Multidimensional Item Response Theory for Score Reporting

Lihua Yao
Defense Manpower Data Center
DoD Center Monterey Bay
Lihua.Yao.civ@mail.mil

October 22, 2012

Introduction

Traditionally, a test is given at the end of a certain period to evaluate the performance of the students. These tests are called summative assessments, and they are administered periodically to determine what students know and do not know at a particular point in time; they measure each student's achievement level as well as provide educators with feedback on the effectiveness of their teaching. For summative assessments, unidimensional item response theory (IRT) models have been used in practice by testing companies to report overall scores indicating students' levels of achievement in broadly defined content area domains such as mathematics and language arts. Oftentimes, subscale scores for different objectives defined by the test design are reported to further diagnose a student's strengths and weaknesses. Simple number of correct score has been reported for that purpose. With the No Child Left Behind Act of 2001, state department are looking for ways of assessing the quality of school teaching, to uncover individual weaknesses and to improve the quality of teaching. Consequently, an increasing number of states are shifting away from using only summative assessments, to using both formative assessments and diagnostic assessments. For each time period, a test is given to students, the results of which can be used both to assess students' achievement and to examine students' strengths and weaknesses; furthermore, results are used as a guide to evaluate and improve teaching

quality. Such tests can be administered multiple times in a school year; indeed, it is desirable to be able to see students' growth over a school year and across school years. The ultimate goal of any assessment is to report valid and reliable scores using fewer items. A model that satisfies the following needs is desirable: (1) can report reliable subscores for each objective or domain; (2) can report overall scores; (3) subscores and overall scores can be compared across time periods, years and forms. Such a model can serve both diagnostic and summative purposes.

Researchers have been studying methods to ensure that subscores are reliable and accurate (e.g., Ackerman & Davey, 1991; Kahraman & Kamata, 2004; Mislevy, 1987; Mislevy & Sheehan, 1989; Wainer, Sheehan, & Wang, 2000). Yen (1987) proposed an empirical Bayes procedure to reduce the error in subscale estimation by incorporating information from the total test score. The resulting objective performance index (OPI) has been widely used at CTB/McGraw-Hill. Wainer et al. (2001) proposed an augmentation to Yen's method by allowing the information from other subscale scores (they note that Yen only used the total test score) to help stabilize each diagnostic subscale score, with each item loading on only one subscale. Wainer et al. reported that this augmentation is well suited for tests that may be more multidimensional in nature. Sinharay, Haberman, and Puhan (2007) suggested methods based on classical test theory to examine whether subscores provided any added value over total scores at an individual level and at the level of institutions that the examinees belong to.

Multidimensional item response theory (MIRT) has the nature of measuring multiple skills with complex interactions of persons and test items (Reckase, 2009). For the last two decades, substantial research on MIRT models has been conducted, addressing such issues as the exploration and detection of test dimensional structure, concurrent versus separate calibration in vertical scaling (due to shifts in dimensional structure across grades), and the detection of benign DIF items that are construct-relevant, to name a few. The dimensional structure of a set of data depends on both the items and the examinees taking the test. The detection and interpretation of the dimensional structure is a complex procedure. Kim (2011) and Zhang et al. (2011) are the two most recent complex studies that used real data sets and analyzed the dimensional structure of the data, examining the model fit, performing parallel analysis and cluster analysis, and BMIRT exploratory and confirmatory analysis, to report meaningful subscores and overall scores and to link the subscores across years. Using confirmatory multidimensional

analysis to compute subscale scores and overall scores increases measurement precision and reduces test length both in traditional paper and pencil format and computer adaptive format (Yao, 2012). MIRT ability estimates have been shown to recover the true values well; they outperformed classical number of correct score and OPI (Yao & Boughton, 2007) and were on par with the argumentation methods (de la Torre & Patz, 2005; Dwyer et al., 2006). MIRT for domain or subscale scores and overall scores and their linking, have been shown to be promising (de la Torre & Hong, 2010; de la Torre & Song, 2009; Haberman & Sinharay, 2010; Puhan & Liang, 2011; Segall, 2001; Sinharay, 2010; Sinharay, Puhan, & Haberman, 2007, 2011; Sinharay & Haberman, 2011; Wang, Cheng, & Chen, 2004; Yao, 2010b, 2011; Yao & Boughton 2009). Moreover, such reported scores can be made comparable by MIRT linking across forms, samples and years. MIRT for score reporting is very promising; for this purpose, the rest of the book chapter will discuss the following: (2) Multidimensional IRT model; (3) Item parameter estimates in the MIRT framework; (4) Methods for score estimates and their standard error of measurement; (5) Linking scores in the MIRT framework; (6) Discussion.

Multidimensional IRT Models

Following the notation of the compensatory MIRT model in Yao & Schwartz (2006), for a dichotomously-scored item j, the probability of a correct response to item j for an examinee with ability $\vec{\theta}_i = (\theta_{i1}, \cdots, \theta_{iD})$ for the multidimensional three-parameter logistic (M-3PL; Reckase, 1997) model is:

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

where $x_{ij} = 0$ or 1 is the response of examinee i to item j. $\vec{\beta}_{2j} = (\beta_{2j1}, \cdots, \beta_{2jD})$ is a vector of dimension D for item discrimination parameters. β_{1j} is the intercept and $\frac{\beta_{1j}}{||\vec{\beta}_{2j}||}$ is the difficulty parameter, β_{3j} is the lower asymptote or the guessing parameter, and $\vec{\beta}_{2j} \odot \vec{\theta}_i^T = \sum_{l=1}^{D} \beta_{2jl}\theta_{il}$. The norm or multidimensional discrimination index MDISC (Reckase & McKinely, 1991) is defined as $||\vec{\beta}_{2j}|| = \sqrt{\sum_{l=1}^{D} \beta_{2jl}^2}$. The parameters for the jth item are $\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{1j}, \beta_{3j})$. For a polytomously-scored item j, the probability of a

response $k-1$ to item j for an examinee with ability $\vec{\theta}_i$ is given by the multidimensional version of the generalized two-parameter partial credit model (M-2PPC; Yao & Schwartz 2006)

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

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

For a polytomous scored item j with response level/category K_j, the multidimensional graded response model (M-GRM; Muraki & Carlson, 1993) is defined below:

First define cumulative response function for $k = 0, \cdots, K_j$ as:

- $P^{\star}_{ijk} = 1$ for $k = 0$.
- $P^{\star}_{ijk} = 0$ for $k = K_j$.
- $P^{\star}_{ijk} = 1 - \frac{1}{1+e^{\vec{\beta}_{2j} \odot \vec{\theta}^T + \beta_{\delta_k j}}} = \frac{1}{1+e^{-\vec{\beta}_{2j} \odot \vec{\theta}^T - \beta_{\delta_k j}}}$ for $k = 1, \cdots, K_j - 1$.

The probability of having response $k-1$ with $k \in \{1, \cdots, K_j\}$ for item j is

$$P_{ijk} = P(x_{ij} = k-1 \mid \vec{\theta}_i, \vec{\beta}_j) = P^{\star}_{ijk-1} - P^{\star}_{ijk}. \tag{3}$$

The parameters for jth item are

$$\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{\delta_1 j}, \ldots, \beta_{\delta_{K_j} j}). \tag{4}$$

Note that $\beta_{\delta_1 j} \leq \beta_{\delta_2 j} \leq \cdots \leq \beta_{\delta_{K_j - 1} j}$.

