Developing a Common Metric in Item Response Theory

Martha L. Stocking and Frederic M. Lord
Educational Testing Service

A common problem arises when independent estimates of item parameters from two separate data sets must be expressed in the same metric. This problem is frequently confronted in studies of horizontal and vertical equating and in studies of item bias. This paper discusses a number of methods for finding the appropriate transformation from one metric to another metric and presents a new method. Data are given comparing this new method with a current method, and recommendations are made.

Suppose that item parameters for a given set of items, all measuring a single ability or trait, have been independently estimated using data obtained from two different groups of examinees. These item parameter estimates will be different because the metric or scale defined by each independent calibration of the items is different. Many applications of item response theory (IRT) not only require adequate fit of the IRT model to the data but also require that item parameter estimates from independent calibrations be expressed in the same metric. Such applications include vertical score-scale equating, horizontal score-scale equating, and item bias studies.

It is possible to find the appropriate transformation of item parameter estimates in one metric to another metric by a number of different methods. This paper will discuss the nature of these scale transformations, survey a number of current methods for finding appropriate transformations, and present a new method and some results of its application.

The Nature of Scale Transformations

Item response theory models $P_i(\theta_a;\alpha_i,\beta_i,\gamma_i)$, the probability of a correct response to item i by a person with ability level θ_a. In typical models, $P_i(\theta_a;\alpha_i,\beta_i,\gamma_i)$ is a function of $\alpha_i(\theta_a - \beta_i)$, where α_i is the item discrimination, β_i is the item difficulty, and γ_i is the probability that an individual of very low ability answers the item correctly. When $P_i(\theta_a;\alpha_i,\beta_i,\gamma_i)$ is a function of $\alpha_i(\theta_a - \beta_i)$, the origin and unit of measurement of the ability (and difficulty) metric are undetermined. That is to say, suppose θ_a is transformed by a linear transformation, producing θ_a^*. Suppose the same linear transformation is applied to β_i to produce β_i^*. Finally, α_i is divided by the multiplicative constant of the linear transformation to produce α_i^*. These transformations will not change the probability of a correct response: $P_i(\theta_a^*;\alpha_i^*,\beta_i^*,\gamma_i)$

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 2, Spring 1983, pp. 201–210
© Copyright 1983 Applied Psychological Measurement Inc.
0146-6216/83/020201-10$1.50

$= P_i(\theta_a;\alpha_i,\beta_i,\gamma_i)$. Notice that no transformation is necessary for the γ_i because γ_i is on the probability metric.

Suppose that there are two tests composed of items all measuring a single trait and that there are some items which appear in both tests. If an item is calibrated (i.e., its parameters are estimated) as part of one test and then calibrated as part of the second test given to a different group, the actual values of the estimates of the parameters will differ because the scales established by the two calibrations differ. However, the relationship between these two scales will be linear, since they differ only in origin and unit of measurement.

If b_{i1} is the estimate of item difficulty from the calibration of item i in Test 1, and b_{i2} is the estimate of the same item difficulty from the calibration of Test 2, b^*_{i2}, the value of b_{i2} transformed to the scale of Test 1, is

$$b^*_{i2} = Ab_{i2} + B \quad , \qquad [1]$$

where A and B are constants of the linear transformation of scale. If estimated item difficulties are transformed by a linear transformation, estimated abilities must be transformed by the same transformation; thus

$$\hat{\theta}^*_{a2} = A\hat{\theta}_{a2} + B \quad . \qquad [2]$$

If estimated item difficulty and ability are transformed by these linear expressions, then estimated item discrimination is transformed by

$$a^*_{i2} = a_{i2}/A \quad . \qquad [3]$$

These transformations do not change $a_{i2}(\hat{\theta}_{a2} - b_{i2})$; consequently, $P_i(\hat{\theta}_{a2};a_{i2},b_{i2},c_{i2}) = P_i(\hat{\theta}_{a2},a^*_{i2},b^*_{i2},c_{i2})$.

