Linking Multidimensional Item Calibrations

Tim Davey, ACT

T. C. Oshima and Kevin Lee, Georgia State University

Invariance of trait scales across changing samples of items and examinees is a central tenet of item response theory (IRT). However, scales defined by most IRT models are truly invariant with respect to certain linear transformations of the parameters. The problem is to find the proper transformation that places separate calibrations on a common scale. A variety of procedures for estimating transformations have been proposed for unidimensional models. This paper explores some issues involved in extending and adapting unidimensional linking procedures for use with multidimensional IRT models. *Index terms: equating, item response theory, linking, metric in IRT, multidimensional IRT, scale linking.*

Many applications of item response theory (IRT) require that parameter estimates obtained from separate calibrations be placed on the same trait scale. For example, IRT-based indexes of differential item functioning cannot be properly interpreted unless the item parameters estimated independently from the reference and focal groups have been linked to the same trait scale (Lord, 1980). Computerized adaptive testing also depends on linking procedures to insure that each item in a bank is calibrated with respect to the same trait metric (Wainer, 1990).

Although invariance of trait scales across changing samples of items and examinees is a central feature of IRT (Lord, 1980), scales defined by most IRT models are truly invariant only with respect to certain linear transformations of the parameter estimates. The same item parameters estimated from separate examinee samples can differ to the extent that the trait distributions differ across samples. Scale linking procedures are intended to correct for these differences by finding the proper linear transformations that place the separate calibrations on a common scale.

A variety of procedures for determining scale transformations have been proposed and evaluated for use with unidimensional IRT models (Divgi, 1985; Haebara, 1980; Lord, 1980; Petersen, Cook, & Stocking, 1981; Stocking & Lord, 1983; Vale, Maurelli, Gialluca, Weiss, & Ree, 1981). Several of these procedures are described below in the course of extending and adapting them for use with multidimensional IRT models.

Indeterminacy in Item Response Models

Item and trait parameters estimated from different examinee samples are likely to be on different metrics because the origin, and often the scale, of the trait metric cannot be uniquely determined. An examination of the unidimensional three-parameter logistic model (3PLM; Birnbaum, 1968) reveals why this is so. The 3PLM gives the probability of a correct answer to item i by an examinee with trait θ_j as

$$P_i(\theta_j) = c_i + \frac{1 - c_i}{1 + \exp\left[-1.7a_i(\theta_j - b_i)\right]}, \quad (1)$$

where

a_i is the item discrimination parameter,
b_i is the item difficulty parameter,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 4, December 1996, pp. 405–416

0146-6216/96/040405-12$1.85

c_i is the probability of an examinee with a very low θ level answering the item correctly by chance,
-1.7 is a scaling constant,
$i = 1, ..., n$, and
$j = 1, ..., N$.

Because item and trait parameters appear in the exponent in the denominator as $a_i(\theta_j - b_i)$, response probabilities are unchanged if a constant, β, is added to both the b and θ parameters. This shifts the location or origin of the trait scale but leaves the response probabilities unaffected. Because these probabilities govern how well the model predicts the observed item responses, model fit is also unaffected. When a single test is administered to a single sample of examinees, there is therefore nothing in the data that allows determination of whether examinees did well because they were capable or because the test was easy. Like the classical item statistics of passing rates and item/test correlations, IRT estimates of b and a are very much sample dependent.

Under the 3PLM, the scale of the θ metric can also be adjusted by multiplying b and θ by a constant, α, and simultaneously dividing a by the same constant. Combining the adjustment of scale with the shift in origin results in the following linear transformations that leave response probabilities invariant:

$$a^* = \frac{1}{\alpha} a, \tag{2}$$

$$b^* = \alpha b + \beta, \tag{3}$$

and

$$\theta^* = \alpha\theta + \beta. \tag{4}$$

Model parameters cannot be uniquely estimated until location and scale indeterminacies have been resolved. This is accomplished by requiring that obtained estimates satisfy certain conditions or constraints. For example, θ or b estimates ($\hat{\theta}$ and $\hat{b}$) may be constrained to have a mean of 0. Imposing these constraints "identifies" the model and allows the remaining parameters to be estimated. The particular constraints imposed determine the scale on which the item and trait parameters are expressed. This dependence is most easily seen when the model is identified by setting the mean and variance of the examinee $\hat{\theta}$s to 0 and 1, respectively.