Let

$$P_{ij} = P_{ij}(X_{ij} \mid \vec{\theta}_i, \vec{\beta}_j) = P_{ij1}^{1(X_{ij}=1)}(1 - P_{ij1})^{1(X_{ij}=0)} \tag{5}$$

for an M-3PL item and

$$P_{ij} = P_{ij}(X_{ij} \mid \vec{\theta}_i, \vec{\beta}_j) = \prod_{k=1}^{K_j} P_{ijk}^{1(X_{ij}=k-1)}, \tag{6}$$

for an M-2PPC item or an M-GRM item, where

$$1_{(X_{ij}=k)} = \begin{Bmatrix} 1 & \text{if } X_{ij} = k \\ 0 & \text{otherwise} \end{Bmatrix}$$

The likelihood equation for responses to J items for N examinees $\mathbf{X} = (\vec{X}_i^T, \cdots, \vec{X}_N^T)$, $\vec{X}_i = (X_{i1}, \cdots, X_{iJ})$, given ability $\tilde{\boldsymbol{\theta}} = (\vec{\theta}_1^T, \cdots, \vec{\theta}_N^T)$ is

$$L(\vec{X} \mid \tilde{\boldsymbol{\theta}}) = \prod_{i=1}^{N} P(\vec{X}_i \mid \vec{\theta}_i, \boldsymbol{\beta}) = \prod_{i=1}^{N} \prod_{j=1}^{J} P_j(X_{ij} \mid \vec{\theta}_i, \vec{\beta}_j). \tag{7}$$

Item Parameter Estimates in the MIRT Framework

With the MIRT model defined, the next important factor for the wide use of the model is the item parameter estimate. The item parameter estimate can be obtained by maximizing the likelihood function or the posterior likelihood function. Researches have been using software that originated from factor analysis (TESTFACT, Wilson, Wood, & Gibbons, 1987; NOHARM, Fraser, 1987) for multiple choice items, until recent 15 years. With the arise of Markov chain Monte Carlo (MCMC) technique, WinBugs (Spiegelhalter, Thomas, Best, & Gilks, 1995), a popular software program used by many researchers in different areas, will output the MCMC sampling of the specified parameters. The drawbacks of WinBugs are that it is too general and may not serve user needs efficiently; that in addition to knowing some programming techniques, users need to have enough knowledge of statistics and enough knowledge of the models and their parameters; and that the run time can be extremely long. For more practical users, there are other ready-to-use software programs that use MCMC methods for multidimensional and unidimensional item response theory models (BMIRT, Yao, 2003a, 2010a, free for download at www.BMIRT.com) and structural equation modeling (Mplus, Muth´en, & Muth´en, 2004); both software has been used and tested by many studies. IRTPRO (Cai, du Toit, & Thissen, 2009 ,2012) was recently developed using a few different techniques to estimate item parameters. Both BMIRT and IRTPRO have the capability of estimating the three-parameter logistic model, the generalized two-parameter partial credit model, and the graded response model in the MIRT framework and in the mutlti-group framework; it is worth mentioning that the multidimensional multigroup feature is very promising in detecting construct-relevant benign

DIF (Yao & Li, 2010; Liu et al, 2012). Moreover, the exploratory mode of the software allows the users to examine the dimensional structure of the data. Lin P (2008) studied and compared the performance of BMIRT, M-plus, NOHARM, and BIOLOG; more studies comparing these softwares are needed.

MIRT Ability Estimation Methods and Standard Error of Measurement (SEM)

Ability estimates in the MIRT framework can be produced by Bayesian or non-Bayesian. Similar to the unidimensional IRT (UIRT), there are three methods that can be used to estimate abilities in the MIRT framework. They are: (a) MLE: Maximum likelihood estimation methods; (b) MAP: Maximize a posterior; (c) EAP, Expected a posterior. With the development of MCMC technique, the abilities can be derived by MCMC sampling for the posterior distribution and the mean of the ability samplings after the burn-in are their estimates; this is similar to EAP. For MAP and EAP, strong priors, standard normal, and noninformative priors can be applied. The following statistics will be used in computing the estimates and their standard error of measurement.

Statistics

First-Derivative.

$$\frac{\partial \log L(\vec{X} \mid \vec{\theta})}{\partial \vec{\theta}} = \sum_{j=1}^{J} \frac{\partial \log P_j}{\partial \vec{\theta}}, \tag{8}$$

where

$$\frac{\partial \log P_j}{\partial \vec{\theta}} = \sum_{k=1}^{K_j} 1_{(X_j=k-1)} \frac{\partial \log P_{jk}}{\partial \vec{\theta}} = \sum_{k=1}^{K_j} 1_{(X_j=k-1)} \frac{\partial P_{jk}}{\partial \vec{\theta}} \frac{1}{P_{jk}} = \sum_{k=1}^{K_j} 1_{(X_j=k-1)} ((k-1)-E_j)\vec{\beta}_{2j}, \tag{9}$$

and

$$E_j = \sum_{k=1}^{K_j} (k-1) P_{jk} = \frac{\sum_{k=1}^{K_j} (k-1) e^{(k-1)\vec{\beta}_{2j} \odot \vec{\theta}^T - \sum_{t=1}^{k} \beta_{\delta_t j}}}{\sum_{m=1}^{K_j} e^{(m-1)\vec{\beta}_{2j} \odot \vec{\theta}^T - \sum_{t=1}^{m} \beta_{\delta_t j}}}, \tag{10}$$

for an M-2PPC item. For an M-3PL item,

$$\frac{\partial \log P_j}{\partial \vec{\theta}} = \left(\frac{1_{(X_j=1)}}{P_{j1}} - \frac{1_{(X_j=0)}}{1-P_{j1}}\right) \frac{\partial P_{j1}}{\partial \vec{\theta}} = \frac{(X_j - P_{j1})(P_{j1} - \beta_{3j})}{P_{j1}(1-\beta_{3j})} \vec{\beta}_{2j}. \tag{11}$$

Second-Derivative.

$$J(\vec{\theta}) = \frac{\partial^2 \log L(\vec{X} \mid \vec{\theta})}{\partial \vec{\theta}^2} = \sum_{j=1}^{J} \frac{\partial^2 \log P_j}{\partial \vec{\theta}^2}, \tag{12}$$

and

$$\frac{\partial^2 \log P_j}{\partial \vec{\theta}^2} = -\sum_{k=1}^{K_j} 1_{(X_j=k-1)} \frac{\partial E_j}{\partial \vec{\theta}} \otimes \vec{\beta}_{2j} = -\sigma_j^2 \vec{\beta}_{2j} \otimes \vec{\beta}_{2j}, \tag{13}$$

where

$$\sigma_j^2 = \sum_{k=1}^{K_j} (k-1)^2 P_{jk} - E_j^2, \tag{14}$$

for an M-2PPC item. Here $\otimes$ is a vector product; $\vec{\beta}_{2j} \otimes \vec{\beta}_{2j}$ is a $D \times D$ matrix, and its mth row and nth column element is the product of the mth and nth element of $\vec{\beta}_{2j}$. For an M-3PL item j,

$$\frac{\partial^2 \log P_j}{\partial \vec{\theta}^2} = \frac{\beta_{3j} X_j - P_{j1}^2}{P_{j1}^2} \frac{1}{1-\beta_{3j}} \frac{\partial P_{j1}}{\partial \vec{\theta}} \otimes \vec{\beta}_{2j} = \frac{(1-P_{j1})(P_{j1}-\beta_{3j})(\beta_{3j} X_j - P_{j1}^2)}{P_{j1}^2 (1-\beta_{3j})^2} \vec{\beta}_{2j} \otimes \vec{\beta}_{2j}. \tag{15}$$

Item and Test Information Function. For an item j, following M-3PL, the information function at $\vec{\theta}$ is

$$I_j(\vec{\theta}) = -E \frac{\partial^2 \log P_j}{\partial \vec{\theta}^2} = \frac{(P_{j1}-\beta_{3j})^2 (1-P_{j1})}{P_{j1}(1-\beta_{3j})^2} \vec{\beta}_{2j} \otimes \vec{\beta}_{2j}. \tag{16}$$

For an item j, following M-2PPC, the information function at $\vec{\theta}$ is

$$I_j(\vec{\theta}) = \sigma^2 \vec{\beta}_{2j} \otimes \vec{\beta}_{2j}, \tag{17}$$

where $\sigma^2 = \sum_{k=1}^{K_j} (k-1)^2 P_{jk} - (\sum_{k=1}^{K_j} (k-1) P_{jk})^2$. The test information for J items at $\vec{\theta}$ is $\mathbf{I}_J(\vec{\theta}) = \sum_{j=1}^{J} I_j(\vec{\theta})$. The directional information in the direction $\vec{\alpha}$ is $\cos(\vec{\alpha}) \mathbf{I}_J \cos(\vec{\alpha})^T$, where $\cos(\vec{\alpha}) = (\cos\alpha_1, \cdots, \cos\alpha_D)$ and α_l is the angle between $\vec{\theta}$ and θ_l.