The problem of transforming the scales reduces to the problem of finding the appropriate A and B of the linear transformation. If concern were with true values of the parameters on their respective scales, it would be simple to find the correct values of A and B; the values of two or more item difficulties could be plotted and the line passing through them determined. But true values are not observable; only estimates of them are available, and these estimates contain error. The estimated item difficulties will not fall on a straight line but will be scattered around some straight line. All methods of transforming scales attempt to estimate the parameters of this line by various techniques and are applicable to any IRT model where $P_i(\theta_a;\alpha_i,\beta_i,\gamma_i)$ is a function of $\alpha_i(\theta_a - \beta_i)$.

Current Methods

Superficially, the problem of finding the linear relationship between two sets of numbers might seem to call for simple regression techniques. The estimated item difficulties (or abilities) from one calibration might be used as the independent variable and those obtained from the second calibration as the dependent variable. This approach would be incorrect. A regression approach (as distinct from a structural relations approach) assumes that the independent variable is measured without error; this is not the case, however. More importantly, a regression procedure is not symmetric with respect to its treatment of the two estimates of item difficulties. Since there is no reason for emphasizing or favoring one estimate of item difficulty over another estimate of the same item difficulty, a symmetric procedure is required.

An approach by Ironson to this problem treats both sets of estimated item difficulties (or abilities) symmetrically. Ironson (1982) used the first principal component as the line giving the transformation from one set of estimated difficulties to another set. This procedure is not suitably invariant under a change of scale: If all estimated item difficulties in one set are divided by a constant, this does not change the slope of the principal component line by a factor equal to the constant.

Another class of symmetric methods uses the first two moments of the distributions of estimated item difficulties. These methods find the parameters of the linear transformation, A and B, such that the mean and standard deviation of the transformed distribution of estimated item difficulties from the second calibration are equal to the mean and standard deviation of the estimated item difficulties from the first calibration. A simple application of this method is found in Marco (1977) and in Cook, Eignor, and Hutton (1979). Poorly estimated item difficulties may have a serious impact on the computation of sample moments, however, producing a linear transformation that does not fit most of the estimated item difficulties very well. Cook et al. (1979) have attempted to solve this by restricting the range of the difficulties used in computing moments.

Bejar and Wingersky (1981) used a more elaborate approach. Robust methods that give smaller weights to outlying points are used to estimate the moments. Linn, Levine, Hastings, and Wardrop (1980) attempted to reduce the influence of outliers by using weighted moments where the weights are inversely proportional to the estimated standard error of the estimates of the item difficulties. The Bejar and Wingersky (1981) procedure treats all outliers in the same fashion, regardless of their standard error. The Linn et al. (1980) procedure treats all points with the same standard error in the same fashion, regardless of their outlier status.

A procedure has been developed by Lord and Stocking (see Appendix) which attempts to overcome these potential problems. This procedure begins with a weighted estimate of the transformation exactly as in Linn et al. (1980). A robust procedure is then used to give small weights to those values whose perpendicular distance from this initial line is large, and a new line is estimated. The robust weighting is repeated until changes in the perpendicular distances become small. Details of this method are presented in the Appendix. Some results of this method will be described in subsequent sections of this paper. A drawback of all of these "mean and sigma" transformation procedures is that they are typically applied only to the estimated item difficulties. That is, the A and B of the linear transformation of scale are estimated using only the b_i and then applied to transform the $\hat{\theta}_a$ and the a_i. Although this is theoretically correct, better methods may exist which use more of the information available from the calibrations.

A class of methods, called "characteristic curve methods" in this paper, uses more information from calibrations. Each calibration of an item yields an estimated item response function or item characteristic curve $\hat{P}_i(\theta_a) \equiv P_i(\theta_a;a_i,b_i,c_i)$. If estimates were error free, the proper choice of A and B for the linear transformation would cause these two curves to coincide. Haebara (1980) averaged the squared difference between the individual item response functions over a suitable distribution of θ, summed over the items common to the two calibrations, and chose A and B to minimize this sum. Divgi (1980) chose the A and B of the linear transformation to minimize the maximum difference between the sum of item response functions for the first calibration and the sum of the item response functions for the second calibration.