Consider two examinee samples—the first of high θ level on average, and the second generally low in θ level. Assume that an identical test has been administered to both groups, and that item and trait parameters have been estimated separately from each sample. Although both sets of $\hat{\theta}$s are constrained to have the same mean, items are answered correctly far more often in the first group than in the second. These differences would be apparent in the b parameters, which would be consistently lower in the first group. However, the two sets of parameter estimates could be scaled jointly by recognizing that in this case item, not examinee, characteristics should be assumed equal across calibrations. Constraining the bs rather than the θs would properly reflect the fact that the items are identical but that the examinee groups differ.

The situation is only slightly more complicated in the multidimensional case in which examinees are characterized by multiple θ dimensions. The compensatory (or linear) multidimensional item response model (McKinley & Reckase, 1983; Reckase, 1985) gives the probability of a correct response to item i by an examinee with the trait vector $\boldsymbol{\theta}_j$ as:

$$P\left(u_{ij} = 1 \middle| \boldsymbol{\theta}_j, \mathbf{a}_i, d_i, c_i\right) = P_i\left(\boldsymbol{\theta}_j\right) = c_i + \left(1 - c_i\right)\mathbf{L}\left(\mathbf{a}_i^{\mathrm{T}}\,\boldsymbol{\theta}_j + d_i\right), \tag{5}$$

where

$L(\cdot)$ is either the logistic or normal distribution function,

the vector $\mathbf{a}_i^T = a_{i1}, a_{i2}, ..., a_{ik}$ characterizes how well the item discriminates with respect to each of m trait dimensions, and

d_i defines item difficulty.

Like the 3PLM, the multidimensional model is underidentified. Item response probabilities are unchanged if the origin and scale of the θ metric are altered by the following linear transformations of the item and trait parameters:

$$\mathbf{a}_i^* = \mathbf{A}^{-T}\mathbf{a}_i, \tag{6}$$

$$\boldsymbol{\theta}_j^* = \mathbf{A}\boldsymbol{\theta}_j + \boldsymbol{\beta}, \tag{7}$$

and

$$d_i^* = d_i - \mathbf{a}_i^T\mathbf{A}^{-1}\boldsymbol{\beta}, \tag{8}$$

where the $m \times m$ rotation matrix $\mathbf{A}$ adjusts the variances, covariances, and orientation of the θ dimensions (scale), and the $m \times 1$ translation vector $\boldsymbol{\beta}$ shifts the means (location).

Identifying the parameters of multidimensional models requires not only that the location and scale of each θ axis be fixed, but that the correlations between the axes be specified as well. The simplest way of orienting the θ axes is to set one or more a parameters to 0 on a given dimension. However, more elaborate constraints are possible, and in fact desirable. For example, the means of sets of a estimates ($\hat{a}$s) can be set to specified constants, or $\hat{a}$s can be rotated to some "objective" criterion of simple structure (Harman, 1967).

Linking Designs

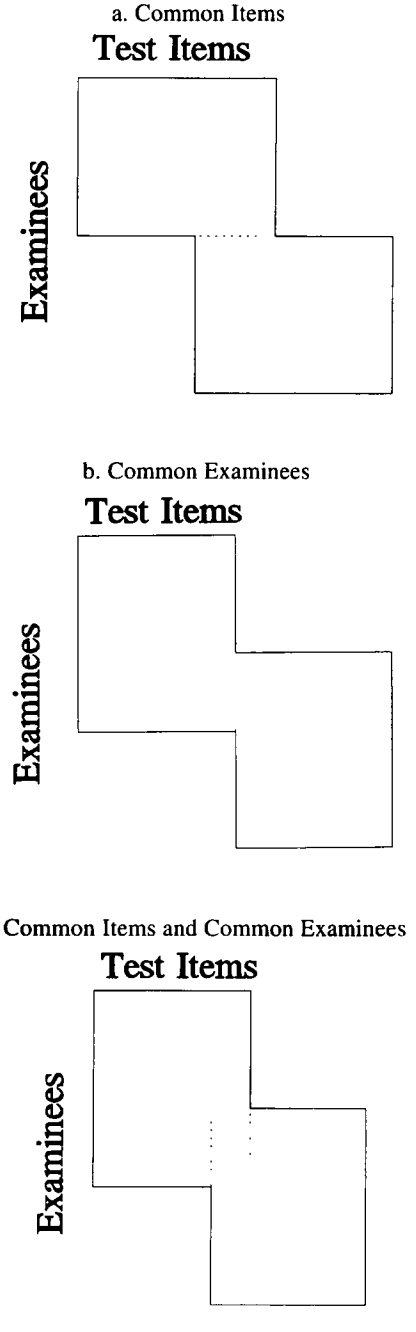
It is possible to link multidimensional item and trait parameters estimated from separate samples to a common scale when the calibration datasets meet one or both of the following conditions (Angoff, 1982):

1. The tests administered to separate examinee samples partially overlap by including some number of items in common.
2. The calibration samples partially overlap by including some number of examinees in common.

