A Minimum Chi-Square Method for Developing a Common Metric in Item Response Theory

D. R. Divgi
Syracuse University

The θ scale in item response theory has arbitrary unit and origin. When a group of items is calibrated twice, estimates from one calibration must be transformed to the metric of the other. A new method is presented for doing so. It is simpler than an earlier method based on test characteristic curves, and makes more complete use of available information.

The metric of the θ scale in item response theory (IRT) is arbitrary. In applications such as equating and bias analysis, a test or subtest is calibrated separately in two samples. It is necessary to transform one set of estimates to the metric of the other. Stocking and Lord (1983) have presented two methods for doing so. The objective of this article is to present a new procedure that combines the desirable features of their methods, and is simpler than the one they recommended.

Method

Estimates from the second calibration are to be transformed to the metric of the first. The transformed estimates of discrimination (a) and difficulty (b) parameters are given by

$$a_2^* = a_2/A \quad , \tag{1}$$

$$b_2^* = Ab_2 + B \quad , \tag{2}$$

for each item, where * indicates a transformed value and the subscript refers to the calibration. (Since true values of parameters do not enter the discussion, there is no need for carets to identify estimates.) The task is to determine the coefficients A and B.

Stocking and Lord's (1983) "mean and sigma" method chooses A and B such that b_2^* has the same mean and variance (over items) as b_1. Means and variances are computed using an iterative robust procedure. The shortcoming of this method is that it ignores the information contained in estimates of discrimination parameters. On the other hand, it does take standard errors of estimates into account.

In their "characteristic curve" method, Stocking and Lord (1983) minimize the mean squared difference between the two test characteristic curves—one based on estimates from the first calibration,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 4, December 1985, pp. 413–415

0146-6216/85/040413-03$1.40

and the other on transformed estimates from the second. This method includes both as and bs in the calculations, but does not involve standard errors of the estimates.

The method being proposed uses the 2×2 covariance matrix of sampling errors for each item (Lord, 1980, p. 191). For a given item, let $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ be the values of this matrix from the two calibrations. When a_2 and b_2 are transformed, so are the diagonal elements of $\boldsymbol{\Sigma}_2$—the aa element is divided and the bb element multiplied by A^2. Denote this transformed matrix by $\boldsymbol{\Sigma}_2^*$. Then, the quadratic form is calculated:

$$(a_1 - a_2^* \quad b_1 - b_2^*)\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2^*)^{-1}\,(a_1 - a_2^* \quad b_1 - b_2^*)' \quad , \tag{3}$$

The transformation coefficients are those values of A and B that minimize the sum Q of these quadratic forms over all items.

The primary advantage of minimizing Q rather than Stocking and Lord's (1983) criterion function is that Q is a quadratic function of B. The equation $dQ/dB = 0$ is linear in B, and hence, easily solved to express B as a function of A. Let $\mathbf{T}$ be the matrix $(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2^*)^{-1}$ for an individual item. Then,

$$B = \Sigma[T_{ab}(a_1 - a_2/A) + T_{bb}(b_1 - Ab_2)]/\Sigma T_{bb} \quad , \tag{4}$$

where each sum is computed over all items. When this value of B is substituted in the expression for Q, there is a minimization problem with only one unknown, which is easy to solve iteratively. (It can be solved by trial and error if an interactive computer program is available.) Stocking and Lord's (1983) characteristic curve method requires a complicated multivariate search procedure.

Moreover, in contrast to the characteristic curve method, which requires a sum over examinees in each iteration, the information matrices are calculated only once. (It is not necessary to use the entire sample for this purpose. As in Stocking and Lord's, 1983, procedure, spaced samples of 200 examinees will suffice.) The proposed method uses estimates of both difficulties and discriminations, and also their standard errors; thus it combines the desirable features of both methods presented by Stocking and Lord. Expression 3 equals the chi-square statistic for item bias when the fitted values of A and B are used (Lord, 1980, chap. 14). Therefore, the new procedure will be called the "chi-square method."

As far as the metric transformation is concerned, the guessing (i.e., c) parameters are a nuisance. Their estimates are not affected by the change of metric, so it can be argued that their role should be minimized. In Stocking and Lord's (1983) characteristic curve method, this could be done by setting c estimates from one calibration equal to those from the other. Another possibility would be to leave them completely out of the calculation of item characteristic curves (ICCs), that is, to compute ICCs using the two-parameter model.

In the chi-square method, estimates of cs enter through the sampling covariance matrices. Initially, the 3×3 information matrix for each item is calculated (Lord, 1980, p. 191). The asymptotic covariance matrix equals the inverse of the information matrix. However, as Wainer and Thissen (1982) have pointed out, the resulting variances are too large. The reason is that estimation algorithms do not use unrestricted maximum likelihood; restrictions are imposed to avoid divergences and unreasonable values. Unfortunately, there are no formulas for obtaining more realistic estimates. A reasonable compromise, used in the example presented below, is the following. Use nonzero estimates of cs while computing the information matrix, but invert only the 2×2 submatrix corresponding to the a and b parameters. (Although this decision is not based on theoretical principles, neither is any method for estimating A and B. It suffices that the procedure and its results be reasonable.)

Illustration and Discussion

The chi-square and characteristic curve methods were applied to 17 items common to two levels of a vocabulary test. (Since ability estimates for examinees were not available, each sum over a sample was

replaced by the corresponding numerical integral over a standard normal distribution of ability.) The mean value of a_2 was divided by mean a_1 to obtain the initial estimate of A. The chi-square method resulted in $A = 1.037$ and $B = -.705$. The values from the characteristic curve method were $A = 1.006$ and $B = -.716$. These changed to 1.025 and $-.718$, respectively, when the latter method was modified by replacing cs from the second calibration with those from the first.

There are no theoretical criteria for choosing among different methods, or for evaluating the quality of a particular method. Stocking and Lord (1983) mentioned the following informal approach. Draw a scatterplot of discrimination estimates from the two calibrations, and add the line represented by the linear transformation. Do the same with difficulty estimates. In both cases, the line should bisect the scatterplot, that is, about half the points should be on either side of the line. In the present illustration there were six plots obtained from three methods for estimating the transformation. With the first calibration along the vertical axis, 7 of the 17 points were above the line in one scatterplot, 8 in each of the others. Thus, all three methods produced satisfactory results.

A proper comparison of methods can be carried out only by applying them to a variety of real data sets. This was not possible in the present study. It is likely that, in most cases, different methods yield practically the same results. Until an extensive comparison is carried out, the choice among available methods remains a matter of judgment. As explained above the chi-square method is simpler and cheaper than the characteristic curve method, and makes use of information about sampling errors.

References

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

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

Wainer, H., & Thissen, D. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397–412.

Author's Address

Send requests for reprints or further information to D. R. Divgi, Center for Naval Analyses, Box 16268, Alexandria VA 22302, U.S.A.