The New Method

The new method falls into the class of characteristic curve methods. An examinee, a, with ability θ_a has a true score ξ_a defined by

$$\xi_a \equiv \xi(\theta_a) \equiv \sum_{i=1}^{n} P_i(\theta_a;\alpha_i,\beta_i,\gamma_i) \quad , \qquad [4]$$

where n is the number of items in the test. The correct linear transformation of scales from two different calibrations of the same test would produce the same true scores for examinee a if the α_i, β_i, γ_i were known. If $\hat{\xi}_a^*$ is the estimated true score obtained from the second calibration of the test after it has been transformed to the scale of the first, then

$$\hat{\xi}^*_a \equiv \hat{\xi}^*(\theta_a) \equiv \sum_{i=1}^{n} P^*_i(\theta) \quad , \qquad [5]$$

where $P^*_i(\theta_a) \equiv P_i(\theta_a;a^*_i,b^*_i,c_i)$. For an examinee the difference $(\hat{\xi}_a - \hat{\xi}^*_a)$ should be small. In practice, it is desirable to choose A and B such that for a suitable group of examinees, the average squared difference between true-score estimates is as small as possible. The function to be minimized is

$$F = \frac{1}{N} \sum_{a=1}^{N} (\hat{\xi}_a - \hat{\xi}^*_a)^2 \quad , \qquad [6]$$

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

This function F, considered as a function of A and B, will be minimized when

$$\frac{\partial F}{\partial A} = \frac{-2}{N} \sum_{a=1}^{N} (\hat{\xi}_a - \hat{\xi}^*_a) \frac{\partial \hat{\xi}^*_a}{\partial A} = 0 \quad , \qquad [7]$$

and

$$\frac{\partial F}{\partial B} = \frac{-2}{N} \sum_{a=1}^{N} (\hat{\xi}_a - \hat{\xi}^*_a) \frac{\partial \hat{\xi}^*_a}{\partial B} = 0 \quad . \qquad [8]$$

Now, using the chain rule of differentiation,

$$\frac{\partial \hat{\xi}^*_a}{\partial A} = \sum_{i=1}^{n} \left(\frac{\partial P^*_i(\theta_a)}{\partial b^*_{i2}} \frac{\partial b^*_{i2}}{\partial A} + \frac{\partial P^*_i(\theta_a)}{\partial a^*_{i2}} \frac{\partial a^*_{i2}}{\partial A} \right) \quad . \qquad [9]$$

Differentiating Equations 1 and 3 gives $\delta b^*_{i2}/\delta A = b_{i2}$ and $\delta a^*_{i2}/\delta A = -a_{i2}/A^2$. Substituting these derivatives into Equation 9 gives the partial derivative

$$\frac{\partial \hat{\xi}^*_a}{\partial A} \equiv \sum_{i=1}^{n} \left(b_{i2} \frac{\partial P^*_i(\theta_a)}{\partial b^*_{i2}} - \frac{a_{i2}}{A^2} \frac{\partial P^*_i(\theta_a)}{\partial a^*_{i2}} \right) \quad . \qquad [10]$$

Also,

$$\frac{\partial \hat{\xi}^*_a}{\partial B} \equiv \sum_{i=1}^{n} \frac{\partial P^*_i(\theta_a)}{\partial b^*_{i2}} \frac{\partial b^*_{i2}}{\partial B} \quad . \qquad [11]$$

From Equation 1, $\delta b^*_{i2}/\delta B = 1$, and substitution into Equation 11 gives

$$\frac{\partial \hat{\xi}^*_a}{\partial B} \equiv \sum_{i=1}^{n} \frac{\partial P^*_i(\theta_a)}{\partial b^*_{i2}} \quad . \qquad [12]$$

The functional form of the partial derivatives of the item response function depends on the mathematical model chosen. Formulas for the partial derivatives for the three-parameter logistic item response function are given in Lord (1980, chap. 4).