Conspicuously absent from this list is a design that administers nonoverlapping test forms to randomly equivalent examinee groups. Although such designs are commonly used to scale and equate unidimensional tests, they are not recommended in the multidimensional case. For it to be possible to link nonoverlapping tests administered to nonoverlapping examinee samples, it would first be necessary to conclude that both tests measured the same number of meaningful traits. No procedures are known for linking tests with different numbers of dimensions. Each dimension would then need to be substantively identified within both tests by inspecting the items "marking" that dimension with high a parameters. The process here is identical to interpreting factor analysis solutions. Once all dimensions are identified, pairs of dimensions would be formed by matching each dimension from one test with its substantive counterpart in the other test. Whether this is possible depends on how clearly the structures of both tests are revealed by the IRT model fit to the data and on how confidently subject matter experts can decide that multiple sets of items are simultaneously parallel. Although all of the above is theoretically feasible, the prospects of successfully accomplishing each step are daunting enough to recommend against randomly equivalent group designs for multidimensional data.

Figures 1a–1c schematically represent some admissible data collection designs. Under the first, and perhaps most frequently encountered, design nonoverlapping groups of examinees are presented tests that have some number of items in common (Figure 1a). The common items are typically called the anchor test. Examinees may or may not be randomly assigned to groups. Test equating applications will often be conducted under this design. Differential item functioning studies also usually follow a special case of this

Figure 1
Linking Designs



design in which the examinee groups are composed of reference and focal examinees, and the anchor test includes all of the items under study.

The second design has nonoverlapping sets of items administered to partially or fully overlapping groups of examinees (Figure 1b). This design might be used when new items are pretested during the administration of an existing operational test. Here, each examinee would receive both the operational test and some set of unscored pretest items.

The final design shown is a combination of the prior two, with both items and examinees partially or fully overlapping (Figure 1c). This design is not uncommon when large numbers of items are pretested and calibrated. Items are often sorted into a set of small item "packets," some number of which are presented to each examinee in an incomplete block design. This results in each item being connected to each other item through both common-item and common-examinee links.

An implicit assumption made by all equating or scale linking procedures is that the tests equated or linked are substantively parallel. Although linking and equating procedures can be computationally applied to arbitrary sets of test items, the results are meaningful only when the items are truly measures of the same trait or traits.

A further qualification, especially relevant to the multidimensional case, is that the sampled examinees differ with respect to the measured traits (Klein & Jarjoura, 1985). It should be emphasized that the observed dimensionality of a test depends on both the items and the examinee population. Consider, for example, a mathematics test with a substantial verbal component. Assume that only a modest level of verbal ability is necessary to comprehend and translate the verbal information into a quantitative form, and that once this translation was properly done excess verbal ability was of no further benefit. The verbal component of this test would be difficult to detect if the test was given only to a group of verbally proficient examinees. The unresolved problem of linking a two-dimensional calibration with a seemingly unidimensional set of item parameter estimates would then be encountered. The same sort of problem occurs when tests are administered to inappropriate samples during an equating study. For example, if high school seniors were used to equate tests designed for a first grade population the tests would be too easy for them and produce score distributions inadequate for equating purposes.

Concurrent Calibration

One application of linking procedures is to put multiple sets of parameter estimates on a common but arbitrary scale. Although this can be accomplished with the common-item or common-examinee approaches described below, it is perhaps best accomplished through concurrent calibration, by which multiple datasets are calibrated simultaneously (Hirsch & Davey, 1990; Mislevy & Bock, 1989). Datasets are simply aggregated or appended and submitted to an item calibration program in a single run. The aggregate data matrices are typically sparse because most examinees will have responded to only some of the items being analyzed. However, because all items are calibrated simultaneously, parameter estimates are automatically reported on a common trait metric.

Concurrent calibration is less conveniently applied to a second basic linking paradigm: putting newly calibrated items on a specific scale defined by some existing item set. This might be required when new items are added to an established item bank. Typically, the scale defined by the previously banked items must remain constant despite the addition of new items. To do this using the concurrent calibration approach, the parameters of the old items must be "fixed" at their banked values while the new items are calibrated. Unfortunately, the item handling capabilities of most calibration programs are quickly outstripped as item banks grow in size. The recourse is to use common-item or common-examinee scaling procedures to estimate the parameters of the linear transformation that places the new items on the trait metric of the established item bank.