Bayesian Statistics

Suppose the prior of population is $f(\vec{\theta})$, then the posterior density function of $\vec{\theta}$ is

$$f(\vec{\theta} \mid \vec{X}) \propto L(\vec{X} \mid \vec{\theta}) f(\vec{\theta}), \tag{18}$$

and if the prior is normal $N(\vec{\mu}, \Sigma)$, then

$$f(\vec{\theta}) = (2\pi)^{-D/2}(\mid \Sigma \mid)^{-1/2} exp(-\frac{1}{2}(\vec{\theta} - \vec{\mu})^T \Sigma^{-1}(\vec{\theta} - \vec{\mu})), \tag{19}$$

where $\vec{\mu}$ and Σ represent the population mean and variance-covariance matrix, respectively.

First-Derivative.

$$\frac{\partial \log f(\vec{\theta} \mid \vec{X})}{\partial \vec{\theta}} = \frac{\partial \log L(\vec{X} \mid \vec{\theta})}{\partial \vec{\theta}} - \frac{\partial(\vec{\theta} - \vec{\mu})}{\partial \vec{\theta}} \Sigma^{-1}(\vec{\theta} - \vec{\mu}), \tag{20}$$

where

$$\frac{\partial(\vec{\theta} - \vec{\mu})}{\partial \theta_k} = (0, \cdots, 1, 0, \cdots, 0)_{1 \times D},$$

and 1 is in the kth position.

Second-Derivative.

$$\frac{\partial^2 \log f(\vec{\theta} \mid \vec{X})}{\partial \vec{\theta}^2} = \frac{\partial^2 \log L(\vec{X} \mid \vec{\theta})}{\partial \vec{\theta}^2} - \Sigma^{-1} = J(\vec{\theta}) - \Sigma^{-1}. \tag{21}$$

Item and test information function. Posterior test information at $\vec{\theta}$ for selected $j-1$ items is

$$\mathbf{I}_{j-1}(\vec{\theta}) = -EJ(\vec{\theta}) + \Sigma^{-1} = -\sum_{j=1}^{J} E\frac{\partial^2 \log P_j}{\partial \vec{\theta}^2} + \Sigma^{-1}. \tag{22}$$

Here bold variable $\mathbf{I}_{j-1}$ indicates the sum of $I_1, \cdots, I_{j-1}$.

The ability estimation methods are described below.

- MLE: MIRT ability is estimated by finding the mode $\hat{\vec{\theta}}$ that maximize the likelihood function $L(\vec{X} \mid \vec{\theta})$, i.e.,

$$\frac{\partial \log L(\vec{X} \mid \vec{\theta})}{\partial \vec{\theta}} \mid \hat{\vec{\theta}} = 0. \tag{23}$$

Using Newton-Raphson method, suppose $\vec{\theta}^m$ is the m-th approximation that maximize $\log L(\vec{X} \mid \vec{\theta})$, then

$$\vec{\theta}^{m+1} = \vec{\theta}^m - \vec{\delta}^m, \tag{24}$$

where

$$\vec{\delta}^m = [\mathbf{J}(\vec{\theta}^m)]^{-1} \times \frac{\partial \log L(\vec{X} \mid \vec{\theta})}{\partial \vec{\theta}}, \tag{25}$$

and $\mathbf{J}(\vec{\theta})$ is the matrix of the second partial derivative.

- MAP: MIRT ability is estimated by finding the mode that maximize the posterior likelihood function $f(\vec{\theta} \mid \vec{X})$. This is similar to MLE method, but the function used is the product of the likelihood and the prior–instead of the likelihood.

  MAP estimates are derived using Equation (23) and (24) for the Bayesian versions.

- EAP: Suppose there are Q quarture points in the θ range $(-3, 3)$ that forms the D-dimensional vector score point $\vec{\theta}_k$, where $k = 1, \cdots, Q^D$. The marginal likelihood function is

$$\int L(\vec{X} \mid \vec{\theta}) f(\vec{\theta}) d\theta_1 \cdots \theta_D = \sum_{k=1}^{Q^D} L(\vec{X} \mid \vec{\theta}_k) f(\vec{\theta}_k). \tag{26}$$

  The expectation is a vector of dimension D, and it is

$$E(\vec{X}) = \int L(\vec{X} \mid \vec{\theta}) f(\vec{\theta}) \vec{\theta} d\theta_1 \cdots \theta_D = \sum_{k=1}^{Q^D} L(\vec{X} \mid \vec{\theta}_k) f(\vec{\theta}_k) \vec{\theta}_k, \tag{27}$$

  where the lth componet of the left hand side is corresponding to the lth component of $\vec{\theta}_k$. The EAP estimates for $\vec{\theta}$ is

$$\hat{\vec{\theta}} = \frac{E(\vec{X})}{\int L(\vec{X} \mid \vec{\theta}) f(\vec{\theta}) d\theta_1 \cdots \theta_D}. \tag{28}$$

For MLE, it is to find $(\hat{\theta_1}, \cdots, \hat{\theta_D})$ such that $\prod_{j=1}^{J} P_j(X_j \mid \vec{\theta}, \vec{\beta}_j) = \prod_{l=1}^{D} \prod_{j \in O_l} P_j(X_j \mid \vec{\theta}, \vec{\beta}_j)$ has the maximum value, which is equivalent to

find $\hat{\theta}_l$ that maximizing the likelihood for domain l for all $l = 1, \cdots, D$. Here O_l contains all the items in domain l. For Bayesian methods with standard normal as the prior, MAP or EAP for deriving $(\hat{\theta_1}, \cdots, \hat{\theta_D})$ is the same as the MAP or EAP for finding $\hat{\theta}_l$ with $N(0, 1)$ as the prior when the items in the test are simple structured, where $l = 1, \cdots, D$, as the joint prior for $\vec{\theta}$ is the product of the priors for each of the domains. When the noninformative prior is used, for example, with variance $var = 10$, and covariance $cov = 0$, the Bayesian methods should yield similar results as those from MLE. Using standard normal or noninformative prior would ignore the correlated information between domains, while using strong priors would allow the information to be borrowed from each other and increase the precision, especially when the test is short. Therefore, compared to the ability estimates from UIRT, the ability estimates from MIRT would have better precision using MAP or EAP when the prior is strong and is neither standard normal nor noninformative (Yao & Boughton, 2007). Such a prior can be derived using the student raw data.

Composite Score and SEM

For a test with J items of known item parameters, for a given score point $\vec{\theta}$, the test information is $\mathbf{I}_J(\vec{\theta})$. the composite score $\theta_{\vec{\alpha}} = \sum_{l=1}^{D} \theta_l w_l$ has a standard error of measurement $SEM(\theta_{\vec{\alpha}}) = V(\theta_{\vec{\alpha}})^{1/2}$, where $V(\theta_{\vec{\alpha}}) = \vec{w}V(\vec{\theta})\vec{w}^T$, $\vec{w} = (w_1, \cdots, w_D) = (cos^2\alpha_1, \cdots, cos^2\alpha_D)$. $V(\vec{\theta})$ can be approximated by $I(\vec{\theta})^{-1}$. The SEM for each domain can be derived by using the angle for that dimension to be 0, and all other angles 90^0.

