 2, \ldots, N$), let $e_{g,a,1}$ be defined by

$$e_{g,a,1} = P_{g,1}(\theta_{1,a}) - P_{g,2}(\theta_{12,a}).$$

(4)

Then by choosing an appropriate loss function $L$ that evaluates the loss caused by the magnitude of $e_{g,a,1}$, the equating error $Q_I_g$ for item $g$ can be written

$$Q_I_g = \sum_{a=1}^{N} L(e_{g,a,1}).$$

Furthermore, the total equating error $Q$ is defined by

$$Q = \sum_{g=1}^{m} Q_I_g.$$  

(5)

If the values of $k$ and $d$ minimizing $Q$ are found, they are the optimum equating coefficients for the loss function $L$.

Previously Developed Methods

Although several equating methods have been developed and used in the past few years, none of them have adopted the optimization procedure described above. Those methods are characterized by the use of the moments of the distributions of item parameters $a_g$ and $b_g$ on the two scales. The basic idea underlying those methods is that, when the equating is defined by (2), the $a_g$ and $b_g$ parameters on Scale 1 must be transformed by

$$a_{g,12} = a_g/k$$

and
to be comparable with the item parameters on Scale 2. Here the second subscript to the item parameter refers to the scale it is defined on.

For example, Marco (1977) and Haebara (1979) estimated \( k \) and \( d \) by

\[
k = \frac{(SD \ of \ \theta_{1,1})}{(SD \ of \ \theta_{1,2})}
\]  

and

\[
d = (Mean \ of \ \theta_{1,2}) - k(Mean \ of \ \theta_{1,1})
\]  

where the means and SD’s are computed by using all common items. Shiba (1978) proposed two different ways of estimating \( k \):

\[
k = \frac{(\sum_{g=1}^{m} a_{g,1})}{(\sum_{g=1}^{m} a_{g,2})}
\]

and

\[
k = \frac{(\sqrt{1+u^2}-1)}{u}
\]

if

\[
u = \frac{(2 \sum_{g=1}^{m} a_{g,1}a_{g,2})}{(\sum_{g=1}^{m} a_{g,2}^2 - a_{g,1}^2)} > 0
\]

and

\[
k = \frac{u}{(1 - \sqrt{1+u^2})}
\]

if \( u < 0 \). More recently, Douglass (1980) claimed that the estimate of \( k \) given by the average of \( a_{g,1}/a_{g,2} \) provided a reasonable equating.

As Douglass (1980) admitted, however, the best method of combining the information contained in the \( a_{g} \) and the \( b_{g} \) for estimating \( k \) and \( d \) is not obvious as far as such heuristic approaches are concerned. If the optimization procedure described in the preceding section is employed, this kind of ambiguity disappears once an appropriate loss function has been selected by the researcher. In the next section, one of the optimum estimation procedures which employs the weighted least squares criterion is discussed.

## Method

### Criterion Function

Let the loss function \( L \) be defined by

\[
L(\theta_{g,1}) = \epsilon_{g,1}^2
\]

where \( \epsilon_{g,1} \) is given by (4). Then \( Q \) in (5) becomes

\[
Q = \sum_{g=1}^{m} \sum_{a=1}^{N} \epsilon_{g,1}^2
\]

If we construct a relative frequency distribution \( h_1 \) of \( \theta_1 \) by dividing the range of \( \theta_1 \) into some \( n_1 \) small intervals with the middle points \( \theta_{1,i} \) (\( i = 1, 2, \ldots, n_1 \)), minimizing \( Q \) is approximately equal to minimizing \( Q_1 \) defined by

\[
Q_1 = \sum_{g=1}^{m} \sum_{i=1}^{n_1} \epsilon_{g,i}^2 h_1(\theta_{1,i})
\]

(This approximation is made here for practical reasons; i.e., to reduce the computation time and the computer storage space needed.)