Common-Item and Common-Examinee Scale Linking

Common-item or common-examinee scale linking are concerned with finding the rotation matrices, **A**, and translation vectors, **β**, that place parameter estimates from separate calibrations onto a common metric. When sets of parameter estimates differ solely due to model indeterminacy, a linear transformation can be identified that maps one set identically onto the other. Of course, estimates always differ as the result of random error as well. No linear transformation will render the two sets equal in this case. All that can be done is to find scaling values that make the separate estimates similar in some sense.

Several measures of similarity and associated procedures for finding scaling values that maximize these measures are described below. Although different procedures will usually produce different scaling values, some of these values are likely to be preferable to others in the sense of being closer to the "true" scaling parameters that could be found if the IRT parameter estimates were free of random error. Three very different approaches to estimating scaling values are outlined below. The description of each assumes a common-item as opposed to common-examinee or mixed design. It is also assumed that the first sample of item parameter estimates defines the base metric. The intent of scaling is then to transform the second sample estimates to make them similar to the first.

Each method will be illustrated by applying it to data simulating the anchor test (common items) linking design. The anchor test consisted of 40 two-dimensional items. To make the simulation more realistic, the generating item parameters for the anchor test were estimates from real data. Table 1 shows these parameters. Simulated responses to the common items were generated from two examinee groups. Traits were distributed as bivariate normal in both groups, with covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}. \tag{9}$$

However, to make the linking problem more interesting, the mean vector differed across groups. Group 1, which defined the base metric, used [0,0]; the mean vector for Group 2 was [.5, .5].

The simulated data were calibrated from the two groups independently using the multidimensional parameter estimation computer program NOHARM (Fraser, 1987). Parameter estimates from Group 2 were then linked to those of Group 1 by each of the three procedures described below, using the computer program IPLink (Lee & Oshima, 1996).

Matching Scaling Functions

The "*b* equating" procedure widely applied to unidimensional models finds the linking transformation that sets the means, and possibly variances, of the common-item $\hat{b}$s equal across calibrations (Hambleton & Swaminathan, 1985). Let $\hat{b}_{1i}$ and $\hat{b}_{2i}$ denote the item difficulty estimates from Group 1 and Group 2 calibrations, respectively. Let $\hat{b}_{2i}$ denote the transformed Group 2 estimates, that is $\hat{b}^*_{2i} = \alpha\hat{b}_{2i} + \beta$. Then scaling constants are found by solving the system of linear equations:

$$\bar{b}_1 = \bar{b}_2^* = \alpha\bar{b}_2 + \beta, \tag{10}$$

and

$$\sigma^2(b_1) = \sigma^2(b_2^*) = \alpha^2(b_2^2), \tag{11}$$

where σ^2 is the variance. This system is solved by

$$\alpha = \sigma_{b_1}/\sigma_{b_2}, \tag{12}$$

and

Table 1
Generating Item Parameters for the Anchor Test

Item	a_1	a_2	d
1	2.3670	0.0000	2.4910
2	.5770	.3810	.9950
3	.6320	.2650	.6550
4	.9900	.4710	.7640
5	.5970	.1830	.2850
6	.7700	.6430	−.0060
7	.7530	.4090	.5680
8	1.6420	.1470	1.2440
9	1.0400	.5160	.6090
10	1.2560	.5140	.2330
11	1.1710	.1980	1.1100
12	1.2560	.3880	.9190
13	1.7110	.4770	.2370
14	.6920	.9140	−.6760
15	.5710	.7210	−.4360
16	.3310	.4270	−.2750
17	2.0970	.6940	.5910
18	1.1900	1.1550	−.9910
19	.6320	.3980	−.2070
20	1.1120	1.3050	−.6880
21	1.0200	1.1760	−.0370
22	.9560	1.2600	−.4850
23	.5970	.8740	−.5960
24	1.0100	.4690	−.1460
25	.8330	.7910	−1.4690
26	.8140	.7740	−1.0770
27	.8690	.8860	−.9680
28	1.7110	1.7590	−.0650
29	1.1170	1.1620	−.8410
30	.9280	1.3770	−1.1980
31	.7940	1.3570	−.9010
32	1.8670	1.5230	−1.3480
33	.6020	.4770	−.6140
34	.4420	.4050	−.9600
35	1.1520	2.1460	−1.8900
36	.6250	.7790	−.8730
37	.5260	.9660	−1.3120
38	.3060	.9880	−1.5290
39	.5710	2.2700	−3.7730
40	.5640	.4580	−.8230

