Journal of Educational Measurement
Summer 2009, Vol. 46, No. 2, pp. 177–197

Multidimensional Linking for Tests with Mixed Item Types

Lihua Yao[1]
Defense Manpower Data Center
Keith Boughton
CTB/McGraw-Hill

Numerous assessments contain a mixture of multiple choice (MC) and constructed response (CR) item types and many have been found to measure more than one trait. Thus, there is a need for multidimensional dichotomous and polytomous item response theory (IRT) modeling solutions, including multidimensional linking software. For example, multidimensional item response theory (MIRT) may have a promising future in subscale score proficiency estimation, leading toward a more diagnostic orientation, which requires the linking of these subscale scores across different forms and populations. Several multidimensional linking studies can be found in the literature; however, none have used a combination of MC and CR item types. Thus, this research explores multidimensional linking accuracy for tests composed of both MC and CR items using a matching test characteristic/response function approach. The two-dimensional simulation study presented here used real data-derived parameters from a large-scale statewide assessment with two subscale scores for diagnostic profiling purposes, under varying conditions of anchor set lengths (6, 8, 16, 32, 60), across 10 population distributions, with a mixture of simple versus complex structured items, using a sample size of 3,000. It was found that for a well chosen anchor set, the parameters recovered well after equating across all populations, even for anchor sets composed of as few as six items.

Although unidimensional models seem to be sufficient in many practical situations, applying multidimensional item response theory (MIRT) models can be very helpful in trying to better understand what an item measures, what an examinee's proficiency level is on each trait, and how accurately the different composites of ability are being assessed. Multidimensional item information and multidimensional item statistics (i.e., discrimination and difficulty) can also be used to help better understand the data structure and improve test development blueprint classifications (Ackerman, 1994a, 1994b, 1996; Muraki & Carlson, 1995; Reckase, 1985, 1997; Reckase & McKinley, 1991; Yao & Schwarz, 2006). A recent application of MIRT to subscore reporting and classification by Yao and Boughton (2007) compared a MIRT approach to the objective performance index (Yen, 1987) and found that the MIRT approach produced more reliable and robust subscores across all conditions studied. Another advantage of the MIRT application over other subscore augmentation methods (Wainer et al., 2001; Yen, 1987) is that equating can be performed which allows one to compare scores on subscales with fewer items across forms, samples, and years, while borrowing information from the other subscales to reduce estimation error. An assessment with subscale diagnostic score reporting can provide

a much more detailed map of where the student needs remediation and, with equating across years, students' growth in each of the subscales can be tracked. Thus, there is a need for research to develop MIRT models and linking procedures for tests that comprise both multiple choice (MC) and constructed response (CR) items (i.e., mixed format), because many assessments use both types.

In reviewing the research literature on MIRT linking, several studies were found (Davey, Oshima, & Lee, 1996; Hirsch, 1989; Li & Lissitz, 2000; Oshima, Davey, & Lee, 2000; Thompson, Nering, & Davey, 1997). Li and Lissitz parameterized the multidimensional two-parameter IRT model (M-2PL) using the TESTFACT program (Wilson, Wood, & Gibbons, 1987) and developed three sets of MIRT approaches to estimating the scaling coefficients: the test response function (TRF), least squares for estimating translation parameters, and the ratio of eigenvalues for estimating the dilation parameters or contraction coefficient that presents the ratio of the two units. Oshima et al. (2000) used the M-2PL model within the NOHARM program (Fraser, 1987) and studied four linking methods (direct, equated function, TRF, and item response function methods) using a program called IPLINK (Lee & Oshima, 1996). All the studies used real or simulated data from tests with only MC items.

The purpose of this research was to explore parameter recovery rates using an extension of the TRF (Stocking & Lord, 1983) matching approach for the multidimensional case, including both dichotomous and polytomous items under different simulation conditions. To make the simulation study realistic, a two-dimensional confirmatory solution was obtained using real data from a mathematics assessment with 45 MC and 15 CR items. Simulated sample sizes of 3,000 were used along with the real data-derived parameters. The conditions varied in the investigation were population means and variance-covariance matrices, anchor item set length, anchor item type (dichotomous and/or polytomous), and anchor set structure (simple and complex). Simple structure refers to items loading only on one dimension while complex structure refers to items loading on more than one dimension. Root mean squared error (RMSE) and absolute bias (BIAS) were used as evaluation criteria to examine the recovery of item parameters, population parameters, and TRFs under varying conditions.

Methods

Multidimensional Models

Suppose we have N examinees and J test items. The observable item response data are contained in a matrix $X = \{X_{ij}\}$, where $i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, J$. The ability parameters for examinees are a matrix of form $\theta = (\vec{\theta}_1, \ldots, \vec{\theta}_N)$, where each $\vec{\theta}_i$ is a D-element vector where D is the number of subscales or the number of dimensions hypothesized.

The probability of a correct response to dichotomous item j for an examinee with ability vector $\vec{\theta}_i$ for the multidimensional three-parameter logistic (M-3PL; Reckase, 1997) model is

$$P_{ij1} = P(x_{ij} = 1 \mid \vec{\theta}_i, \vec{\beta}_j) = \beta_{3j} + \frac{1 - \beta_{3j}}{1 + e^{(-\vec{\beta}_{2j}^T \vec{\theta}_i + \beta_{1j})}},$$

where $x_{ij} = 0$ or 1 is the response of examinee i to item j. $\vec{\beta}_{2j} = (\beta_{2j1}, \ldots, \beta_{2jD})^T$ is a D-element vector of item discrimination parameters. T here indicates the transpose of the vector. β_{1j} is the difficulty parameter, β_{3j} is the guessing parameter, and $\vec{\beta}_{2j}^T\vec{\theta}_i = \sum_{l=1}^{D} \beta_{2jl}\theta_{il}$ is a dot (inner) product of two vectors. The parameters for the jth item are $\vec{\beta}_j = (\vec{\beta}_{2j}^T, \beta_{1j}, \beta_{3j})^T$.

The probability of a response of $k-1$ to polytomous item j for an examinee with ability vector $\vec{\theta}_i$ is given by the multidimensional version of the partial credit model (M-2PPC; Yao & Schwarz, 2006)

$$P_{ijk} = P(x_{ij} = k-1 \mid \vec{\theta}_i, \vec{\beta}_j) = \frac{e^{(k-1)\vec{\beta}_{2j}^T\vec{\theta}_i - \sum_{t=1}^{k}\beta_{\delta_t j}}}{\sum_{m=1}^{K_j} e^{(m-1)\vec{\beta}_{2j}^T\vec{\theta}_i - \sum_{t=1}^{m}\beta_{\delta_t j}}},$$

where $x_{ij} = 0, \ldots, K_j - 1$ is the response of examinee i to item j. $\beta_{\delta_k j}$, for $k = 1, 2, \ldots, K_j$, are the threshold parameters. $\beta_{\delta_1 j} = 0$, and K_j is the number of response categories for the jth item. The parameters for jth item are $\vec{\beta}_j = (\vec{\beta}_{2j}^T, \beta_{\delta_2 j}, \ldots, \beta_{\delta_{K_j} j})^T$.