Moreover, it may be desirable that the criterion function also takes into account the effect of equating errors on the examinees whose ability parameters are defined on Scale 2. This can be achieved by first defining \( Q_2 \) by

\[
Q_2 = \sum_{g=1}^{m} \sum_{i=1}^{n_1} \epsilon_{g,i}^2 h_2(\theta_{2,i})
\]

where \( h_2 \) is the relative frequency distribution of \( \theta_2 \), \( n_2 \) is the number of intervals, and

\[
\epsilon_{g,i} = P_{g,2}(\theta_{2,i}) - P_{g,1}(\theta_{2,i})
\]

where

\[
\theta_{2,i} = (\theta_{1,i} - d)/k
\]

which defines a transformation of ability parameters on Scale 2 to Scale 1. Then the final form of the criterion function is given by

\[
Q^* = Q_1 + Q_2
\]
Equating Logistic Ability Scales

Similarly, the equating error for item g is

\[ Q_{g}^* = \sum_{t=1}^{n_{1}} e_{t1} q_{2} h_{1}(\theta_{1,t}) + \sum_{w=1}^{n_{2}} e_{w1} q_{2} h_{2}(\theta_{2,w}) \]  

Equation (8) is recognized as the criterion function for a weighted least squares estimation of k and d with the weights being \( h_{1}(\theta_{1,t}) \) and \( h_{2}(\theta_{2,w}) \).

Solution

The weighted least squares solution for \( k \) and \( d \) is obtained by solving the simultaneous equations

\[ \partial Q^*/\partial k = 0 \]  

and

\[ \partial Q^*/\partial d = 0 \]

for \( k \) and \( d \). Equations (10) and (11) are to be solved numerically, since no closed form of analytical solution is available.

Among various techniques for solving such equations, we apply the Gauss-Newton method. For convenience of illustration, let us introduce some vector and matrix expressions:

\[ \mathbf{x} = \begin{bmatrix} k \\ d \end{bmatrix} \]

\[ \mathbf{f} = [P_{1,p}(\theta_{1,1}), P_{1,p}(\theta_{1,2}), \ldots, P_{w,s}(\theta_{w,1}), \ldots, P_{w,s}(\theta_{w,n})] \]

\[ \hat{\mathbf{f}} = [P_{1,p}(\theta_{1,1}), P_{1,p}(\theta_{1,2}), \ldots, P_{w,s}(\theta_{w,1}), \ldots, P_{w,s}(\theta_{w,n})] \]

\[ \mathbf{W} = \text{Diagonal matrix with elements} \]

\[ h_{1}(\theta_{1,1}), \ldots, h_{t}(\theta_{1,n_{1}}), h_{2}(\theta_{2,1}), \ldots, h_{2}(\theta_{2,n_{2}}). \]

Using these expressions, (8) can be rewritten as

\[ Q^* = (\mathbf{f} - \hat{\mathbf{f}})^T \mathbf{W} (\mathbf{f} - \hat{\mathbf{f}}). \]

And (10) and (11) are combined in the form

\[ \mathbf{g}(\mathbf{x}) \equiv (\partial \hat{\mathbf{f}}/\partial \mathbf{x})^T \mathbf{W} (\mathbf{f} - \hat{\mathbf{f}}) = 0. \]

Then the vector of estimates are updated successively by

\[ \mathbf{x}^{(j+1)} = \mathbf{x}^{(j)} - \mathbf{H}(\mathbf{x}^{(j)})^{-1} \mathbf{g}(\mathbf{x}^{(j)}), \]

where

\[ \mathbf{H}(\mathbf{x}) = (\partial \hat{\mathbf{f}}/\partial \mathbf{x})^T \mathbf{W} (\partial \hat{\mathbf{f}}/\partial \mathbf{x}), \]

and superscripts \( (j) \) and \( (j+1) \) refer to iteration numbers (see, for example, Dixon, 1972). A scalar \( t \) in (12) is a step size factor which is used to keep the criterion function \( Q^* \) from increasing. At the beginning of each iteration, \( t \) is set to one. And if \( Q^* \) increases with this step size, then \( t \) is repeatedly halved until \( Q^* \) decreases.