The composite score for the D-dimensinal domain scores can be obtained by simply averaging the domain scores with a set of predetermined weights, by the weighted sum of the domain scores and the optimized weight (Yao, 2011, 2012), or by higher-order MIRT (de la Torre & Hong, 2010; de la Torre & Song, 2009; Yao, 2010b) model using MCMC that simultaneously estimate the domain abilities and the composite abilities. For the optimized weight, it yields the composite score with the smallest SEM, and therefore has a better prediction or estimation for an examinee's overall ability. The optimized weight can be derived by formula (Yao, 2012) or by computer program (BMIRT, 2010).

Comparing the Performance of MLE, MAP, and EAP

To compare the performance of the three methods MLE, MAP, and EAP, a simulation study was conducted through running BMIRT.

Paper and pencil ASVAB Armed Forces Qualification Test (AFQT) data for four contents were collected, with each student taking four test AR (30 items), WK (35 items), PC (15 items), MK (25 items). In total, there were 105 items with about 6000 responses for each item. To derive the true item parameters for the simulation study, the following analysis was conducted: BMIRT four-dimensional calibration with simple structure, with each item load on each dimension presenting each of the four contents; the term dimension, domain, and content are used interchangeably. Scale was fixed by specifying that the prior of the ability distribution be multivariate with mean $\vec{B}_1 = (0, 0, 0, 0)$ and variance-covariance matrix of $\mathbf{A}_1$, an identity matrix. A sample size of 1000 was simulated from $N(\vec{\mu}, \Sigma)$, where the mean was $\vec{\mu} = (-1, 0.0, 1, -0.7)$, and the variance-covariance matrix was

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.7 \\ 0.5 & 1 & 0.6 & 0.4 \\ 0.5 & 0.6 & 1 & 0.4 \\ 0.7 & 0.4 & .4 & 1 \end{pmatrix}_{4\times 4}.$$

Responses generated from the MIRT model using the true item parameters and the true abilities were produced with 30 replications using 30 random seeds. For MAP and EAP estimates, a strong prior that was the generating distribution was applied; in practice the prior information can be obtained from previous knowledge about the population or from the population raw response data. Reliability, RMSE, and BIAS for the abilities were computed across the 30 replications. They were defined as follows: let f_{true} be the value of a function obtained from true parameter and f_l be the value of a function obtained from the estimated parameter from sample l. Here the function f can represent ability parameters. RMSE was calculated by $RMSE = \sqrt{\frac{1}{n}\sum_{l=1}^{n}(f_l - f_{true})^2}$, where n is the number of replications. BIAS was defined by $BIAS = \mid f_{true} - \overline{f} \mid$, where $\overline{f} = \frac{1}{n}\sum_{l=1}^{n} f_l$ was the final estimate. The reliability for the subscore is computed below: For each of the $N = 1000$ simulated examinee with true ability value $\vec{\theta}_i$, compute the mean and variance of the estimated subscores for the N simulees. Let $\mathbf{x}_l(\vec{\theta}_i)$, $l = 1, \cdots, n = 30$ indicates the estimated score for the lth replication; it is a vector of dimension D. The following computations are for each coordinate

of the vector. Then the mean and variance of the scores are

$$\mathbf{m}(\vec{\theta}_i) = m(x \mid \vec{\theta}_i) = \frac{\sum_{l=1}^{n} \mathbf{x}_l(\vec{\theta}_i)}{n}, \tag{29}$$

$$\mathbf{Var}(\vec{\theta}_i) = var(x \mid \vec{\theta}_i) = \frac{\sum_{l=1}^{n} [\mathbf{x}_l(\vec{\theta}_i) - \mathbf{m}(x \mid \vec{\theta}_i)]^2}{n-1}. \tag{30}$$

Finally, the overall mean and variance can be obtained by

$$\mathbf{E} = \sum_{i=1}^{N} \mathbf{m}(\vec{\theta}_i), \tag{31}$$

$$\mathbf{Var} = \sum_{i=1}^{N} [\mathbf{m}(\vec{\theta}_i) - \mathbf{E}]^2, \tag{32}$$

Let mean variance or the error variance be computed by

$$\mathbf{E}(Var) = \sum_{i=1}^{N} \mathbf{Var}(\vec{\theta}_i). \tag{33}$$

In general, test reliability is defined by

$$r = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \tag{34}$$

and in this case it is computed for each dimension by

$$\mathbf{r} = \frac{\mathbf{Var}}{\mathbf{Var} + \mathbf{E}(Var)}. \tag{35}$$

The reliability for the four contents were computed and they are presented in Table 1, along with RMSE and BIAS. Since the true item and ability parameters were known, the true SEMs were Derived; these were used to examine the BIAS. In practice, the estimated SEM can be derived using the estimated item and ability parameters. Figure 1 shows SEM and BIAS for the three methods for the four subscores. The mean for the true item difficulties for the four domains were 0.90, 0.45, -0.10, and 0.9, respectively; this is the cause for the observable high SEM for AR and MK on the left side in Figure 1. The population had means (-1, 0, 1, -0.7), which were quite

different from the item difficulties, specially for the first-, third-, and fourth-dimensions. For the second dimension dimension (WK), the reliability are much higher than the other dimensions–the population mean was closer to the item difficulty for that dimension than those of the other dimensions. The Non-Bayesian method MLE has the smallest BIAS, largest RMSE, and the smallest reliability. The Bayesian method with strong prior has the largest reliability and the smallest RMSE, followed by Bayesian methods with standard normal as the prior. Bayesians with noninformative priors perform similarly with MLE. Overall EAP performed similarly to MAP; however, the estimation time for EAP was much longer than MAP, with 21 quarture points from $(-3, 3)$ being used. Since the total number of items in estimating the ability is 105, the performace of MLE should be good with this lengthy test. It is expected that MAP with strong prior would perform even better than MLE for a short test. To arrive at a conclusion on the relative performances of the three methods, a much better design with varying test length, number of quarture points, time used, and population distribution and prior and size is needed; this study only sheds light on the topic.

Table 1

Reliability, RMSE, and BIAS for the MIRT subscale estimates.

Method	AR	WK	PC	MK
		Reliability		
MLE	.65	.85	.63	.66
MAP Strong Prior	.84	.88	.79	.81
MAP Standard Prior	.78	.86	.71	.77
MAP Noninformative Prior	.70	.85	.66	.68
EAP Strong Prior	.85	.88	.79	.83
EAP Standard Prior	.79	.86	.71	.79
EAP Noninformative Prior	.78	.86	.71	.77
		RMSE		
MLE	.73	.42	.88	.70
MAP Strong Prior	.46	.35	.49	.47
MAP Standard Prior	.57	.36	.60	.54
MAP Noninformative Prior	.60	.41	.70	.64
EAP Strong Prior	.45	.35	.49	.46
EAP Standard Prior	.52	.36	.59	.50
EAP Noninformative Prior	.53	.40	.61	.56
		BIAS		
MLE	.176	.071	.282	.150
MAP Strong Prior	.256	.128	.275	.255
MAP Standard Prior	.387	.118	.391	.340
MAP Noninformative Prior	.129	.051	.151	.113
EAP Strong Prior	.267	.117	.265	.258
EAP Standard Prior	.322	.112	.368	.289
EAP Noninformative Prior	.243	.057	.171	.244

Linking Scores in the MIRT Framework

For security purposes, many test forms are developed. Scores from different test forms should be made comparable by putting them onto the same coordinate system. This is called horizontal linking. Because policy makers and schools are interested in knowing how much the students have changed, the scores from different grade levels should be put onto the same coordinate system. This is called vertical scaling or linking. Setting up the base coordinate system for MIRT score reporting is much more complex than for the UIRT. It involves the number of dimensions, dimensional structure, the location of the system, the orientation of coordinate axes, and the unit of measurement for each coordinate axes; careful research is needed for each project (Kim, 2011).

For a given set of base coordinate system, there are a few methods for linking new coordinate systems to the base coordinate system.