$$\beta = \bar{b}_1 - \alpha\bar{b}_2. \tag{13}$$

b equating can be extended to the multidimensional case by defining and then solving a more general system of scaling equations (Davey, 1991). These equations specify functions of the common-item (or common-examinee) parameter estimates that are made equal across calibrations by applying the proper choice of $\mathbf{A}$ and $\boldsymbol{\beta}$ to one set of estimates. The left-hand side of each scaling equation is a function of the parameter estimates from the base calibration sample. The right-hand side is the same function of the Group 2 parameter estimates (or, more precisely, the transformed versions of these estimates). For example, let $\mathbf{a}_{1i}$, d_{1i}, and $\boldsymbol{\theta}_{1j}$ denote the parameter estimates for common items and/or examinees from Group 1, and $\mathbf{a}_{2i}$, d_{2i}, and $\boldsymbol{\theta}_{2j}$ represent estimates of these same parameters from the Group 2 calibration. The system of scaling equations then takes the form

$$
\begin{aligned}
h_1(\mathbf{a}_{11}, d_{11}, \ldots, \mathbf{a}_{1n}, d_{1n}; \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{1N}) &= h_1(\mathbf{A}^{-\mathrm{T}}\mathbf{a}_{21}, d_{21} - \mathbf{a}_{21}^{\mathrm{T}}\mathbf{A}^{-1}\boldsymbol{\beta}, \ldots, \mathbf{A}^{-\mathrm{T}}\mathbf{a}_{2n}, d_{2n} - \mathbf{a}_{2n}^{\mathrm{T}}\mathbf{A}^{-1}\boldsymbol{\beta}; \mathbf{A}\boldsymbol{\theta}_{21} + \boldsymbol{\beta}, \ldots, \mathbf{A}\boldsymbol{\theta}_{2N} + \boldsymbol{\beta}) \\
h_2(\mathbf{a}_{11}, d_{11}, \ldots, \mathbf{a}_{1n}, d_{1n}; \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{1N}) &= h_2(\mathbf{A}^{-\mathrm{T}}\mathbf{a}_{21}, d_{21} - \mathbf{a}_{21}^{\mathrm{T}}\mathbf{A}^{-1}\boldsymbol{\beta}, \ldots, \mathbf{A}^{-\mathrm{T}}\mathbf{a}_{2n}, d_{2n} - \mathbf{a}_{2n}^{\mathrm{T}}\mathbf{A}^{-1}\boldsymbol{\beta}; \mathbf{A}\boldsymbol{\theta}_{21} + \boldsymbol{\beta}, \ldots, \mathbf{A}\boldsymbol{\theta}_{2N} + \boldsymbol{\beta}) \\
&\vdots \\
h_q(\mathbf{a}_{11}, d_{11}, \ldots, \mathbf{a}_{1n}, d_{1n}; \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{1N}) &= h_q(\mathbf{A}^{-\mathrm{T}}\mathbf{a}_{21}, d_{21} - \mathbf{a}_{21}^{\mathrm{T}}\mathbf{A}^{-1}\boldsymbol{\beta}, \ldots, \mathbf{A}^{-\mathrm{T}}\mathbf{a}_{2n}, d_{2n} - \mathbf{a}_{2n}^{\mathrm{T}}\mathbf{A}^{-1}\boldsymbol{\beta}; \mathbf{A}\boldsymbol{\theta}_{21} + \boldsymbol{\beta}, \ldots, \mathbf{A}\boldsymbol{\theta}_{2N} + \boldsymbol{\beta}),
\end{aligned}
\tag{14}
$$

where q is the number of elements of **A** and $\boldsymbol{\beta}$ that are to be estimated.

Considerable flexibility is allowed in defining the scaling equations. For example, one equality may require that after transformation of the Group 2 estimates, the average $\hat{a}$ on the first dimension for Items 1–5 is to be the same as that in the base set. Means of the $\hat{d}$ parameters on a given dimension may also be set equal across calibrations. The resulting system of equations is then solved simultaneously for the unknown elements of **A** and $\boldsymbol{\beta}$.

Different systems of linear equations will likely produce different estimates of the scaling parameters. Several factors influence the quality of these estimates. The first is the choice of scaling functions themselves. At a minimum, the functions selected must allow each scaling parameter to be estimated. Systems of scaling functions must therefore not only be consistent, or solvable, but must also include each scaling parameter as an unknown.