To implement this algorithm, one must specify the initial values for \( k \) and \( d \). It is quite reasonable to use the values obtained by (6) and (7) or by any other previously developed formulas. Or if one has some prior knowledge about the approximate values of \( k \) and \( d \), he may use it.

Applications

To illustrate the proposed method, equating was carried out for two sets of data.

Equating 1

Applying the three-parameter logistic model with fixed non-zero \( c_{2} \) values, item parameters were estimated for a 25-item intelligence test (Cognitive Abilities Test, Verbal Battery, Test 1, Form 3, Level E), using a black seventh grade group (600 examinees) and a white seventh grade group (600 examinees), separately. Computer program LOGIST (Wood, Wingersky, & Lord, 1976) was used for the estimation of the item parameters. Then the scale for the black group was equated to the scale for the white group, using a computer program EQUATOR which was developed by the author to perform
equating by the proposed method. The initial values were computed by (6) and (7).

**Equating 2**

The test used in Equating 2 was a 60-item achievement test (Tests of Achievement and Proficiency, Reading Comprehension Subtest, Form T, Level 15). In the same way as in Equating 1, item parameters were estimated for a black ninth grade group (400 examinees) and a white ninth grade group (400 examinees), separately. Then the black group scale was equated to the white group scale as in Equating 1.

**Results and Discussion**

The convergence processes in Equatings 1 and 2 are summarized in Tables 1 and 2, respectively. The convergence criterion adopted was the sum of the absolute values of changes in $k$ and $d$ being less than $10^{-4}$. The last column of each table shows the values of an index of average error $QM$ defined by

$$QM = (Q^*/2m)^{1/2},$$

where $Q^*$ is defined by (8). In the actual computation, only those examinees with ability parameter estimates between 3.1 standard deviations below and above the group mean were used in each group. This range was then divided into 31 intervals of size 0.2 to construct relative frequency distributions $h_1$ and $h_2$ in (8).

The tables show rather quick convergence. In each equating only a few iterations were needed to reach the minimum value of $QM$ to the fifth decimal place. A complete execution of EQUATOR took 7.44 s on an IBM 360/65 in Equating 1 and 12.86 s in Equating 2.

The ratios of the final $QM$ to the initial $QM$ are .403 in Equating 1 and .988 in Equating 2. In addition to the values of $QM$, the equating error for each item $g$ was computed by

$$QMI_g = (QI^*_g/2)^{1/2},$$

where $QI^*_g$ is defined by (9). Using this index, it was found that the amount of error decreased, from the initial stage to the final stage of iterations, in 22 out of 25 items in Equating 1, and in 37 out of 60 items in Equating 2.

The foregoing evaluation shows that the proposed method for equating worked more effectively in Equating 1 than in Equating 2. This implies an advantage of the method: The big reductions of $QM$ and $QMI_g$ in Equating 1 are due to poor estimates of initial values (a previously proposed solution) for $k$ and $d$ caused by such extremely high $b_g$ values as 169.0 and 115.6, and the present method seems to be more robust to the existence of such outliers and errors in the item parameter estimation than the previously developed method. This aspect of ro-
bustness is closely related to another advantage of the method. That is, those items with very high $b_i$ values tend to have very low discriminating power as graphically represented by flat item characteristic curves, implying little influence on the change of the criterion function and the result of equating. In other words, highly discriminating items tend to affect equating more strongly than less discriminating items.

A possible modification of the method to make it more robust to the existence of outliers may be to remove those items that have very high $QMI_i$ values from the equating process. This may be achieved either at the computation of initial values or during the iteration process.

Incidentally, the value of the criterion function $QM$ can be used as an index of goodness of fit of the logistic model to the items and the examinees involved. Similarly, $QMI_i$ can be regarded as an index of the contribution of item $g$ to the lack of fit of the model. Hence, removing those items with high $QMI_i$ values not only improves equating but also helps to construct an item bank which the logistic model fits better.

**References**


(Received April 25, 1980)