Like unidimensional IRT models, the scale for examinees' ability (or item parameters) in the MIRT models has indeterminacy; that is, the parameters are determined up to a linear transformation. The transformation matrix $\mathbf{A}_{D\times D}$ and location vector $\vec{B}_{1\times D}$ can be determined by the following: For an M-3PL item j, let

$$\vec{\beta}^*_{2j} = \vec{\beta}_{2j}\mathbf{A}^{-1}, \tag{36}$$

$$\beta^*_{1j} = \beta_{1j} + \vec{\beta}_{2j}\mathbf{A}^{-1}\vec{B}^T, \tag{37}$$

$$\beta^*_{3j} = \beta_{3j}. \tag{38}$$

For an M-2PPC item, let

$$\beta^*_{\delta_{kj}} = \beta_{\delta_{kj}} + \beta_{2j}\mathbf{A}^{-1}\vec{B}^T, \tag{39}$$

for $k = 1, \cdots, K_j$. Let $\vec{\theta}^*_i = \vec{\theta}_i\mathbf{A}^T + \vec{B},$, then the probability of obtaining a certain score on the jth item is not altered, that is $P_{ijk}(\vec{\theta}^*_i, \vec{\beta}^*_j) = P_{ijk}(\vec{\theta}_i, \vec{\beta}_j)$. Reckase (2009) described how to produce the transformations from item and person parameters. There are also available linking software (IPLINK, Lee & Oshima, 1996; LinkMIRT, Yao, 2004) that links two sets of item parameters onto the same scale and there are some studies on MIRT linking (Davey, Oshima, & Lee, 1996; Hirsch, 1989; Li & Lissitz, 2000; Oshima, Davey, & Lee, 2000). Basically, there are three types of linking design: common-item design, common-person design, and random equivalent-groups design. For the common-item design, common items or anchor items are shared between the two tests. For the common-person design, a sample of examinees take both tests. For the random equivalent-groups design, the two populations taking the two tests are similar or have the same population distribution. The three designs and a study comparing the last two designs are discussed below.

Common-Item Design

For the nonequivalent common-item design (NEAT), the transformation matrix and the location vector can be derived mathematically from the two sets of item parameters (Chapter 8, Reckase, 2009); it involves matrix computations. Other methods for linking for subscores have been studied based on MIRT model using matching test response function, mean/mean,

mean/sigma (Yao & Boughton, 2009; Yao, 2011)–they are extensions for the linking methods for the UIRT. There is also a linking study using linear equating methods based on classical analysis (Puhan & Liang, 2011). The following will discuss the three linking methods using the MIRT model.

Matching TRF. The Stocking-Lord (Stocking & Lord, 1983) method was generalized to the multidimensional case. Let J_1 be the number of multiple choice items and J_2 be the number of constructed response items. The test response function is then defined by the following

$$TRF(\vec{\theta},\vec{\beta}) = \frac{1}{J_1 + \sum_{j=1}^{J_2}(K_j - 1)} \sum_{i=1}^{N}[\sum_{j=1}^{J_1} P_{ij1} + \sum_{j=1}^{J_2}\sum_{k=1}^{K_j}(k-1)P_{ijk}]. \quad (40)$$

Let the lowest and highest point for the ability scale be m_1, m_2. If Q quadrature points are chosen between (m_1, m_2), then there are Q^D possible choices of $\vec{\theta}$, where D is the dimension. The transformation matrix $\mathbf{A}_{D\times D}$ and location vector $\vec{B}_{1\times D}$ can be determined by finding the minimum difference between the TRFs: $min\{\frac{1}{Q^D}\sum_{i=1}^{Q^D}[TRF(\vec{\theta}_i,\vec{\beta}) - TRF(\vec{\theta}_i,\vec{\beta}^*)]^2\}$.

Mean/Sigma. Let $\mu = (\mu_1,\cdots,\mu_D), \sigma = (\sigma_1,\cdots,\sigma_D), \mu^* = (\mu_1^*,\cdots,\mu_D^*), \sigma^* = (\sigma_1^*,\cdots,\sigma_D^*)$ be the mean and standard deviation of the difficulty parameter $\beta_{1j}/||\vec{\beta}_{2j}||$ in two different matrixes. For $i \in \{1,\cdots,D\}$, the ith element in each of the vectors is obtained from the items contributing to that dimension or domain. Let $a_i = \frac{\sigma_i^*}{\sigma_i}$, and $b_i = \mu_i^* - a_i\mu_i$. For a test of simple structure, for example, where an item contributes to only one domain, the transformation matrix $\mathbf{A}_{D\times D}$ and location vector $\vec{B}_{1\times D}$ can be obtained below:

$$\mathbf{A} = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ \cdots & & & \\ 0 & \cdots & 0 & a_D \end{pmatrix}_{D\times D},$$

and $\vec{B}_{1\times D} = (b_1,\cdots b_D)$.

Mean/Mean. Similarly to the Mean/Sigma method, let $\mu_a = (\mu_{a1},\cdots,\mu_{aD})$, $\mu_b = (\mu_{b1},\cdots,\mu_{bD})$, $\mu_a^* = (\mu_{a1}^*,\cdots,\mu_{aD}^*)$, $\mu_b^* = (\mu_{b1}^*,\cdots,\mu_{bD}^*)$ be the means of the discrimination parameter and the difficulty parameter $\beta_{1j}/||\vec{\beta}_{2j}||$ in two different matrixes. Then, for $i = 1,\cdots,D$, $a_i = \frac{\mu_{ai}^*}{\mu_{ai}}$, and $b_i = \mu_{bi}^* - a_i\mu_{bi}$.

Common-Person Design

In the literature, there are no studies in the MIRT framework for linking by the common-person design and linking by random equivalent-groups design or population distributions.

For the common-person design, suppose there are N examinees that take both tests. Each examine has two $1 \times D$ vector ability estimates $\vec{\theta}_i = (\theta_{i1}, \cdots, \theta_{iD})$ and $\vec{\theta}_i^* = (\theta_{i1}^*, \cdots, \theta_{iD}^*)$ for the base coordinate system and the new coordinate system, respectively. Let $\boldsymbol{\theta} = (\vec{\theta}_1^T, \cdots, \vec{\theta}_N^T)_{N\times D}^T$ and $\boldsymbol{\theta}^* = (\vec{\theta^*}_1^T, \cdots, \vec{\theta^*}_N^T)_{N\times D}^T$. Let the estimated population mean denoted by $\vec{\mu}$ and $\vec{\mu^*}$, respectively, for the two coordinate system. Let $\boldsymbol{\mu} = (\vec{\mu}^T, \cdots, \vec{\mu}^T)_{N\times D}^T$, $\boldsymbol{\mu}^* = (\vec{\mu^*}^T, \cdots, \vec{\mu^*}^T)_{N\times D}^T$, then

$$(\boldsymbol{\theta}^* - \boldsymbol{\mu}^*)_{N\times D} = (\boldsymbol{\theta} - \boldsymbol{\mu})_{N\times D}\mathbf{A}^T, \tag{41}$$

$$(\boldsymbol{\theta} - \boldsymbol{\mu})_{N\times D} = (\boldsymbol{\theta}^* - \boldsymbol{\mu}^*)_{N\times D}(\mathbf{A}^T)^{-1}. \tag{42}$$

The transformation matrix $\mathbf{A}$ can be derived by

$$\mathbf{A}^T = [(\boldsymbol{\theta} - \boldsymbol{\mu})^T(\boldsymbol{\theta} - \boldsymbol{\mu})]^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})^T(\boldsymbol{\theta}^* - \boldsymbol{\mu}^*), \tag{43}$$

or by

$$(\mathbf{A}^T)^{-1} = [(\boldsymbol{\theta}^* - \boldsymbol{\mu}^*)^T(\boldsymbol{\theta}^* - \boldsymbol{\mu}^*)]^{-1}(\boldsymbol{\theta}^* - \boldsymbol{\mu}^*)^T(\boldsymbol{\theta} - \boldsymbol{\mu}). \tag{44}$$

The location vector $\vec{B}$ is then derived by

$$\vec{B} = \vec{\mu}^* - \vec{\mu}\mathbf{A}^T. \tag{45}$$

Random Equivalent-Groups Design

For the random equivalent-groups design, the population distributions for the two coordinate system are the same.