Multidimensional Linking

Like unidimensional IRT models, the scale for examinees' ability (or item parameters) in MIRT models has indeterminacy. The D by D transformation matrix (rotation) A and D-element location vector (aligning) $\vec{B}$ can be determined by the following: For an M-3PL item j, let

$$\vec{\beta}_{2j}^* = A^{-1}\vec{\beta}_{2j}, \tag{1}$$

$$\beta_{1j}^* = \beta_{1j} + \beta_{2j}^T A^{-1}\vec{B}, \tag{2}$$

$$\beta_{3j}^* = \beta_{3j}, \tag{3}$$

or

$$\beta_{\delta_k j}^* = \beta_{\delta_k j} + \beta_{2j}^T A^{-1}\vec{B}, \tag{4}$$

for $k = 1, \ldots, K_j$, for M-2PPC items, and let $\vec{\theta}_i^* = A^T\vec{\theta}_i + \vec{B}$, then the probability of obtaining a particular score on jth item is not altered, that is $P_{ijk}(\vec{\theta}_i^*, \vec{\beta}_j^*) = P_{ijk}(\vec{\theta}_i, \vec{\beta}_j)$.

The Stocking and Lord (1983) method was generalized to the multidimensional case for this study. Let J_1 be the number of MC items and J_2 be the number of CR items. We must select a set of quadrature points for numerical integration to estimate the transformation constants in A and $\vec{B}$. Suppose we select the lowest and highest points on the ability scales to be m_1 and m_2, respectively. If Q quadrature points

are chosen in the interval (m_1, m_2) on each dimension then there are a total of Q^D quadrature points, where D is the number of dimensions. The transformation matrix, A, and location vector, $\vec{B}$, can be determined by finding the minimum difference between the TRFs.

The TRF is defined as

$$TRF(\vec{\theta}, \vec{\beta}) = \frac{1}{J_1 + \sum_{j=1}^{J_2}(K_j - 1)} \sum_{i=1}^{N} \left[\sum_{j=1}^{J_1} P_{ij1} + \sum_{j=1}^{J_2} \sum_{k=1}^{K_j} (k-1) P_{ijk} \right].$$

The minimum we must find is

$$\min \left\{ \frac{1}{Q^D} \sum_{i=1}^{Q^D} \left[TRF(\vec{\theta}_i, \vec{\beta}) - TRF(\vec{\theta}_i, \vec{\beta}^*) \right]^2 \right\}.$$

The linking procedure just described was implemented in the LinkMIRT (Yao, 2004) program.[2] In searching for A and $\vec{B}$, the starting point, the number of quadrature points, and the number of iterations all affect the accuracy and can be specified by users.

Simulation Study Design

There are several factors that may affect the accuracy of item and ability parameter recovery when multidimensional linking is employed. The first factor is the accuracy of the parameter estimation software itself. A program called BMIRT (Yao, 2003), which uses Markov chain Monte Carlo (MCMC) methods, was used to estimate the parameters.[3] BMIRT has been found to produce accurate item and ability parameter estimates for tests consisting of both dichotomous and polytomous items (Yao & Boughton, 2007; Yao & Schwarz, 2006) and has been found to be comparable to NOHARM (Yao & Boughton, 2005) under the M-3PL model, with a fixed c-parameter for NOHARM. The second factor is that of sample size. Yao and Boughton (2007) found that a sample size of 3,000 was needed for accurate and stable parameter estimation for tests that are similar to the ones used here and was therefore used for all conditions in this study. The third factor is the number of items and the structure (i.e., simple vs. complex) of those items that load on each dimension. The items loading on each dimension were aligned with the information about objectives assessed as defined in the test blueprint. The fourth factor is the effect of the population distribution across the dimensions (Kahraman & Kamata, 2004; Oshima et al., 2000). Given the limited amount of previous research using real data, this study used 10 different sets of population means and variance-covariance matrices in order to cover a broad range of population distributions that might be found using real data. In fact, it is difficult to find tests that were created to be truly multidimensional and thus it is difficult to know what types of distributions would best represent reality. Thus, the authors used some distributions similar to Davey et al. (1996), in order to build on their research. Some distributions were based on what was found using the "real" mathematics data from this study, which had a correlation of .8 between the two

dimensions, with a variance of .9. The 10 populations were varied by the population mean, correlation, and variances, in order to determine which of these factors has the greatest impact on equating. The populations cover a range of distributions and represent an attempt to cover what might be found in reality. The fifth and final factor is that of the linking program and methodology. Oshima et al. (2000) found that the item and TRF linking procedures produced accurate results (Kolen & Brennan, 1995). Therefore, this simulation study used the TRF method within the LinkMIRT (Yao, 2004) program (IPLINK cannot currently incorporate polytomous items).

Item Parameters

The simulated examinee responses were based on actual dichotomous and polytomous item parameter estimates obtained from a large-scale grade 8 mathematics assessment, with four objectives or subscales. The calibration sample consisted of responses from 10,000 examinees to 60 items, which were fit to a two-dimensional solution represented in Figure 1. As shown in the figure, items for two of the objectives loaded on a single dimension (simple structure) and items for the other two objectives loaded on both dimensions (complex structure).

The item parameters used in this study were based on a two-dimensional confirmatory calibration using BMIRT from the data described above with the specified dimensional loadings shown in Tables 1 (MC items) and 2 (CR items). In running BMIRT, the indeterminacies of the model were solved by fixing the population distributions to multivariate normal with mean vectors of zeros and identity matrices as the variance-covariance matrices, as employed by NOHARM.

The vector plots (Ackerman, 1996; Reckase & McKinley, 1991; Yao & Schwarz, 2006) for the 60 items are shown in Figure 2, with MC and CR items plotted separately for visual clarity. Items loading on dimension one are on the x-axis; items loading on dimension two are on the y-axis; with items of complex structure in the first quadrant having positive difficulty, and in the third quadrant having negative difficulty. Items that are on the 45-degree line measure both dimensions equally well, while items that fall between the 45-degree line and the x-axis are more aligned with dimension one and items that fall between the 45-degree line and y-axis are more aligned with dimension two.

Items only Loading on Dimension 1	**Items Loading on Dimensions 1 and 2**	**Items only Loading on Dimension 2**
Objective 3 Statistics & Probability (12 Items)	***Objective 1*** Number Sense & Computational Techniques (15 Items) ***Objective 2*** Algebra & Functions (15 Items)	***Objective 4*** Geometry & Measurement (18 Items)

FIGURE 1. *Structure of the design.*

TABLE 1
M-3PL Item Parameters

Item	Number of Levels	Item Parameters: Discrimination $\vec{\beta}_{2j}$		Difficulty β_{1j}	Asymptote β_{3j}
1	1	.946	1.600	4.977	.053
2	1	.967	1.475	.476	.164
4	1		2.305	3.537	.107
5	1	.797		−2.143	.175
6	1	1.269		.049	.192
7	1	1.691		.681	.187
8	1	.225	.212	.143	.207
10	1	1.092	1.432	1.552	.249
11	1	.597	.850	.593	.287
13	1		1.941	.975	.042
15	1	.670	.877	5.115	.140
16	1		1.717	1.400	.127
17	1		1.221	.195	.299
18	1	.775	.978	−.656	.105
19	1		1.228	.372	.194
21	1	.921		1.434	.040
22	1	.750	1.363	2.894	.245
23	1	1.140	1.476	1.067	.208
24	1	.291	.588	−.788	.245
26	1	1.629		−.689	.124
27	1	1.576		3.150	.072
29	1		1.364	3.206	.193
31	1		2.518	1.302	.247
32	1	.594	.942	2.125	.172
33	1		3.556	4.132	.246
35	1	1.041	1.345	1.367	.116
36	1		2.252	4.143	.217
37	1	.678	1.315	.686	.265
38	1	.986	1.480	.170	.175
39	1	1.173	1.363	1.846	.213
41	1	.579	.772	1.799	.107
42	1		1.672	.898	.474
44	1		1.361	−.623	.251
45	1	.685	.554	.471	.116
46	1	1.257		−1.765	.139
47	1	1.083	1.358	−1.964	.037
49	1	1.634		.691	.434
51	1		1.992	1.882	.032
52	1	.989	1.193	−1.180	.225
54	1	.676	1.016	−.148	.321
55	1		1.752	.387	.245
56	1	.699		.288	.210
57	1	1.112	1.400	.791	.169
58	1		.572	−1.297	.188
59	1	.747	1.772	3.648	.307