Considerations in selecting scaling functions. A major consideration in selecting scaling functions is their stability across random examinee samples. It seems intuitive that stable functions of the common parameter estimates should yield stable estimates of the scaling parameters. Thus, functions based on larger numbers of parameter estimates are preferable to those based on fewer. In a simulation study, Oshima & Davey (1994) found that the quality of the resulting scale linkage does depend strongly on the scaling equations used, with equations involving many estimates strongly preferred to those involving fewer.

A second important influence on scaling parameter estimates is the character of the common-item sets or common-examinee groups. More common items or examinees are preferable to fewer, all other things being equal. Items yielding stable $\hat{d}$s or $\hat{a}$s are also desirable, as are examinees providing stable $\hat{\boldsymbol{\theta}}$s. The nature of the common-parameter set also interacts with the choice of scaling functions. For example, if the scaling functions equated the means and variances of ds, a common-item set composed of items with well estimated $\hat{d}$s would be ideal.

The third and most subtle influence on the quality of scaling parameter estimates are the values of the scaling constants being estimated. The sampling errors of the scaling parameter estimates depend on the true values of those parameters, and extreme scaling values are generally more poorly estimated than less extreme values (Davey, 1992). Like test equating, scale linking is likely to be most successfully applied when it is least necessary.

Example. The matched scaling function method was applied by defining six scaling functions of the two-dimensional item parameters shown in Table 1. Six functions were needed to estimate the four values of the 2×2 **A** matrix and the two values of $\boldsymbol{\beta}$. The functions specified were:

$$
h_1 = \frac{1}{20}\sum_{i=1}^{20} \hat{\mathbf{a}}_{i1} \quad h_2 = \frac{1}{20}\sum_{i=1}^{20} \hat{\mathbf{a}}_{i2} \quad h_3 = \frac{1}{20}\sum_{i=1}^{20} \hat{d}_i \quad h_4 = \frac{1}{20}\sum_{i=21}^{40} \hat{\mathbf{a}}_{i1} \quad h_5 = \frac{1}{20}\sum_{i=21}^{40} \hat{\mathbf{a}}_{i2} \quad h_6 = \frac{1}{20}\sum_{i=21}^{40} \hat{d}_i \quad , \tag{15}
$$

where the first subscript refers to the item and the second subscript refers to the dimension.

The 40 items were split into two blocks of 20 items each, and the means of the $\hat{a}$s and $\hat{b}$s were obtained within each block. These six functions were defined for both groups. **A** and $\boldsymbol{\beta}$ were then found so that the functions computed on the transformed estimates from Group 2 were equal to the same functions of Group 1. IPLink was used to estimate the scaling parameters as

$$\mathbf{A} = \begin{bmatrix} .99 & .08 \\ .02 & 1.00 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} -.58 \\ -.37 \end{bmatrix}. \tag{16}$$

Both calibration examinee samples were drawn from population distributions with the same covariance matrix. Accordingly, no rotation was necessary to link the estimates from Group 2 to Group 1. The obtained **A** was therefore close to identity. However, because the population means were offset, $\boldsymbol{\beta} = [-.5, -.5]$ was expected. The obtained translation was reasonably near that target.

Matching Test Response Functions or Surfaces

Stocking & Lord (1983) proposed linking unidimensional parameter estimates by minimizing differences between test response functions (TRFs), or response functions summed across common items. They began in the usual way, with two estimates of each common item's parameters, one set of estimates from each calibration sample. Each item's parameter estimates define a response function (Equation 1). These can be summed across the common items to produce TRFs. These functions relate examinee θ level to expected number-correct scores on the summed item set:

$$T(\theta) = \sum_{i=1}^{n} P_i(\theta). \tag{17}$$

TRFs, denoted $T_1(\theta)$ and $T_2(\theta)$, are computed for both sets of common item parameter estimates. Scaling values are then determined so that, when applied to the Group 2 parameter estimates, the transformed $T_2(\theta)$ $[T_2^*(\theta)]$ is maximally similar in some sense to $T_1(\theta)$. For example, scaling constants might be found that minimize the difference:

$$\sum_{q=1}^{Q} w_q \left[T_1(\theta_q) - T_2^*(\theta_q) \right]^2 \tag{18}$$

over some grid of θ_qs. The w_q are allowed to differentially weight differences taken at different θ values to recognize that some regions of the θ scale are more important than others. A common choice would be to use w_q proportional to the θ distribution in the base calibration sample.

