

IRT Equating Using Parameterized Test Characteristic Curves

(Presented at the annual meeting of NCME, May 2010, Denver, CO)

Alan Nicewander
Pacific Metrics, Inc.

IRT equating methods may be thought of as having a basis in the fundamental indeterminacy in IRT models. IRT models for multiple-choice (MC) and polytomous, constructed-response (CR) contain exponents of the type,

$$a_i(\theta - b_i). \quad (1)$$

Where a_i is the slope parameter for the i^{th} item and b_i is a location parameter (actually, a category boundary parameter for CR items). The measurement scale for θ is indeterminate, which can be seen by linearly transforming θ in the above to produce,

$$\theta^* = A\theta + B. \quad (2)$$

One may now absorb this linear transformation of θ into the items parameters a_i and b_i , viz.,

$$a_i^* = Aa_i \quad (3)$$

$$b_i^* = (b_i - B)/A. \quad (4)$$

Then

$$a_i(\theta^* - b_i) = a_i^*(\theta - b_i^*) \quad (5)$$

Because the two exponents in (5) are equal, this equation establishes that the scale for θ is indeterminate. In practice, this indeterminacy is resolved by fixing the mean and variance for θ —usually at values of zero and one, respectively. If a group of common anchor items are imbedded in two test forms administered to different groups of persons, one would not expect the parameter estimates, a_i and b_i , for the common items to be the same for the two groups—even though, θ has been scaled to have zero mean and unit variance in both groups during item calibration. If, using the notation above, θ^* and θ actually have different means and variances, these differences will be absorbed in the estimates of a_i and b_i , as in (3) and (4), during item calibrations. In order to place the parameter estimates on the same scale, say the scale for θ , we need to solve for A and B in Equations (3) and (4) above:

$$A = a_i^*/a_i \quad (6)$$

$$B = b_i - Ab_i^* \quad (7)$$

A and B are often referred to as the “equating constants”; specifically, the equating slope and intercept, respectively. These equating constants allow one to place a_i^* and b_i^* on the same scale as a_i and b_i (which is the scale for θ) as follows:

$$\frac{a_i^*}{A} \quad (8)$$

places a_i^* on the scale of a_i , and

$$Ab_i^* + B, \quad (9)$$

places b_i^* on the same scale as b_i . In order to obtain the above results, solve (6) for a_i , and (7) for b_i .

Since common-item sets for equating invariably contain more than a single item, two, IRT equating methods based on a *set* of common items are reviewed. For what is to follow, assume that one has item parameter estimates for a set of common items that are imbedded in two different test forms that have been administered to two different groups. Designate one of these data sets the *anchor* data, and the other the *target* data. There are two well-known methods for solving for the equating constants, A and

B, that are based on the first two moments of the a-and-b values for the anchor-and-target data. The Mean/Mean Method uses the means of the a-and-b values for the two data sets, and the Mean/Sigma uses the mean and standard deviation of the b-values for the anchor-and-target data. Detailed descriptions of these two methods are given by Kolen and Brennan (2004). The so-called Characteristic Curve Methods are of major interest here, and are briefly described below.

Characteristic Curve Methods for IRT Equating

Haebara and Stocking-Lord Equating Methods. The Haebara and Stocking-Lord methods are similar in that they solve for the A and B that minimize the sum of squared distances between the characteristic curves for the common items in the anchor and target data sets. The difference is that Haebara minimizes the sum of the distances between each pair of common-item characteristic curves, and Stocking-Lord minimizes the sum of the distances between the two, *test* characteristic curves. The advantage of these methods is that all an item's parameters are simultaneously considered in the process of estimating the equating constants, A and B. These methods generalize to common-item sets that contain CR items. (See Kolen & Brennan, 2004).

Marginal Maximum Likelihood IRT Equating using Bilog-MG and Bilog-MG-3. This very flexible IRT software package (Zimowsky, Muraki, Mislevy & Bock, 2007) estimates the equating constants, A and B, for two or more test forms containing a set of common items. The equating constants are estimated simultaneously with the item parameters using the method of marginal maximum likelihood. At the present time, Bilog-MG-3 is limited to MC items only, but by fall of 2010, a modified version should be available that will handle a mix of MC and CR items.

A New Characteristic Curve Method for Common-Item Equating—Parameterized TCC Equating. The characteristic curve equating method proposed here is based on parameterized approximations to the two TCCs for the common items. Nonlinear least squares is used to fit three-parameter logistic (3-PL) curves to the proportion-correct TCCs for the anchor-and-target data. Once parameterized, it is easy to solve for the equating constants, A and B. Let the approximate, fitted, proportion-correct TCCs for the anchor items be represented as,

$$\frac{1}{n} \sum_i P_i(\theta) \cong \gamma + \frac{1 - \gamma}{1 + \exp(-\alpha(\theta - \beta))} \quad (10)$$

For the anchor data, and

$$\frac{1}{n} \sum_i P_i(\theta^*) \cong \gamma + \frac{1 - \gamma}{1 + \exp(-\alpha^*(\theta - \beta^*))}, \quad (11)$$

for the target data. It should be noted that these TCCs are not exact 3-PL curves unless all common items are MC and all have identical item parameters. However, experience with this method has shown that 3-PL curves fit TCCs with remarkable accuracy. The non-linear R^2 s associated with the fitted curves are invariably above .999 (and frequently above .9999)—even when the common items are all CR items. This proposed method is quite simple to apply, as shown below:

Once one has least squares estimates, $\hat{\alpha}$, $\hat{\alpha}^*$, $\hat{\beta}$, $\hat{\beta}^*$ and $\hat{\gamma}$, of the parameters in Equations (10) and (11) above, these estimates may then be used to solve for the equating constants, A and B:

$$A = \frac{\hat{\alpha}^*}{\hat{\alpha}};$$

$$B = \hat{\beta} - A \widehat{\beta}^*.$$

(Note: γ can be fitted to both TCCs and an average value used as the lower-asymptote for both curves, but what we have found works best is to estimate γ for the anchor data, and fix this value when fitting the TCC to the target data.) SAS's[®] PROC NLIN is quite convenient for fitting 3-PL curves to TCCs, and requires only a few lines of code.

An advantage of the parameterized TCC method is that it fulfills the *Symmetry Property* of equating transformations (Lord, 1980), and Stocking-Lord does not. The symmetry property requires that if anchor and target data sets are interchanged, then the equating function is the inverse of the original function. In some applications of the Stocking-Lord method, the two equating functions are computed, and the equating constants are averaged.

Performance of the Parameterized TCC (PTE) Equating Relative to Stocking-Lord (S-L). P-TCC equating is now part of our equating software and has been partially tested in operational equating. In our implementation, the nonlinear least squares fitting of the TCCs was done using the Marquardt method (See Press, Flannery, Teukolsky & Vetterling, 1992). In general, the two methods for equating give very similar results. A small simulation study is presented in which the P-TCC method is compared to the S-L method in situations where the equating constants are known. The common items are a set of eight items—five MC items, two 3-category CR items and one 5-category CR items. The item parameters for the anchor data are shown in *Appendix A*; the target item parameters were generated in consonance with (3) and (4) using two sets of known equating constants, $A=.8$; $B=-.5$, and $A=1.5$; $B=.5$. In the simulations containing 10 replications per method, the known equating constants were built into the target data sets plus normally distributed error was added to target item parameter in the following manner:

- 1) Target a-values were computed as $a^*=aA + .1\epsilon$;
- 2) Target b-values were computed as $b^*=(b-B)/A + .05\epsilon$, where $\epsilon \sim N(0,1)$.

The two equating methods were compared by:

- a) Computing the root average squared distance between the equated curves (a weighted average, where the weights were normal probability densities);
- b) The plot of the predicted p-values from the two sets of equated item parameters;
- c) The r^2 associated with the p-value plot;
- d) The root mean squared error (RMSE) associated with the p-value plot;
- f) The bias or mean difference between the two sets of predicted p-values.

Tables 1 summarizes the means and standard deviations for SL and PTE equatings based on the eight-item data sets.

Table 1. Means and standard deviations of the outcome variables for SL and PTE equating in two situations where the true equating parameters were known (n-replications=10 per situation).

True A = .8; True B = -.5						
Method	Mean A (s.d.)	Mean B (s.d.)	Mean r^2 (s.d.)	Mean Dist. (s.d.)	Mean Bias (s.d.)	Mean RMSE (s.d.)
SL	.77(.09)	-.47(.07)	.99(.01)	.00(.00)	.00(.00)	.02(.01)
PTE	.77(.09)	-.49(.04)	.98(.01)	.00(.00)	.01(.00)	.02(.01)
True A = 1.5; True B = .5						
SL	1.49(.08)	.50(.03)	.99(.00)	.00(.00)	.00(.00)	.01(.01)
PTE	1.49(.08)	.49(.03)	.99(.00)	.00(.00)	.00(.00)	.01(.01)

r^2 is the squared correlation between the two sets of predicted p-values for the anchor data and the equated target data (p-values are predicted using a grid of equally-spaced θ -values which are used in conjunction with the anchor item parameters and the equated, target item parameters)

Dist. Is the root, average weighted distance between the equated TCCs.

Bias is the difference in the averages of the two sets of predicted p-values.

RMSE is the root mean squared difference between the two sets of predicted p-values.

References

- Kolen, M. J., Brennan, R. L., (2004). *Test equating methods and practice* (2004). New York: Springer-Verlag
- Lord, F. M., (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Press, W.H., Flannery, B.P., Teukolsk, S. A. & Vetterling, W.T. (1992). *Numerical recipes in pascal*. New York: Cambridge.
- Zimosky, M.F., Muraki, E., Mislevy, R.J. & Bock, R.D. (2007). BILOG-MG-3. Chicago: Scientific Software, Inc.

Appendix A

Item Parameters for the Eight-Items (Anchor Data)

Item	n-categories	a	w ₁	w ₂	w ₃	w ₄
1	2	1.2	-.5	.25	.	.
2	2	1	0	.26	.	.
3	2	.8	-1	.22		
4	2	2	2	.21	.	.
5	2	1	1	.25	.	.
6	3	.8	-1	2	.	.
7	3	.5	-.5	.5	.	.
8	5	.4	-1	-.02	.5	1.5

For two-category (MC) items, w_1 is the b-value, and w_2 is the c-value; for three-category CR items, w_1 and w_2 are the category boundaries, and for the 5-category CR item, w_1 - w_4 are the category boundaries.