TABLE 2
M-2PPC Item Parameters

Item	Number of Levels	Parameters: Discrimination β_{2j1}	β_{2j2}	Threshold $\beta_{\delta_2 j}$	$\beta_{\delta_3 j}$	$\beta_{\delta_4 j}$	$\beta_{\delta_5 j}$
3	3	.474	.601	.734	−.527		
9	4	1.339		−1.445	.371	−.405	
12	3	.624	.752	2.131	.541		
14	4	.206	.561	−.706	2.162	1.689	
20	5	.514	.990	1.539	1.067	2.117	1.207
25	3	.667	1.016	3.463	1.963		
28	4	1.365		−1.469	−.425	.276	
30	3		.780	2.837	−1.563		
34	4		1.369	−.246	2.109	2.348	
40	5		1.182	−1.964	.186	.686	.897
43	3	1.259		1.228	2.293		
48	4		.973	−.300	−.894	.145	
50	3	.257	.477	2.285	−1.323		
53	4	.610	.769	.125	.662	1.213	
60	5	.468	.729	−2.946	−1.123	2.252	.062

Simulation Conditions

Ten populations with multivariate normal distributions were used to generate examinees' abilities and can be found in Table 3. For each population, 20 replications were sampled, each with a sample size of 3,000 examinees.

Population design. Table 3 lists the means and variance-covariance matrices of the 10 population distributions. This variety of populations allows for a direct comparison of the impact of different means, variances, and correlations across equating samples in order to better generalize to what might be found in real data with multiple dimensions. The populations were all assumed to be multivariate normal. Population 1 was a standard multivariate normal distribution. Compared to population 1, populations 2 and 3 differ only in means, populations 4 and 5 differ only in correlations, and populations 6 and 7 differ only in variances. Compared to population 7, populations 8, 9, and 10 differ in correlations, means, and variances.

Anchor item design. Each anchor set was selected to ensure full coverage of the difficulty distribution. To ensure that we included anchor items with equal loadings on both dimensions and that these items were of good quality (high discrimination), we chose items that had similar a-parameters for each dimension, with slightly higher than average discrimination. Table 4 shows the anchor sets and the number of items in each objective in that anchor set. The anchor sets were ordered by the number of items in each and the objectives covered. Because it is important to span a range of plausible conditions the design of the anchor set varied by anchor length, item type,

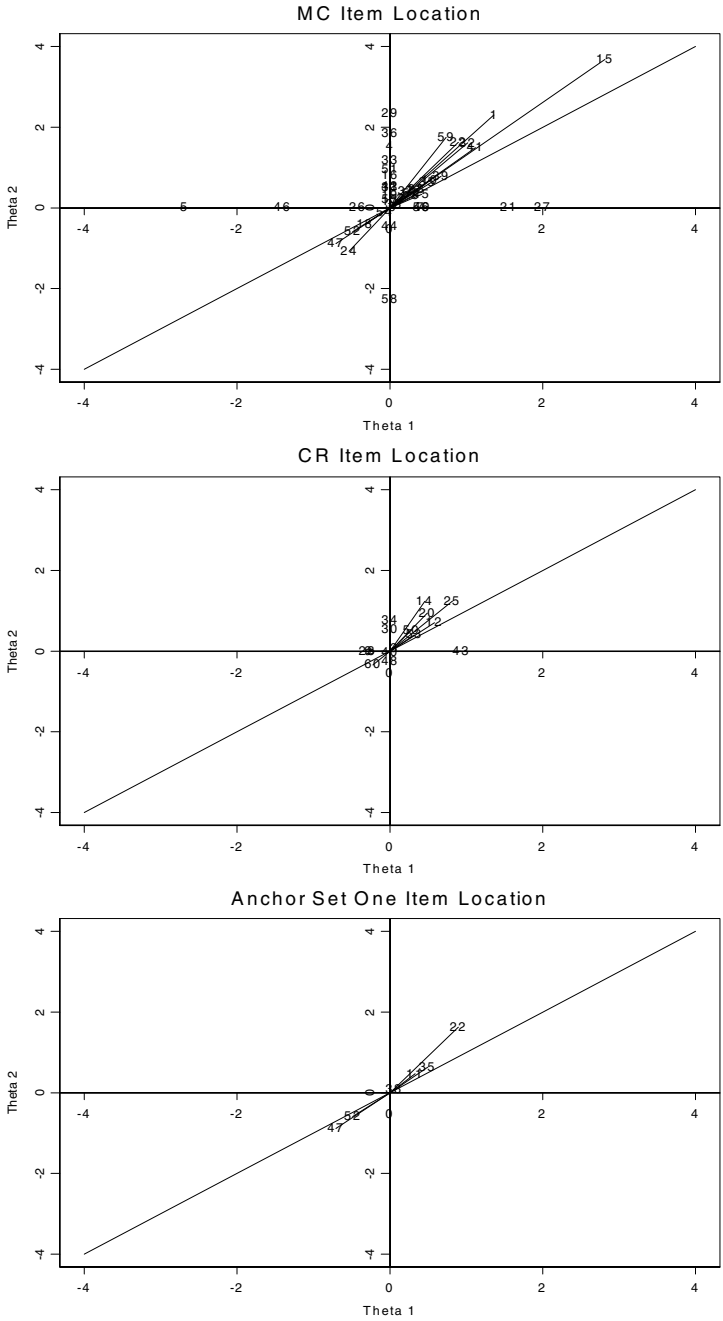

FIGURE 2. *Vector plots for multiple choice (MC) items (top), constructed response (CR) items (middle), and the six items from anchor set one (bottom).*

TABLE 3
Population Parameters

Population	Mean 1	Mean 2	Variance-Covariance	Matrix
1	0	0	1	0
			0	1
2	.5	.5	1	0
			0	1
3	0	−1	1	0
			0	1
4	0	0	1	.5
			.5	1
5	0	0	1	.9
			.9	1
6	0	0	1.5	0
			0	1.5
7	0	0	2	0
			0	2
8	.5	.5	2	1
			1	2
9	−1	0	2	.8
			.8	1.5
10	1	1	2	1
			1	2