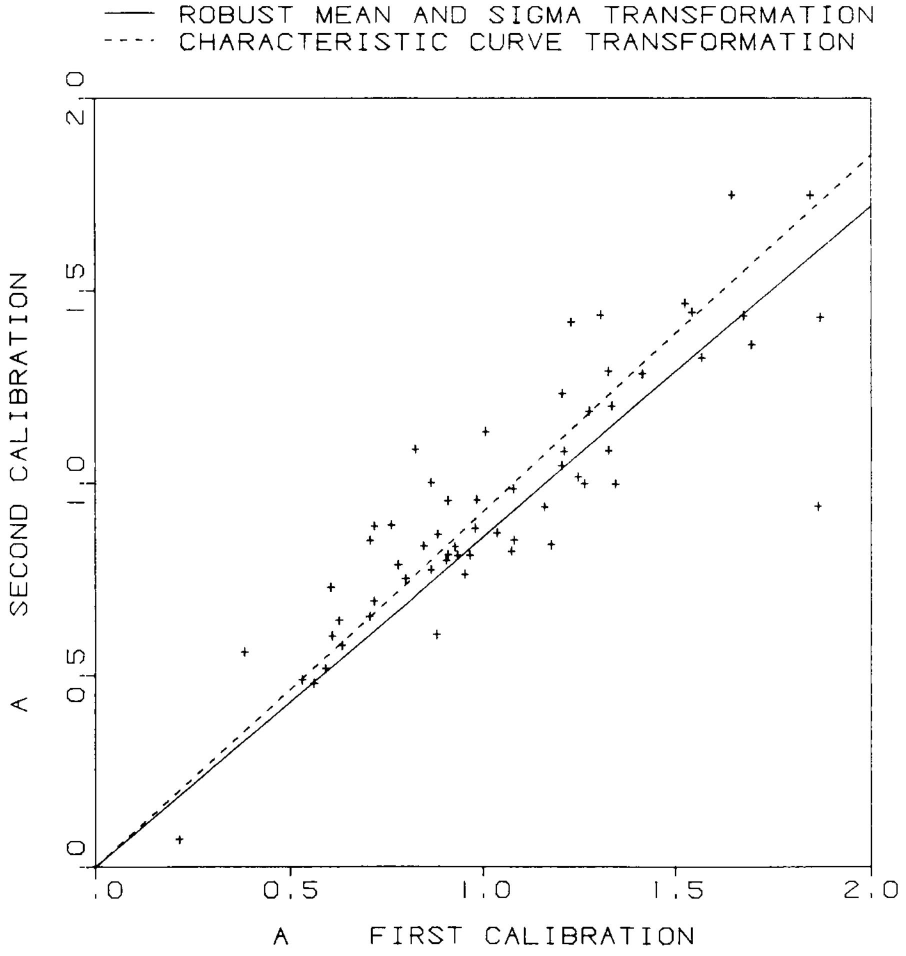
Once the functional form for the item response function is chosen, its derivatives are substituted into Equations 10 and 12. These new expressions are then substituted into Equations 7 and 8 to find the location of the minimum of F in Equation 6.

In practice, the arbitrary group of examinees over which the function is minimized is chosen to have true θ_a equal to the estimated θ_a of a spaced sample of about 200 examinees from the first calibration of a test. The parameters A and B of the linear transformation are found by minimizing F using the multivariate search technique by Davidon (1959) and Fletcher and Powell (1963).

Results

Two transformation techniques—the robust mean and sigma method, developed by Lord and Stocking (see Appendix), and the new characteristic curve method—were compared on more than 20 pairs of tests. Each test was calibrated using the computer program LOGIST (Wingersky, in press; Wingersky, Barton, & Lord, 1982) on data from over 2,000 examinees. Each pair of tests represented either (1) two 125-item tests with 40 or 85 items in common or (2) two 85-item tests with 25 or 60 items in common. The two transformation methods were, of course, applied to the common items.