Suppose the estimates for the variance-covariance matrix for the population distribution for the two coordinate systems are $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^*$, respectively, then

$$\boldsymbol{\Sigma}^* = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T, \tag{46}$$

and

$$\vec{\mu}^* = \vec{\mu}\mathbf{A}^T + \vec{B}. \tag{47}$$

Suppose the decomposition for the two matrixes are $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ and $\boldsymbol{\Sigma}^* = \mathbf{C}\mathbf{C}^T$, then

$$\mathbf{A} = \mathbf{C}\mathbf{L}^{-1}. \tag{48}$$

The location vector $\vec{B}$ is derived using equation 38.

For the estimates $\vec{\theta}^*$, going back to the base coordinate system, it will become

$$\vec{\theta} = \vec{\theta}^*(\mathbf{A}^T)^{-1} + (\vec{\mu} - \vec{\mu}^*(\mathbf{A}^T)^{-1}). \tag{49}$$

For Σ^* and $\vec{\mu}^*$, going back to the base coordinate system, they will become $\Sigma = A^{-1}\Sigma^*(A^{-1})^T$, and $\vec{\mu} = \vec{\mu}^*(A^T)^{-1} - \vec{B}(A^T)^{-1}$. For $\vec{\beta}^*$, going back to the base coordinate system, they will become

$$\vec{\beta}_{2j} = \vec{\beta}^*_{2j}\mathbf{A}, \tag{50}$$

$$\beta_{1j} = \beta^*_{1j} - \vec{\beta}^*_{2j}\vec{B}^T = \beta^*_{1j} - \vec{\beta}^*_{2j}\mathbf{A}\mathbf{A}^{-1}\vec{B}^T, \tag{51}$$

$$\beta^*_{3j} = \beta_{3j}, \tag{52}$$

for an M-3PL item. For an M-2PPC item, we have

$$\beta_{\delta_k j} = \beta^*_{\delta_k j} - \vec{\beta}^*_{2j}\vec{B}^T = \beta^*_{\delta_k j} - \vec{\beta}^*_{2j}\mathbf{A}\mathbf{A}^{-1}\vec{B}^T \tag{53}$$

$$-\vec{B}(\mathbf{A}^T)^{-1} = \vec{\mu} - \vec{\mu}^*(\mathbf{A}^T)^{-1}. \tag{54}$$

Those linking parameters $\mathbf{A}$ and $\vec{B}$ for the three designs, common-item, common-person, and random equivalent-groups, and the transformed parameters are the output of LinkMIRT (Yao, 2004), which is available for free download at www.BMIRT.com.

A Simulation Study Comparing Linking by Common-Person and by Random Equivalent-Groups

The same real data set described earlier were used to derive the simulated item parameters. BMIRT three-dimensional confirmatory analysis was conducted with each dimension presenting General, Verbal, and Math - composite scores for the four contents (AR, WK, PC, and MK). In running BMIRT, the population prior was fixed using the standard multivariate normal distribution. The dimensional structure loading on each of the three composite scores is displayed in Table 2.

Table 2

Dimensional structure for the composite score

Content	General	Verbal	Math
AR	X		X
WK	X	X	
PC	X	X	
MK	X		X

The resulting three-dimensional item parameters were used as the true value for the simulation. Four sets of examinees' true abilities were generated from four sets of population distributions described in Table 3.

Table 3

Four sets of populations for the simulation study

Population	*Variance-Covariance*	*Mean*
P1	$\mathbf{I} = \{var = 1, cov = 0\}$	$\vec{\mu}_1 = (0, 0, 0)$
P2	$\mathbf{I} = \{var = 1, cov = 0\}$	$\vec{\mu}_2 = (0, .5, -.5)$
P3	$\Sigma_1 = \{var = 1, cov = .5\}$	$\vec{\mu}_3 = (0, 0, 0)$
P4	$\Sigma_2 = \{var = 1, cov = .7\}$	$\vec{\mu}_4 = (-1, 0, 1)$

For each of the four populations, examinees' true abilities were generated with sample sizes of 100, 300, 500, 1000, and 6000, respectively. Correspondingly, responses were generated using the item parameters following the model described in Equation 1 (SimuMIRT, Yao, 2003b). Sample sizes of 100, 300, 500, and 1000 were used as the common-person in the linking. For each of the four populations, the responses for sample sizes of 100, 300, 500, and 1000 were merged with the response for sample size of 6000. For each of the generated response data, three-dimensional BMIRT confirmatory calibrations following the same design as Table 2 were conducted and the two linking procedures described below were applied through running LinkMIRT (2004).

- Linking by common-person, M1: For each response data set, BMIRT confirmatory runs were conducted by fixing the population prior to be standard multivariate normal. The sampling for the item parameters and ability parameters were output of BMIRT, and the means of the MCMC sampling were the final parameter estimates. Linking to the scale of the true values were performed by LinkMIRT and the transformation matrix $\mathbf{A}$ and location vector $\vec{B}$ were obtained from the true

abilities and the estimated abilities of size 100, 300, 500, and 1000, respectively, following Equations 36 and 38. For each data set, the final item and ability estimates for the whole population were obtained by transforming the BMIRT output using $\mathbf{A}$ and $\vec{B}$, following Equations 42 and 43-47.

- Linking by population, M2: For each response data set of size 100, 300, 500, 1000, 6100, 6300, 6500, and 7000, BMIRT confirmatory run were conducted by fixing the population prior to be standard multivariate normal. The posterior population ability means and variance-covariance were computed for each MCMC iteration and the final population means and variance-covariances were computed based on the means of MCMC posterior means and variance-covariances. The linking transformation matrix $\mathbf{A}$ and location vector $\vec{B}$ were computed based on the the estimated posterior means and variance-covariance matrix and the true values, following Equations 41 and 38.

It is observed that

- For P1, $\mathbf{A} = \mathbf{I}$, $\vec{B} = (0, 0, 0)$;
- For P2, $\mathbf{A} = \mathbf{I}$, $\vec{B} = (0, -.5, .5)$;
- For P3, $\mathbf{A} = \mathbf{L}^{-1}$, where $\mathbf{L}\mathbf{L}^T = \Sigma_1$, $\vec{B} = (0, 0, 0)$. It is derived that

$$\mathbf{L} = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.5 & 0.87 & 0.00 \\ 0.51 & 0.29 & 0.82 \end{pmatrix}_{3\times 3},$$

  and

$$\mathbf{A} = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ -0.58 & 1.15 & 0.00 \\ -0.41 & -0.41 & 1.22 \end{pmatrix}_{3\times 3},$$

- For P4, $\mathbf{A} = \mathbf{L}^{-1}$, where $\mathbf{L}\mathbf{L}^T = \Sigma_2$. It is derived that

$$\mathbf{L} = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.70 & 0.71 & 0.00 \\ 0.70 & 0.29 & 0.65 \end{pmatrix}_{3\times 3},$$

and

$$\mathbf{A} = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ -0.98 & 1.40 & 0.00 \\ -0.63 & -0.63 & 1.54 \end{pmatrix}_{3\times 3},$$

$\vec{B} = (0, 0, 0) - (-1, 0, 1), \mathbf{A} = (1, -0.98, -2.17)$;

In running LinkMIRT, the transformation matrix and the location vectors for sample size of 1000 are listed below: For linking by common-person, we have:

- For population P1, $\vec{B} = (0.0130, -0.0038, -0.0078)$, and

$$\mathbf{A} = \begin{pmatrix} 0.8927 & 0.1010 & 0.1385 \\ 0.0955 & 0.6753 & -0.1098 \\ 0.0966 & -0.0985 & 0.5294 \end{pmatrix}_{3\times 3},$$