TABLE 4
Anchor Sets

Anchor	Objective-One Complex		Objective-Two Complex		Objective-Three Simple-Dimension 1		Objective Four Simple-Dimension 2		
Set	*MC*	*CR*	*MC*	*CR*	*MC*	*CR*	*MC*	*CR*	Type
1	3	0	3	0	0	0	0	0	MC
2	0	0	0	0	3	0	3	0	MC
3	4	0	4	0	0	0	0	0	MC
4	3	1	3	1	0	0	0	0	MC + CR
5	3	0	3	0	1	0	1	0	MC
6	1	0	1	0	3	0	3	0	MC
7	6	0	6	0	0	0	0	0	MC
8	8	0	8	0	0	0	0	0	MC
9	6	2	6	2	0	0	0	0	MC + CR
10	3	1	3	1	3	1	3	1	MC + CR
11	6	2	6	2	6	2	6	2	MC + CR
12	11	4	11	4	9	3	14	4	MC + CR
13(fix)	3	0	3	0	0	0	0	0	Mod 1
14(fix)	0	0	0	0	3	0	3	0	Mod 2
15(fix)	6	2	6	2	0	0	0	0	Mod 9

and item structure. For example, anchor set 1 consists of three items in objectives one and two, respectively, and is complex in structure. Anchor set 2 consists of three items in objectives three and four and is simple in structure. In order to check the effect of adding CR items in the anchor set, anchor set 4 was obtained by replacing 1 MC item with 1 CR item in each objective from anchor set 3. Similarly, anchor set 9 was obtained by replacing several MC items with CR items from anchor set 8. Compared to anchor set 4, anchor set 8 was also created based on anchor set 3, but by adding more MC items in order to retain the same number of score points. Anchor set 12 contains all 60 items and served as a baseline for comparing recovery. The last three anchor sets, which are marked as fixed, had their item parameters set to their true values while estimating the other parameters in the test (i.e., the multidimensional Stocking-Lord was not used). This allowed for a direct comparison of fixed anchor scaling versus the Stocking-Lord extension. To show how the anchor items cover the dimensional composite being measured the six anchor items from anchor set one are plotted in Figure 2 (last graph in series).

Estimation and Linking

Using LinkMIRT, 10 starting points were chosen and, with each, 500 iterations and 30 quadrature points per dimension between −4 and 4 were used. These numbers are large enough to ensure good estimation, with TRF error rounding to .01. Item, ability, and population parameters for all the conditions were obtained by first using BMIRT (10 groups × 20 replications = 200 times) to estimate the parameters and then using LinkMIRT (10 groups × 20 replications × 12 anchor sets = 2,400 times) to scale the estimated item parameters back onto the true parameter scale using the various anchor sets.

Item parameter estimation. For each model condition, using BMIRT, 20,000 iterations were run. The starting values of the parameters were arbitrarily set. The number of iterations was chosen after reviewing the sampling histories for convergence. Given that it was not feasible to check each individual outcome for convergence, a conservative burn-in length of 10,000 iterations was set for each run.

For the multiple choice items (M-3PL), the priors were $\beta_{1j} \sim N(\mu_{\beta_{1j}}, \sigma^2_{\beta_{1j}})$, $\log(\beta_{2jl}) \sim N(\log(\mu_{\beta_{2j}}), \sigma^2_{\beta_{2j}})$, for $l = 1, 2$. $\beta_{3j} \sim beta(a, b)$, and $\mu_{\beta_{1j}} = 0$, $\mu_{\beta_{2j}} = 1.2$, $a = 6$, $b = 16$, $\sigma_{\beta_{1j}} = 1.0$, and $\sigma_{\beta_{2j}} = 0.9$.

For the constructed response items (M-2PPC), the priors were taken to be log-normal for each component of $\vec{\beta}_{2j}$ and normal for $\beta_{\delta_k j}$. The means and standard deviations of the prior distributions were $\mu_{\beta_{2j}} = 1.2$, $\mu_{\beta_{\delta_k j}} = 0$, $\sigma_{\beta_{\delta_k j}} = 1.0$, $\sigma_{\beta_{2j}} = 0.9$, where $k = 2, \ldots, K_j$. The parameters for the priors gave the best acceptance rate.

The indeterminacies of the model in this simulation study were solved by fixing discrimination vector and difficulty of two item parameters (item 5 and 13, which were of simple structure, loading on dimensions one and two, respectively) to $\vec{\beta}_{2,5} = (1, 0)$, $\beta_{1,5} = 0$; $\vec{\beta}_{2,3} = (0, 1)$, $\beta_{1,3} = 0$. The indeterminacies of the model can also be resolved by fixing the population parameters (Yao & Boughton, 2007), although the two methods may produce different levels of accuracy in estimation. Because the

focus of this study is to investigate the accuracy of linking by TRF, using one method for all the conditions should be sufficient.

For each model condition, the transformation matrix A and the transformation location vector $\vec{B}$ were produced using LinkMIRT, and estimates of $\vec{\beta}_j^*$ were obtained using the estimates of $\vec{\beta}_j$ from BMIRT and Equations 1, 2, 3, and 4. The final item parameter estimates (after equating) for each condition ($10 \times 12 = 120$) were obtained by averaging the estimates after transformation across the 20 replications.

Population parameter estimation. For each of the model conditions population distribution parameters $\lambda^* = (\mu^*, \Sigma^*)$ after equating were estimated using the transformation matrix A and the transformation location vector $\vec{B}$ by substituting estimates into

$$\mu^* = A^T\mu + \vec{B}, \quad \Sigma^* = A^T\Sigma A,$$

where $\lambda = (\mu, \Sigma)$ contains the population parameters estimated from BMIRT.

Criteria for Evaluating Results

RMSE was computed for all parameters and was used to examine the parameter recovery rates. Let f_{true} be the true parameter and let f_j be the estimated parameter from sample j, then $RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(f_j - f_{true})^2}$, where n is the number of replications. The RMSE can be decomposed into mean variance and mean squared bias as: $RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(f_j - \overline{f})^2 + (\overline{f} - f_{true})^2}$, where $\overline{f} = \frac{1}{n}\sum_{j=1}^{n} f_j$, is the final estimate. The mean variance measures the variability across replications, while the mean bias measures the deviation of the estimate from the true parameter. The absolute mean bias, $BIAS = \frac{1}{n}\sum_{j=1}^{n}|(f_j - f_{true})|$, was also used to assess the parameter recovery.

Results

Item Parameter Recovery After Equating

The item parameter recovery rates from BMIRT after equating were examined by RMSEs and BIAS. Table 5 lists the BIAS for the MC items, with columns indicating the 12 anchor sets, and rows the 10 populations. For each population and each anchor set, BIAS for the average of the two discriminations and BIAS for the difficulties are listed. Column 15 presents the average BIAS across all the parameters and columns 16–19 show the individual item parameter's average BIAS across the 12 anchor sets for each population. Rows present the results by population and row 23 shows the average BIAS across all populations for each anchor set. Rows 24 through 27 show the individual item parameters' average BIAS for each anchor set across all populations. One can see that anchor sets 2, 6, 10, 11, and 12 performed the best, followed by anchor set 5. Note that the exact same conclusions would be drawn using the average BIAS over all item parameters, as compared to the individual item parameters. It is