The extension of Stocking & Lord's (1983) procedure to multidimensional models is straightforward. The multidimensional version of the TRF is a surface formed by summing item response functions. Differences between these surfaces can then be minimized by proper choice of **A** and $\boldsymbol{\beta}$. For example, in the two-dimensional case the difference

$$\sum_{q_1=1}^{Q_1} \sum_{q_2=1}^{Q_2} w_{jk} \left[T_1(\theta_{q_1}, \theta_{q_2}) - T_2^*(\theta_{q_1}, \theta_{q_2}) \right]^2 \tag{19}$$

might be minimized.

Experience in the unidimensional case suggests that minimizing differences between TRFs can produce quite stable estimates of scaling parameters. Reckase, Davey, & Ackerman (1989a, 1989b) reported encouraging results for extensions to multidimensional models.

Example. TRF matching was applied to the same data used previously with the scaling function method. The two-dimensional test response surfaces were evaluated at 49 θ values evenly spaced on the square defined by the corners (–3,–3), (3,3), (3,–3), (–3,3). Differences were equally weighted across the θ values. IPLink estimated the scaling parameters that minimized the squared differences between the surfaces as

$$\mathbf{A} = \begin{bmatrix} .94 & .04 \\ .02 & .99 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} -.53 \\ -.42 \end{bmatrix}. \tag{20}$$

Again, these values were reasonably close to the "true" parameters of the identity matrix and the vector [−.5, −.5].

Minimizing Differences Between Common-Item Parameter Estimates

Divgi (1985) proposed linking unidimensional item calibrations by simply finding scaling values that minimized the sum of squared differences between the two sets of common item parameter estimates. Thus, α and β were found such that

$$\min(\alpha,\beta)\sum_{i=1}^{n}\left(a_{1i} - a_{2i}^{*}\right)^2 + \sum_{i=1}^{n}\left(b_{1i} - b_{2i}^{*}\right)^2, \tag{21}$$

where a_2^* and b_2^* are the transformed parameter estimates from Group 2 (i.e., $a_2^* = a_2/\alpha$ and $b_2^* = b_2\alpha + \beta$.

The straightforward extension of this procedure to the multidimensional case works with the function

$$\min(\mathbf{A},\boldsymbol{\beta})\sum_{i=1}^{n}\left(\mathbf{a}_{1i} - \mathbf{a}_{2i}^{*}\right)^{\mathrm{T}}\mathbf{W}_i\left(\mathbf{a}_{1i} - \mathbf{a}_{2i}^{*}\right) + \sum_{i=1}^{n} w_i\left(d_{1i} - d_{2i}^{*}\right)^2, \tag{22}$$

where $\mathbf{a}_2^*$ and d_2 are, again, the transformed common-item parameter estimates from Group 2. The matrices $\mathbf{W}_i$ and scalars w_i allow differences to be weighted differentially across items and θ dimensions. A reasonable choice would be to make these weights inversely proportional to the asymptotic sampling variances of the parameter estimates.

Equation 22 has strong parallels to methods termed Procrustes rotations (Hurley & Cattell, 1962) that have long been used by factor analysts. Procrustes methods seek a rotation matrix $\mathbf{A}$ that transforms a matrix of factor loadings, $\boldsymbol{\lambda}_2$, to make it as similar as possible in some sense to some specified target loading matrix, $\boldsymbol{\lambda}_1$. A common choice is to find $\mathbf{A}$ that minimizes the least squares difference function

$$\min(\mathbf{A})\mathrm{trace}\left(\boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_2\mathbf{A}\right)^{\mathrm{T}}\left(\boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_2\mathbf{A}\right), \tag{23}$$

subject to the condition that columns of $\mathbf{A}$ are of unit length (i.e., that the diagonal elements of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ are each unity).

Procrustes rotations can theoretically be used to link multidimensional θ scales by first appending the $\hat{\mathbf{a}}$ vectors for the common items as calibrated from both examinee groups into matrices $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, where

$$\boldsymbol{\Lambda}_1 = \begin{bmatrix} \hat{\mathbf{a}}_{11} \\ \hat{\mathbf{a}}_{12} \\ \vdots \\ \hat{\mathbf{a}}_{1n} \end{bmatrix} \quad \boldsymbol{\Lambda}_2 = \begin{bmatrix} \hat{\mathbf{a}}_{21} \\ \hat{\mathbf{a}}_{22} \\ \vdots \\ \hat{\mathbf{a}}_{2n} \end{bmatrix}. \tag{24}$$

$\mathbf{A}$ is then estimated by solving Equation 23. Once $\mathbf{A}$ has been determined, $\boldsymbol{\beta}$ is then found as

$$\boldsymbol{\beta} = \bar{\hat{\boldsymbol{\theta}}}_1 - \mathbf{A}\bar{\hat{\boldsymbol{\theta}}}_2, \tag{25}$$

where $\bar{\hat{\boldsymbol{\theta}}}_1$ and $\bar{\hat{\boldsymbol{\theta}}}_2$ are the mean θ estimates from Group 1 and Group 2, respectively.