- For population P2, $\vec{B} = (0.0234, -0.3962, 0.3090)$, and

$$\mathbf{A} = \begin{pmatrix} 0.8872 & 0.1058 & 0.1355 \\ 0.0999 & 0.6754 & -0.1158 \\ 0.0980 & -0.1062 & 0.4988 \end{pmatrix}_{3\times 3},$$

- For population P3, $\vec{B} = (0.0142, -0.0073, 0.0011)$, and

$$\mathbf{A} = \begin{pmatrix} 0.7689 & 0.1331 & 0.1722 \\ -0.1549 & 0.6613 & -0.2314 \\ -0.0886 & -0.1549 & 0.4437 \end{pmatrix}_{3\times 3},$$

- For population P4, $\vec{B} = (0.4575, 0.1920, -0.4055)$, and

$$\mathbf{A} = \begin{pmatrix} 0.64167 & 0.2014 & 0.1815 \\ -0.0943 & 0.5474 & -0.2865 \\ -0.0172 & -0.2507 & 0.3861 \end{pmatrix}_{3\times 3},$$

For linking by M2, the population distribution, we have

- For population P1, $\vec{B} = (-0.0060, 0.0050, -0.0040)$, and

$$\mathbf{A} = \begin{pmatrix} 1.0010 & 0.0000 & 0.0000 \\ 0.0070 & 0.9999 & 0.000 \\ 0.0030 & -0.0110 & 0.9979 \end{pmatrix}_{3\times 3},$$

- For population P2, $\vec{B} = (-0.0060, -0.4922, 0.4978)$, and

$$\mathbf{A} = \begin{pmatrix} 1.0005 & 0.0000 & 0.0000 \\ 0.0060 & 1.0000 & 0.0000 \\ 0.0040 & -0.0070 & 0.9975 \end{pmatrix}_{3\times 3},$$

- For population P3, $\vec{B} = (-0.0080, 0.0000, -0.0030)$, and

$$\mathbf{A} = \begin{pmatrix} 1.0015 & 0.0000 & 0.0000 \\ -0.5746 & 1.1535 & 0.0000 \\ -0.4245 & -0.3652 & 1.2203 \end{pmatrix}_{3\times 3},$$

- For population P4, $\vec{B} = (0.9820, -0.9767, -2.1755)$, and

$$\mathbf{A} = \begin{pmatrix} 1.0010 & 0.0000 & 0.0000 \\ -0.9757 & 1.3982 & 0.0000 \\ -0.6409 & -0.6157 & 1.5327 \end{pmatrix}_{3\times 3},$$

It was found that the transformation matrix $\mathbf{A}$ and location vector $\vec{B}$ derived from LinkMIRT using the population distribution based on the estimates are very close to those derived in theory. However, those constants derived based on the common-person design were very different from those derived in theory. Figures 2 and 3 show the BIAS and correlations between the true ability values and the estimated ability values before linking and after linking. The x-axis represents the four populations and the four sample sizes. The y-axis represents the BIAS or the correlations for the three abilities General, Verbal and Math. It can be observed that population 4 had the largest bias and the smallest correlations before linking, followed by population 3; the two populations had different variance-covariance matrix from those (identity) used as the prior in BMIRT estimation. All the BIAS were reduced and correlations increased after population linking. For population P4, where a rotation is needed, linking by the population outperformed

linking by common-person and both reduced the BIAS and increased the correlations. Linking by common-person reduced the BIAS only for population P4; for other populations, the BIAS became even larger after linking by common-person. Sample size had some effect on the linking by common-person, but had no effect on the linking by population. Contrary to the general belief, large sample size may not necessarily increase the linking precision for the common-person design; in fact, the opposite was observed in this study. This suggests that using MAP ability estimates for linking by common-person may not be the right choice.

Table 4 shows the BIAS and correlations between the true item parameters and their estimates before linking and after linking for the sample size of 1000; there are three columns for the three discriminations and one column "D" for the item difficulty. For population P1, the true population distribution was the standard normal, which was the same as the priors used in the estimation; the results before linking and after linking were similar. For populations P2 and P4, the true population means were different from 0's, therefore, the BIAS for the item difficulty parameters were large and the correlations were small. However, after linking, the BIAS became smaller and the correlations became larger, with linking by the population performing better than linking by common-person. For populations P3 and P4, the true population variance-covariance matrices were different from the identity, therefore, the BIAS for the discrimination parameters were large and the correlations were small. However, after linking, the BIAS became smaller and the correlations became larger, with linking by the population performing better than linking by common-person.

Table 4

BIAS and correlations between the true values and the estimates for the item discrimination and difficulty parameters before linking and after linking.

Condition	Before Linking				Linking-M1				Linking-M2			
	G	V	M	D	G	V	M	D	G	V	M	D
					BIAS							
P1-100	.07	.03	.03	.10	.11	.15	.14	.11	.07	.03	.03	.10
P2-300	.08	.02	.04	.38	.12	.13	.13	.15	.08	.02	.04	.12
P3-500	.41	.08	.13	.09	.16	.21	.19	.10	.14	.03	.09	.09
P4-1000	.53	.17	.18	.99	.26	.32	.26	.41	.20	.08	.10	.31
					Correlations							
P1-100	.98	.99	.99	.99	.97	.98	.95	.99	.98	.99	.99	.99
P2-300	.98	.99	.98	.97	.97	.98	.95	.99	.98	.99	.98	.99
P3-500	.90	.99	.98	.99	.94	.96	.93	.99	.97	.99	.98	.99
P4-1000	.79	.98	.95	.94	.81	.90	.89	.95	.92	.98	.95	.97

Figure 4 shows the estimated item parameters before and after linking against their true values for population P4 with sample size 1000. "BeforeE" means before linking, and "AfterE" means after linking. a1, a2, a3, and b represent the discrimination parameters for the General, Verbal, and Math, and the difficulty, respectively. The closer the dot is to the diagonal line, the better the estimate is. It is clear that linking improved the item parameter estimates, and linking by population (M2) produced the largest improvements.

Because the ability estimates from Bayesian were biased, linking by common-person would improve if the ability estimates were from MLE. Item parameter estimates from BMIRT MCMC run were used as fixed and the three dimensional ability estimates from MLE were derived, then used to obtain the linking constant. Below are the transformation matrices and location vectors by linking using common-person based on MLE estimates from sample size of 1000.

- For population P1, $\vec{B} = (-0.0030, 0.0221, -0.0475)$, and

$$\mathbf{A} = \begin{pmatrix} 1.0427 & -0.0003 & 0.0157 \\ -0.0222 & 0.9725 & -0.0058 \\ 0.0242 & -0.0085 & 1.0259 \end{pmatrix}_{3\times 3},$$

- For population P2, $\vec{B} = (0.0112, -0.4619, 0.4372)$, and

$$\mathbf{A} = \begin{pmatrix} 1.0320 & -0.0007 & 0.0164 \\ -0.0094 & 0.9869 & -0.0181 \\ -0.0026 & -0.0281 & 0.9972 \end{pmatrix}_{3\times 3},$$

- For population P3, $\vec{B} = (0.0014, 0.0128, -0.0241)$, and

$$\mathbf{A} = \begin{pmatrix} 0.9863 & 0.0060 & 0.0965 \\ -0.5830 & 1.1931 & -0.1543 \\ -0.5389 & -0.0258 & 1.0815 \end{pmatrix}_{3\times 3},$$

- For population P4, $\vec{B} = (0.6967, -0.4310, -1.7831)$, and

$$\mathbf{A} = \begin{pmatrix} 0.8337 & 0.0567 & 0.1323 \\ -0.6668 & 1.3097 & -0.2867 \\ -0.5487 & -0.2124 & 1.1987 \end{pmatrix}_{3\times 3},$$

It is clear that such derived $\mathbf{A}$ and $\vec{B}$ are similar to those derived by population distribution, which are similar to those derived in theory. Therefore, when linking by common-person is conducted, the ability estimates need to be derived by MLE methods. The results for different sample sizes vary slightly; a better study design and more replications are needed to determine how many common-persons are needed for linking.