TABLE 5
BIAS for 10 Populations and 12 Anchor Sets for MC Items

Population		Anchor Set 1	2	3	4	5	6	7	8	9	10	11	12	Average	β_{2j1}	β_{2j2}	β_{1j}	β_{3j}
1	β_{2j}	.23	.15	.24	.25	.15	.14	.24	.21	.21	.14	.14	.14	.14	.16	.21	.38	.05
	β_{1j}	.48	.31	.54	.40	.31	.31	.46	.46	.40	.29	.30	.29					
2	β_{2j}	.31	.14	.28	.23	.14	.12	.24	.24	.23	.13	.12	.12	.16	.17	.21	.40	.06
	β_{1j}	.53	.36	.54	.41	.34	.36	.41	.41	.46	.33	.33	.32					
3	β_{2j}	.28	.21	.28	.25	.20	.20	.26	.26	.23	.20	.20	.19	.17	.17	.28	.45	.04
	β_{1j}	.68	.34	.61	.45	.33	.33	.60	.62	.48	.31	.32	.32					
4	β_{2j}	.22	.14	.23	.22	.14	.14	.20	.23	.20	.14	.13	.14	.14	.16	.19	.35	.04
	β_{1j}	.53	.30	.42	.37	.30	.29	.41	.39	.40	.28	.28	.27					
5	β_{2j}	.26	.18	.26	.24	.19	.18	.24	.26	.24	.18	.18	.18	.16	.18	.25	.41	.04
	β_{1j}	.54	.30	.52	.44	.37	.29	.53	.58	.49	.28	.27	.28					
6	β_{2j}	.20	.12	.21	.18	.13	.12	.19	.16	.19	.12	.12	.12	.12	.14	.18	.36	.04
	β_{1j}	.50	.28	.51	.38	.30	.27	.44	.38	.39	.28	.27	.27					
7	β_{2j}	.22	.10	.22	.16	.12	.10	.18	.18	.16	.10	.10	.10	.13	.13	.17	.35	.05
	β_{1j}	.50	.27	.55	.36	.30	.27	.36	.41	.39	.27	.28	.28					
8	β_{2j}	.20	.11	.18	.16	.11	.10	.16	.16	.17	.11	.10	.09	.11	.12	.15	.33	.04
	β_{1j}	.46	.26	.42	.37	.27	.25	.41	.41	.39	.24	.24	.24					
9	β_{2j}	.22	.12	.20	.18	.12	.12	.18	.17	.17	.12	.12	.12	.14	.13	.18	.40	.03
	β_{1j}	.63	.25	.63	.44	.31	.24	.57	.53	.48	.25	.25	.26					
10	β_{2j}	.22	.12	.23	.16	.11	.11	.18	.18	.16	.11	.10	.10	.12	.12	.17	.35	.05
	β_{1j}	.47	.30	.43	.36	.29	.30	.39	.45	.37	.27	.27	.26					
Average		.21	.11	.21	.15	.11	.10	.16	.17	.15	.10	.10	.10					
β_{2j1}		.18	.12	.18	.19	.13	.12	.16	.16	.17	.12	.12	.12					
β_{2j2}		.28	.15	.28	.22	.15	.15	.25	.25	.22	.15	.15	.14					
β_{1j}		.53	.30	.52	.40	.31	.29	.46	.46	.43	.28	.28	.28					
β_{3j}		.04	.04	.04	.04	.04	.04	.04	.04	.04	.04	.04	.04					

important to notice that anchor sets with the largest BIAS all contained MC items with complex structure.

Overall, populations 6, 7, 8, and 10 had the smallest BIAS. The means for groups 8 and 10 were closer to the difficulty of the MC items (which was 1.05) and the correlation between the dimensions was .5. Populations 1 and 4 had very similar recovery, both had the same mean, but one had a correlation of 0, while the other had a correlation of .5. Population 5, which had the same mean as groups 1 and 4, had a larger error, possibly due to the higher correlation between the dimensions ($r = .9$). Populations 6 and 7, which had the same mean and correlation as population 1, but higher variance, produced smaller errors. Population 3 had the largest BIAS, most likely a result of the fact that the two means (0 and -1) were farther away from each other. Again, the same conclusion would be drawn, regardless if one used the overall average or the individual item parameter BIAS results.

Table 6, similar to Table 5, lists the BIAS results for the CR items across the 10 populations (indicated by row), for the 12 anchor sets (indicated by column), with the BIAS broken out for the two discriminations and each of the item threshold parameters. Similar observations to the MC items were found for the CR items.

Figure 3 presents the MC item recovery rates in RMSE values on the y-axis and anchor set type on the x-axis for all populations, across all parameters, while Figure 4 shows the CR recovery rates. The recovery rates for both the MC and CR items clearly indicate that anchor sets 2, 5, 6, 10, 11, and 12 had the lowest error. Thus, if a strong anchor set is used (i.e., at least one simple structured item per dimension), then item parameters are well recovered. Comparing the performance of anchor sets 3 and 4 and anchor sets 8 and 9, as expected, having CR items in the anchor set produces better results than using only MC items. However, including more MC items that produce the same number of score points would also reduce the error, although not to the same degree as adding CR items, as shown when comparing anchor sets 8 and 4.

After completing the item parameter recovery using the multidimensional Stocking-Lord procedure, anchor sets 1, 2, and 9 were chosen along with populations 1 and 9 in a fixed anchor scaling for comparison purposes. As shown in Figures 3 and 4, all three sets had the greatest errors and the recovery was not as good as for any of the 12 anchor item sets. From these results, one can see that MIRT TRF linking approach performed better than the fixed item anchor approach.

Population Parameter Recovery After Equating

Table 7 shows the BIAS for the population means. For almost all populations anchor set 12 gave the smallest BIAS. Overall, anchor sets 10 and 11 gave comparable results to anchor set 12, with anchor sets 6 and 2 having slightly larger BIAS than anchor set 12. Anchor sets 1 and 3 produced the largest BIAS, with anchor set 8 following close behind. Notice that anchor sets 1, 3, and 8 comprised only MC items with complex structure. Overall, population 2 had the smallest error rate for the mean, followed by populations 10, 7, and 8. Population 3 had the largest error, followed by populations 9 and 5. Similar BIAS was observed for the variance-covariance matrices.

TABLE 6
BIAS for 10 Populations and 12 Anchor Sets for CR Items

	Anchor Set												Overall Anchor Set						
Population	1	2	3	4	5	6	7	8	9	10	11	12	Avg	β_{2j1}	β_{2j2}	$\beta_{\delta_2 j}$	$\beta_{\delta_3 j}$	$\beta_{\delta_4 j}$	$\beta_{\delta_5 j}$
1	.19	.08	.23	.16	.08	.08	.18	.17	.14	.07	.07	.06	.13	.06	.10	.15	.15	.17	.14
2	.23	.10	.22	.14	.09	.08	.14	.15	.17	.06	.07	.06	.13	.07	.11	.14	.15	.16	.12
3	.30	.11	.27	.18	.11	.10	.26	.27	.19	.09	.10	.09	.17	.05	.11	.19	.20	.25	.23
4	.23	.08	.18	.15	.09	.08	.16	.16	.15	.07	.07	.07	.12	.06	.10	.14	.14	.17	.15
5	.24	.09	.23	.19	.14	.09	.23	.26	.21	.08	.08	.08	.16	.08	.15	.18	.18	.21	.16
6	.20	.07	.21	.14	.08	.06	.17	.14	.15	.06	.06	.06	.12	.05	.07	.14	.14	.17	.13
7	.22	.07	.24	.13	.09	.06	.14	.15	.15	.06	.07	.07	.12	.05	.08	.15	.15	.16	.14
8	.21	.07	.18	.15	.09	.07	.17	.17	.16	.06	.06	.06	.12	.05	.08	.14	.15	.17	.13
9	.30	.08	.30	.20	.12	.08	.26	.24	.21	.08	.08	.09	.17	.05	.08	.20	.22	.26	.20
10	.20	.08	.19	.13	.08	.08	.15	.18	.13	.06	.07	.06	.12	.05	.09	.14	.14	.16	.13
Average	.23	.08	.22	.16	.10	.08	.19	.19	.17	.07	.07	.07							
β_{2j1}	.09	.04	.08	.08	.05	.04	.07	.07	.07	.04	.04	.04							
β_{2j2}	.18	.04	.18	.13	.06	.04	.14	.15	.12	.04	.04	.04							
$\beta_{\delta_2 j}$	.27	.09	.26	.18	.11	.09	.22	.22	.20	.08	.08	.08							
$\beta_{\delta_3 j}$	.28	.10	.27	.18	.11	.10	.23	.23	.20	.08	.09	.08							
$\beta_{\delta_4 j}$	.33	.11	.32	.21	.13	.10	.27	.27	.24	.09	.09	.09							
$\beta_{\delta_5 j}$	.24	.11	.23	.15	.13	.10	.19	.20	.17	.09	.10	.10							