Example. Item parameter matching was applied to the evaluation data by using IPLink to directly minimize Equation 22. No differential weighting across items and dimensions was used, so $\mathbf{W}_i$ was an identity matrix and all w_i were unity. The resulting scaling parameters were

$$\mathbf{A} = \begin{bmatrix} 1.02 & .06 \\ .06 & 1.16 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} -.53 \\ -.59 \end{bmatrix}. \tag{26}$$

Like those reported above for the other methods, these values were suitably close to the "true" parameters of the identity matrix and the vector [−.5, −.5].

Conclusion

Although considerable theoretical and practical work needs to be conducted before specific procedures for linking multidimensional trait metrics can be recommended, substantial progress has been made and is continuing. Although the few evaluation studies referenced are limited in scope and generality, work to extend these studies is ongoing.

A software package for carrying out all of the above described procedures has recently become available (Lee & Oshima, 1996). However, each procedure could also be implemented with a middle-level programming language like that found in the more complete statistical packages (SAS, 1991).

Most tests continue to be scored by procedures that assume, either explicitly or implicitly, that a single trait underlies and influences examinee performance. However, this view is generally accepted as a simplistic convenience, with item responses being in fact functions of multiple traits. An item designed to measure a student's skill in mathematics may also require some level of verbal ability. If analyzed in sufficient detail, even relatively simple test items can be shown to tap more than a single trait. It is not claimed that these concerns mandate general application of multidimensional models, because unidimensional models remain sufficient in most practical situations. However, unidimensional models should no longer be assumed simply because more realistic models are not available or are burdensome to apply.

References

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. Rubin, *Test equating* (pp. 55–70). New York: Academic Press.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading MA: Addison-Wesley.

Davey, T. C. (1991, June). *Some issues in linking multidimensional item calibrations.* Presented at the Office of Naval Research Contractors Meeting on Model-Based Psychological Measurement, Princeton NJ.

Davey, T. C. (1992, April). *Optimal common-item anchors in ability metric linking.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9,* 413–415.

Fraser, C. (1987). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory* [Computer program]. Center for Behavioural Studies, The University of New England, Armidale, New South Wales, Australia.

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Harman, H. H. (1967). *Modern factor analysis.* Chicago: University of Chicago Press.

Hirsch, T. M., & Davey, T. C. (1990, June). *Maintaining ability metrics in the face of contradictory information.* Paper presented at the annual meeting of the Psychometric Society, Princeton NJ.

Hurley, J. R., & Cattell, R. B. (1962). The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science, 7,* 258–262.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22,* 197–206.

Lee, K., & Oshima, T. C. (1996). *IPLink: Multidimensional and unidimensional IRT linking* [Computer program]. Atlanta: Georgia State University.

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

McKinley, R. L., & Reckase, M. D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Rep. ONR 83-2). Iowa City IA: The American College Testing Program.

Mislevy, R. J., & Bock, R. D. (1989). *PC-BILOG: Item analysis and test scoring with binary logistic models.* Mooresville IN: Scientific Software, Inc.

Oshima, T. C., & Davey, T. C. (1994, April). *Evaluation of procedures for linking multidimensional item cali-*

brations. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

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

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Reckase, M. D., Davey, T. C., & Ackerman, T. A. (1989a, June). *Evaluating the multidimensional parallelism of five forms of the ACT Assessment mathematics test.* Paper presented at the Office of Naval Research Contractors Meeting on Model-Based Psychological Measurement, Oklahoma City.

Reckase, M. D., Davey, T. C., & Ackerman, T. A. (1989b, April). *Similarity of the multidimensional space defined by parallel forms of a mathematics test.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

SAS Institute Inc. (1991). *SAS language guide* [Computer program manual]. Cary NC: Author.

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

Vale, C. D., Maurelli, V. A., Gialluca, K. A., Weiss, D. J., & Ree, M. J. (1981). *Methods for linking item parameters* (AFHRL-TR-81-10). Brooks Air Force Base TX: U.S. Air Force Human Resources Laboratory.

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillsdale NJ: Erlbaum.

Author's Address

Send requests for reprints or further information to Tim Davey, American College Testing, 2201 North Dodge Street, Iowa City IA 55423, U.S.A. Email: davey@act.org.