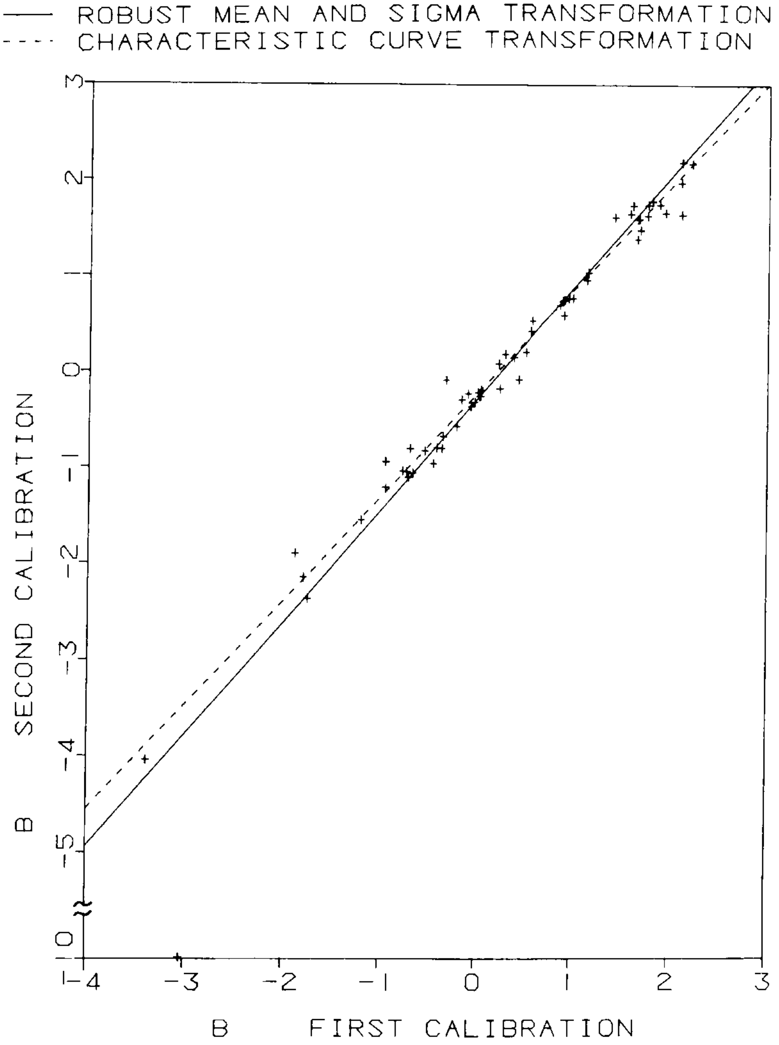
Figure 1
The Two Transformation Methods Compared for Estimated Item Discriminations



In all comparisons, the robust mean and sigma method never provided a better fit to the estimated item difficulties and discriminations; in some cases it provided a worse fit. An example of the latter is shown in Figure 1. The horizontal axis in Figure 1 is the scale of the estimated item discriminations from one calibration of an 85-item test. The vertical axis is the scale of the estimated item discriminations from the calibration of the second 85-item test in the pair. The points plotted are the estimated item discriminations from the 60 items that are common between the two tests of this pair. The solid line through the points is the linear transformation estimated by the robust mean and sigma method. The dashed line is the linear transformation estimated by the new characteristic curve method.

Figure 1 demonstrates that the robust mean and sigma method sometimes produces unsatisfactory results. The solid line does not bisect the point cloud; there are only 17 out of 60 points below the line.

Figure 2
The Two Transformation Methods Compared for Estimated Item Difficulties



The characteristic curve transformation was better; 32 of the 60 points are below the line.