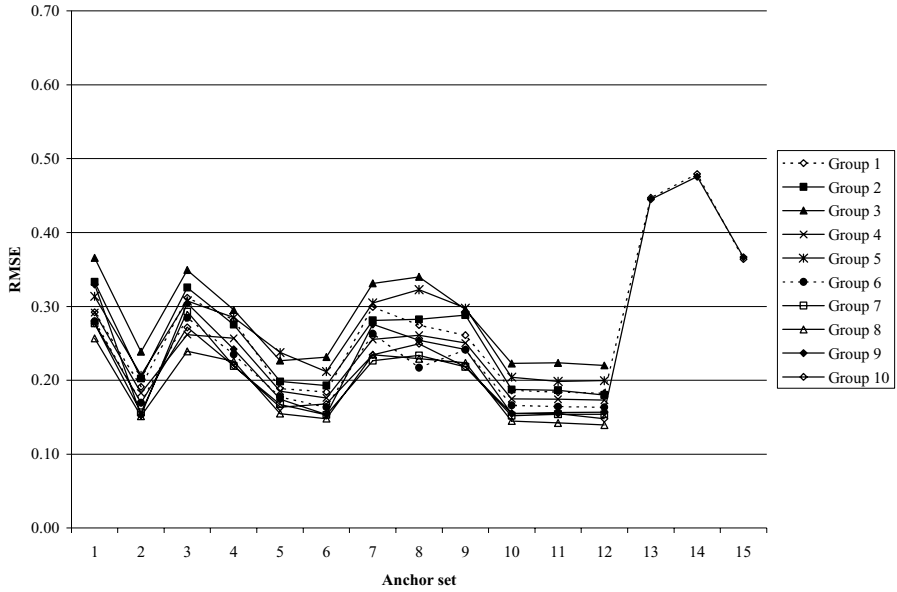

FIGURE 3. *Root mean squared error (RMSE) for populations by anchor sets for multiple choice items.*

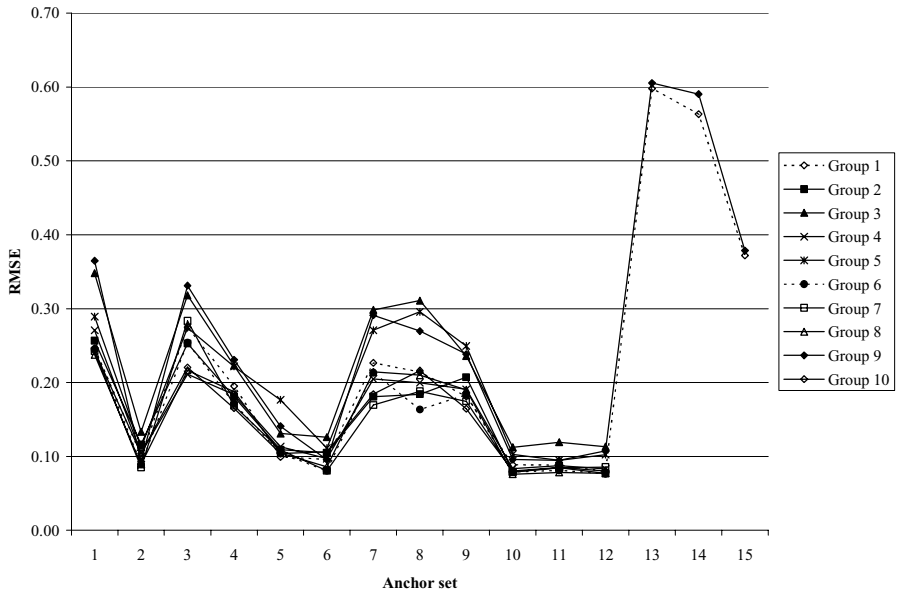

FIGURE 4. *Root mean squared error (RMSE) for populations by anchor set for constructed response items.*

TABLE 7
BIAS for 10 Populations and 12 Anchor Sets for Population Mean

Anchor Set	Population 1	2	3	4	5	6	7	8	9	10	Overall Population
1	.34	.05	.92	.44	.70	.35	.14	.20	.66	.10	.39
2	.08	.11	.12	.08	.07	.05	.06	.05	.05	.08	.07
3	.43	.04	.91	.34	.62	.41	.19	.15	.65	.14	.39
4	.11	.09	.26	.14	.38	.13	.03	.07	.45	.05	.17
5	.05	.03	.09	.05	.14	.07	.05	.06	.17	.05	.08
6	.08	.09	.10	.07	.06	.03	.03	.04	.05	.08	.06
7	.16	.03	.80	.28	.58	.25	.08	.10	.61	.12	.30
8	.28	.02	.94	.26	.69	.17	.11	.22	.62	.05	.33
9	.13	.02	.40	.21	.49	.18	.16	.13	.49	.06	.23
10	.04	.02	.03	.02	.06	.01	.02	.04	.04	.05	.03
11	.05	.06	.02	.03	.02	.04	.02	.04	.06	.05	.04
12	.02	.02	.01	.02	.06	.02	.02	.03	.07	.04	.03
Overall	.15	.05	.38	.16	.32	.14	.08	.09	.33	.07	

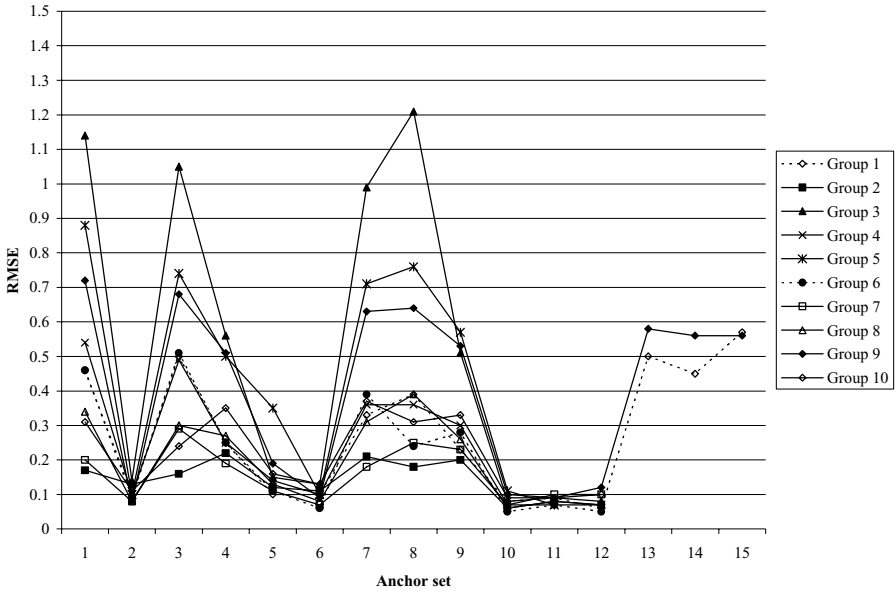

FIGURE 5. *Root mean squared error (RMSE) for populations by anchor sets for population distribution means.*

Figure 5 shows RMSEs across all populations (y-axis), for each anchor set on the x-axis for the means, while Figure 6 shows the RMSEs across all populations and anchor sets for the variance-covariance recovery rates. The best recovery rates were found for anchor sets 2, 5, 6, 10, 11, and 12 across both means and variance-covariance matrices. Interestingly, all population means and variance-covariance values were recovered well for the six best performing anchor sets. For example, after

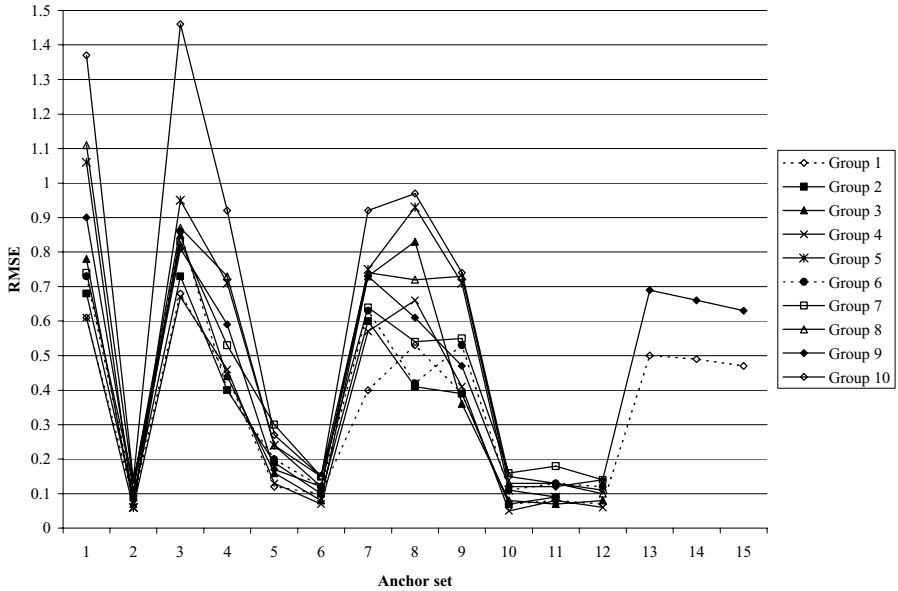

FIGURE 6. *Root mean squared error (RMSE) for populations by anchor sets for population distribution variance-covariance.*

equating using anchor set 12, the true population mean (0, −1) for population 3 was estimated at (−.11, −.91) and the true variance-covariance matrix that was an identity matrix was estimated at $\sigma = \left(\begin{smallmatrix} 1.02 & -.03 \\ -.03 & 1.00 \end{smallmatrix}\right)$. The true mean (0, 0) for population 4 was estimated at (−.07, .05) and the true variance-covariance that had 1s on the diagonals and .5s on the off diagonals was estimated at $\sigma = \left(\begin{smallmatrix} 1.07 & .48 \\ .48 & 1.01 \end{smallmatrix}\right)$.

The true mean (1, 1) for population 10 was estimated at (.89, 1.04) and the true variance-covariance estimates that had 2 on the diagonals and 1 for the off diagonals was estimated at $\sigma = \left(\begin{smallmatrix} 2.03 & .91 \\ .91 & 1.91 \end{smallmatrix}\right)$. Note that these three groups represent the range of recovery rates that were found. The RMSEs of the transformation matrix A and the transformation location $\vec{B}$ across replications are not shown here, but they were very small, ranging from .00 to .09, for all the conditions.

TRF Recovery After Equating

Figure 7 shows the difference between the estimated and the true TRFs for anchor set 12, for populations 3 and 8, with the left-hand graphs showing the contour plots and the right-hand graphs being the surface plots. For the contour plots, 10 levels were plotted. For most of the two-dimensional theta region the TRF differences are small with only a small curve at the lower left with a value of .01. For the surface plots, the range on the z-axis (vertical) is from −.015 to .015. The true TRFs were recovered well for both populations (as expected), with flat surfaces that were close to $z = 0$. Population 8 was better recovered compared to population 3.

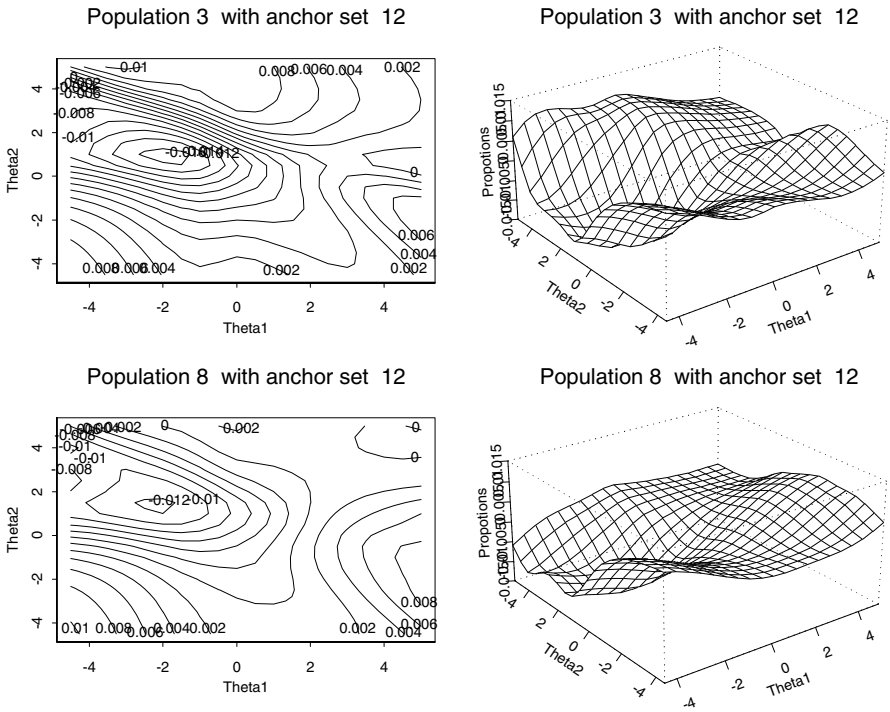

FIGURE 7. *Contour plot and surface plot for the difference of test response functions (TRFs) between the estimated and the true from anchor set 12 for groups 3 (top) and 8 (bottom).*