Figure 2 shows the results of this same transformation for the estimated item difficulties. Note that the vertical axis is broken in Figure 2 in order to plot a single point for which the estimated item difficulty from the first calibration was about -3.0 and from the second calibration about -10.0. As an outlier, this point received zero weight in determining the best-fitting line for the robust mean and sigma method. The characteristic curve transformation appears to provide a better fit to the estimated item difficulties here, as well as to the estimated discrimination parameters of Figure 1. These results are not surprising, since the charcteristic curve method uses all of the estimated item parameters in obtaining the transformation, whereas the robust mean and sigma method uses only estimated item difficulties.

The pairs of tests studied could be divided into two groups. Within each group the pairs of tests were related to each other in such a way as to make it possible to construct a chain of 12 transformations performed sequentially. The final test in the chain was the same test (but given to a different sample of people) as the first in the chain. The transformation of estimates from this final test to the scale of the first test should be an identity transformation.

Both transformation methods were run on each chain and the results compared to an identity transformation. The characteristic curve method performed slightly better on one chain, and slightly worse on the other, when compared to the robust mean and sigma method. This result is possibly due to sampling fluctuations.

Conclusions

The characteristic curve method is logically superior to the robust mean and sigma method in that it uses more of the information available from each calibration. This superiority is not always demonstrated in long chains of transformations, probably due to sampling fluctuations. The authors prefer the new method because it avoids serious misfits like that shown in Figure 1.

References

Bejar, I., & Wingersky, M. S. *An application of item response theory to equating the Test of Standard Written English* (College Board Report No. 81-8). Prince ton NJ: Educational Testing Service, 1981. (ETS No. 81-35)

Cook, L. L., Eignor, D. R., & Hutton, L. R. *Considerations in the application of latent trait theory to objectives-based criterion-referenced tests*. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1979.

Davidon, W. C. *Variable metric method for minimization* (Research and Development Report ANL-5990; rev. ed.). Argonne IL: Argonne National Laboratory, U.S. Atomic Energy Commission, 1959.

Divigi, D. R. *Evaluation of scales for multilevel test batteries*. Paper presented at the meeting of the American Educational Research Association, Boston, April 1980. (Revised)

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

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

Ironson, G. Chi-square and item response theory techniques. In R. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore MD: Johns Hopkins University Press, 1982.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. *An investigation of item bias in a test of reading comprehension* (Technical Report No. 163). Urbana IL: Center for the Study of Reading, University of Illinois, 1980.

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

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

Mosteller, F., & Tukey, J. W. *Data analysis and regression: A second course in statistics*. Reading MA: Addison-Wesley, 1977.

Petersen, N. S., Cook, L. L., & Stocking, M. L. *IRT versus conventional equating methods: A comparative study of scale stability*. Paper presented at the meeting of the American Educational Research Association, Los Angeles, April 1981.

Wingersky, M. S. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *ERIBC monograph on applications of item response theory*. Vancouver BC: Educational Research Institute of British Columbia, in press.

Wingersky, M. S., Barton, M. A., & Lord, F. M. *LOGIST user's guide*. Princeton NJ: Educational Testing Service, 1982.

Appendix
Transforming Logistic Scales Using a Robust Iterative Weighted Mean and Sigma Method

This transformation method uses a function of the estimated standard errors of the estimated item difficulties for common items as weights to determine an initial transformation line based on mean and sigma equating of weighted estimates of item difficulties for the common items. A new set of weights is computed using a combination of the estimated standard error weights and robust (Tukey) weights based on perpendicular distances to the line. A new transformation line is computed and the procedure iterates until the maximum change in the perpendicular distances is less than some criterion.

Method

Computing the Standard Errors

The inverse of the information matrix I (Lord, 1980, p. 191) is an approximation to the variance-covariance matrix for the item parameter estimates. The diagonal element of the inverse corresponding to the item difficulty is the estimated variance of the estimate of item difficulty. The square root of this quantity is the estimated standard error of the estimate of item difficulty.

Each item has two estimated item difficulties, one from each calibration. Therefore, each item has two estimated standard errors. The initial weight for an item to be used in the iterative procedure is the reciprocal of the larger estimated squared standard error of the estimated item difficulty.