Discussion and Conclusion

This research used real data item parameters from a confirmatory MIRT modeling approach for two dimensions. The study crossed 12 anchor item sets, including different combinations of simple and complex structured items for 10 populations, with both MC and CR item types. The results of this study show the item parameter recovery rates in comparison to the 60-item anchor baseline (anchor set 12). Anchor set 10, consisting of 16 items (8 items per dimension), and anchor set 11, consisting of 32 items (16 items per dimension), both having a combination of simple and complex structured items, resulted in RMSE or BIAS values that were comparable to the baseline linking recovery rates. Anchor set 5 (2 simple and 6 complex) and anchor set 6 (6 simple and 2 complex items) recovered their true item parameters equally well, with anchor set 6 producing slightly better results. Anchor set 2 (simple structure items) had only 6 items, 3 per dimension, and was also found to have good recovery rates. Overall, anchor sets with simple structured items on both dimensions outperformed anchors with only complex structure or with fewer simple structured items. Replacing some of the MC items with CR items produced better results when linking in two-dimensional space. Adding more MC items to have the same number of score points also produced better results, although to a lesser degree than with CR

items with the same number of score points. Populations with means that are very different (3 and 9) and populations with higher correlations (9) produced larger estimation errors. Populations that have means closer to the difficulty of the test (2, 8, and 10) and populations that have lower correlations between dimensions (4, 8, and 10) produced smaller estimation errors.

The results for the population means and variance-covariance matrices showed that the best recovery rates were for populations 2, 8, and 10 with means of .5 or 1.0 for each dimension and correlations of .0 or .5 between the dimensions. Population 1 recovery rates were not as good as population 2, with a mean of .0 per dimension and a correlation of .0 between the dimensions. This is a direct result of the test mean difficulty being closer to .5, then .0. The population recovery decreases for conditions that have a higher correlation between dimensions (populations 5 and 9), and for means that substantially differ (populations 3 and 9 with a possible prior effect) from each other.

Many assessments are designed to report on a multitude of subskills based on the test blueprint, as seen in the area of diagnostic score reporting. Such data, especially those with CR items may measure the construct at a greater depth and may tend to be more multidimensional in structure (Ackerman & Smith, 1988; Perkhounkova & Dunbar, 1999; Rosa, Swygert, Nelson, & Thissen, 2001). Exploring the multidimensional structure of these assessments and fitting a model to match should provide a much richer profile of information about the construct and student abilities than is currently provided by a unidimensional IRT approach. In fact, Yao and Boughton (2007) found that the MIRT approach produced more reliable and robust subscores than a unidimensional approach. Also, unlike other subscore augmentation methods, as found in this research, a MIRT approach can be supported by actually equating at the subscore level (with only a few items) while borrowing information from the other subscales to reduce estimation error so that subscores can be compared across forms, samples, and years. Writing test items for the purpose of diagnostic subscore and overall composite score reporting across years is very difficult, especially for anchor items (Luecht & Miller, 1992). As found in their study, items that measure only one single trait well are in fact better choices as anchor items, and thus more research is needed on how to best write such items.

Many testing organizations (e.g., CTB/McGraw-Hill and ETS) use a combination of a unidimensional three-parameter logistic and two-parameter partial credit models for testing programs, and to move into a MIRT system would require programs with the capability of simultaneously calibrating MC and CR items using multidimensional models. For multidimensional models to be employed successfully in practice, linking procedures that include both dichotomous and polytomous models are also required and need to be studied systematically. This research builds on two previous linking studies (Li & Lissitz, 2000; Oshima et al., 2000), by extending their research to tests with mixed item formats, different numbers and types of anchor items (MC vs. CR), with simple to complex structured items, and across multiple multidimensional population distributions by using a matching TRF procedure. Although two-dimensional data were studied here, the algorithm and software for both parameter estimation and parameter equating/linking would not need to be modified for use in higher dimensional space; a major strength of the MCMC approach.

The reader is warned, however, that with current computing hardware the run times for more than two dimensions will be significantly longer than for two. For future research, it will be important to compare the approach used here to other linking procedures, as well as other criteria for choosing anchor item sets, all with extensions to more then two dimensions. Using nonnormal multivariate distributions such as multivariate *t* for the populations is another area for future study.

Notes

[1] This paper was written while the first author was at CTB/McGraw-Hill.

[2] Requests for reprints or information about LinkMIRT should be addressed to the first author.

[3] Requests for BMIRT should be addressed to Richard Patz at CTB/McGraw-Hill or to the first author.

References

Ackerman, T. A. (1994a). Creating a test information profile in a two-dimensional latent space. *Applied Psychological Measurement*, *18*, 257–275.

Ackerman, T. A. (1994b). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *20*, 309–310.

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, *20*, 311–329.

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement*, *12*, 117–128.

Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement*, *20*, 405–416.

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

Hirsch, T. M. (1989). Multidimensional equating. *Journal of Educational Measurement*, *26*, 337–349.

Kahraman, N., & Kamata, T. A. (2004). Increasing the precision of subscale scores by using out-of-scale information. *Applied Psychological Measurement*, *28*, 407–428.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Lee, K., & Oshima, T. C. (1996). IPLINK: Multidimensional and unidimensional item parameter linking in item response theory. *Applied Psychological Measurement*, *20*, 230.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement*, *24*, 115–138.

Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement*, *16*, 279–293.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*, 73–90.

Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement*, *31*, 357–373.

Perkhounkova, Y., & Dunbar, S. B. (1999, April). *Influence of item content and format on the dimensionality of tests combining multiple choice and open-response items: An application of the Poly-DIMTEST procedure*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

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

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25–36.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*, 361–373.

Rosa, K., Swygert, K. A., Nelson, L., & Thissen, D. (2001). In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 253–292). Mahwah, NJ: Lawrence Erlbaum.

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

Thompson, T. D., Nering, M., & Davey, T. (1997, June). *Multidimensional IRT scale linking without common items or common examinees*. Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., et al. (2001). Augmented scores: "Borrowing Strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Mahwah, NJ: Lawrence Erlbaum.

Wilson, D., Wood, R., & Gibbons, R. D. (1987). *TESTFACT: Test scoring, item statistics, and item factor analysis* [Computer software and manual]. Mooresville, IN: Scientific Software.

Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory* [Computer software and manual]. Monterey, CA: CTB/McGraw-Hill.

Yao, L. (2004). *LinkMIRT: Linking of multivariate item response model* [Computer software]. Monterey, CA: DMDC DoD Center.

Yao, L., & Boughton, K. A. (2005). *Multidimensional parameter recovery: Markov chain Monte Carlo versus NOHARM*. Unpublished manuscript.

Yao, L., & Boughton, K. A. (2007). A Multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*, 1–23.

Yao, L., & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement*, *30*, 469–492.

Yen, W. M. (1987, June). *A Bayesian/IRT index of objective performance*. Paper presented at the annual meeting of the Psychometric Society, Montreal, Québec, Canada.

Authors

LIHUA YAO is a Mathematical Statistician at Defense Manpower Data Center, 400 Gigling Road, Seaside, CA 93955; Lihua.Yao@osd.pentagon.mil. Her primary research interests include psychometric methods, mathematics and statistical methods, and software development.

KEITH BOUGHTON is a Senior Research Scientist at CTB/McGraw-Hill, #4-4109 Garry Street, Richmond, B.C., Canada, V7E 2T9; Keith_Boughton@ctb.com. His areas of specialization include MIRT, MIRT Linking, DIF, MCMC, Mixture/Latent Class Models.