The accuracy with which an estimated standard error of *b* is computed is the ratio of the determinant to the product of the diagonals of the information matrix. If this ratio is less than .0001, the estimated standard error is not accurate. The item is given a standard error weight of zero.

All people are included in the computation, except those who did not reach the item.

Computing the Mean and Sigma Transformation

There are two distributions of weighted estimated item difficulties, one from each calibration. Let $\underset{\sim}{b}_1$ be the distribution from the first calibration, and b be the distribution from the second calibration and compute

$\overline{X}_{\underset{\sim}{b}_1}$, the mean of $\underset{\sim}{b}_1$;
$\sigma_{\underset{\sim}{b}_1}$, the standard deviation of $\underset{\sim}{b}_1$;
$\overline{X}_{\underset{\sim}{b}_2}$, the mean of $\underset{\sim}{b}_2$; and
$\sigma_{\underset{\sim}{b}_1}$, the standard deviation of $\underset{\sim}{b}_2$.

The mean and sigma transformation (line) to put the second calibration estimated item difficulties onto the scale of the first is

$$\underset{\sim}{b}_2' = A * \underset{\sim}{b}_2 + B \quad , \qquad [A1]$$

where $\underset{\sim}{b}_2'$ is the transformed distribution from the second calibration. For this transformation

$$A = \sigma_{\underset{\sim}{b}_1} / \sigma_{\underset{\sim}{b}_2} \quad ,$$
$$B = \bar{X}_{\underset{\sim}{b}_1} - A * \bar{X}_{\underset{\sim}{b}_2} \quad . \qquad [A2]$$

Computing the Tukey Weights

A method of computing a robust estimate of location by weighting data with differential weights is given by Mosteller and Tukey (1977, p. 205). Only one part of this process, namely, the formula for the weights, will be used.

For purposes of this paper, $Y*$ is the transformation line that has tentatively been found. Tukey's $(Y(i) - Y*)$ is replaced with the perpendicular distance of a point to the line.

Let $D(i)$ equal the absolute value of the perpendicular distance. Then the weights, $T(i)$, are

$$T(i) = \begin{cases} \{1 - (D(i)/CS)^2\}^2 & \text{when} \quad (D(i)/CS)^2 < 1 \\ 0 & \text{otherwise} \end{cases} \qquad [A3]$$

where S is the median of the $D(i)$ and C is a constant equal to 6.

The Iterative Procedure

The iterative procedure is as follows:

Step 1: For the item difficulty of each common item, compute

$$W(i) = [SE(B(i))]^{-2} \quad , \qquad [A4]$$

where $SE(B)$ is the larger of the two estimated standard errors.

Step 2: Compute scaled weights

$$W(i)' = W(i)/(\text{sum of } W(i)) \quad . \qquad [A5]$$

Step 3: Compute the mean and sigma transformation line, using Equation A2, between the two sets of estimated item difficulties weighted by W' from Equation A5, and get the slope, A, and the intercept, B.

Step 4: Compute the perpendicular distances of each point to the line.

Step 5: Compute the Tukey weights, $T(i)$ from Equation A3 for each item, using these perpendicular distances.

Step 6: Reweight each point by a combined weight $U(i)$, where

$$U(i) = (W(i) * T(i))/(\text{sum of } W(i) * T(i)) \quad . \qquad [A6]$$

Step 7: Compute the weighted mean and sigma transformation line using these new weights.

Step 8: Repeat Steps 4, 5, and 6 until the maximum change in the perpendicular distances is less than .01.

Result

This procedure gives low weights to poorly determined item difficulties or to item difficulties which are outliers. Once the final transformation is found for the estimated item difficulties, the estimated item discriminations, as well as the ability estimates, are transformed.

Acknowledgment

This work was supported in part by Contract N00014-80-C-0402, Project designation NR150-453, between the Office of Naval Research and Educational Testing Service.

Author's Address

Send requests for reprints or further information to Martha L. Stocking or Frederic M. Lord, Educational Testing Service, Princeton NJ 08541, U.S.A.