Discussion

Reporting subscores and overall scores are the ultimate goals for any Assessment. Ideally, the assessment not only evaluates the achievement level for each examinee, but also provides the students and teachers with the strengths and weaknesses of each examinee, which are useful for improving educational quality. However, certain important factors must be considered before making a decision on whether to report subscores at either the individual or institutional level. Reported scores should be valid, comparable, and reliable, with the standard applying to subscores as well. The final decision is determined by the purpose of the test. The development of the test is crucial. For the purpose of reporting accurate subscores, there should be enough items for each subskill. Inaccurate overall score information can lead

to inaccurate pass and fail decisions; similarly inaccurate information at the subscore level can lead to incorrect instructional and remedial decisions, resulting in large and needless expense for states or training sites. Similar to score estimates in the UIRT, there are Bayesian and non-Bayesian methods for the subscore estimates by MIRT models. The subscore estimates by MLE are unbiased, while the estimates by MAP and EAP are biased but have better precision than MLE. For the Bayesian methods, priors play an important role. Strong priors yield better precisions. MAP and EAP perform similarly, however, EAP takes more time than MAP; for EAP, as the number of quarture points increases, the time increases nonlinearly. The length of the test, the number of items in each subscale, and the population distributions are clearly factors that affect the accuracy of the estimation. A better design varying these factors and examining the performance of the three methods is needed.

For the subscore to be comparable across years, forms and grades or time periods, the common anchor item design, the common-person design and the randomly equivalent-groups design are possible. There have been some linking studies regarding the common-item design; a certain number of items in each subskill should be present. The common-person design and randomly equivalent-groups design are discussed and compared through simulation study in this book chapter. It was found that linking by population outperformed linking by common-person when the common-person ability estimates are based on Bayesian, especially when the population distributions are different from the standard multivariate normal - which is the default scale or matrix of many estimation softwares. For linking by common-person, the ability estimates for the common-person needs to be derived using MLE; using Bayesian does not yield accurate linking constants. The linking constants using MLE estimates for linking by common-person are found to be similar to those derived from linking by population distribution.

For tests spanned across multiple years, the objectives and the dimensional structure may vary; new objectives may sprout and objectives from previous years may diminish. Then how do we keep track of students' growth in the subscore level across years? There are many research questions that need to be answered before MIRT model can be applied to keep track of students' growth in the subscore level.

References

Ackerman, T. A., & Davey, T. C. (1991, April). *Concurrent adaptive mea-*

surement of multiple abilities. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

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

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size: A higher-order IRT approach. *Applied Psychological Measurement*, *34*, 267-285.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional IRT in test scoring. *Journal of Educational and Behavioral Statistics, 30*, 295-311.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33*, 620-639.

Dwyer, A., Boughton, K. A., Yao, L., Steffen, M., & Lewis, D. (2006, April). A comparsion of subscale score augmentation methods using empirical data. Paper presented at the meeting of the National Council on Measurement in Education, San Fransisco, CA.

Cai, L., du Toit, S. H. C., & Thissen, D. (2009). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International.

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

Haberman, J. S., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75* , 331-354.

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

Kahraman, N., & Kamata, A. (2004). Increasing the precision of subscores by using out-of scale information. *Applied Psychological Measurement, 28*, 407–426.

Kim, Y.Y. (2011). *A MIRT Application to Inform Trend Decisions in NAEP*. Study Report, Washington, D.C.: NAEP Education Statistics Services Institute. December 2011.

Kolen, M.J. & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, *29*, 8-14.

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.

Lin, P (2008). IRT vs. factore analysis approaches in analyzing multigroup multidimensional binary data: the effect of structural orthogonality, and the equivalence in test structure, item difficulty, and examinee groups. Doctor of Philosophy Dissertation at University of Maryland.

Liu, H. Y., Li, C., Zhang, P., & Luo, F. (2012). Testing Measurement Equivalence of Categorical Items' Threshold/Difficulty Parameters: A Comparison of CCFA and (M)IRT Approaches. *Acta Psychologica Sinica*, *44 (8)*, 1124-1136.

Mislevy, R.J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11,* 81-91.

Mislevy, R.J., & Sheehan, K.M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54,* 661-679.

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

Muth´en, L. K. & Muth´en, B. O. (2004). Mplus user's guide, version 3. Los Angeles, CA: Mplus.

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

Puhan, G., & liang, L. (2011). Equating subscores under the nonequivalent anchor test. *Educational measurement: Issues and Practice, 30(1)*, 23-35.

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

Reckase, M. D. (2009). Multidimensional item response theory. New York: Springer.

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

Segall, D. O. (2001). General ability measurement: an application of multidimensional item response theory. *Psychometrika*, *66*, 79-97.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47,* 150-174

Sinharay, S., & Haberman, J. S. (2011). Equating of augumented subscores. *Journal of Educational Measurement, 48*, 122-145.

Sinharay, S., Puhan, G., & Haberman, J. S. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issures and Practice, 26(4)*, 21-28.

Sinharay, S., Puhan, G., & Haberman, J. S. (2011). An NCME instructional module on subscores. *Educational Measurement: Issures and Practice, 30(3)*, 29-40.

Spiegelhalter, D. L., Thomas, A., Best, N. G., & Gilks, W. R. (1995). *BUGS: Bayesian inference using Gibbs sampling, version 3.03* (Technical Report). Cambridge, UK: Biostatistics Unit-MRC.

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

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education, 17(2)*, 89-112.

Wainer, H., Sheehan, K., & Wang, X. (2000). Some paths toward making praxis scores more useful. *Journal of Educational Measurement, 37,* 113-140.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., & Nelson, L. (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: Erlbaum.

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*, 116-136.

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

Yao, L. (2003a). *BMIRT: Bayesian multivariate item response theory.* [Computer software]. Monterey, CA: DMDC.

Yao, L. (2003b). *SimuMIRT: Simulation of multidimensional item response theory.* [Computer software]. Monterey, CA: DMDC.

Yao, L. (2004). *LinkMIRT:Linking of multidimensional item response theory.* [Computer software]. Monterey, CA: DMDC.

Yao, L. (2010a). *BMIRTII: Bayesian multivariate item response theory.* [Computer software]. Monterey, CA: DMDC.

Yao, L. (2010b). Reoporting valid and reliability overall score and domain scores. *Journal of Educational Measurement. 47*, 339-360.

Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement, 35.* 48-66.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and Applications. *Psychometrika*, 2012, DOI: 10.1007/s11336-012-9265-5, 77(3), 495-523.

Yao, L., & Boughton, K. A. (2007). A Multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement. 31*, 83-105.

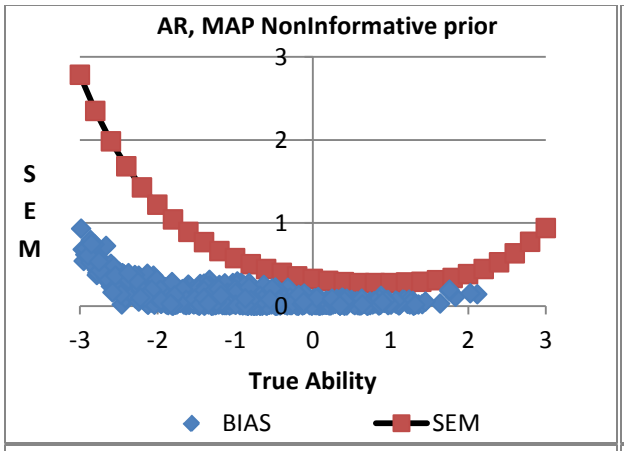
Yao, L., & Boughton, K. A. (2009). Multidimensional linking for tests containing polytomous items. *Journal of Educational Measurement, 46,* 177–197.

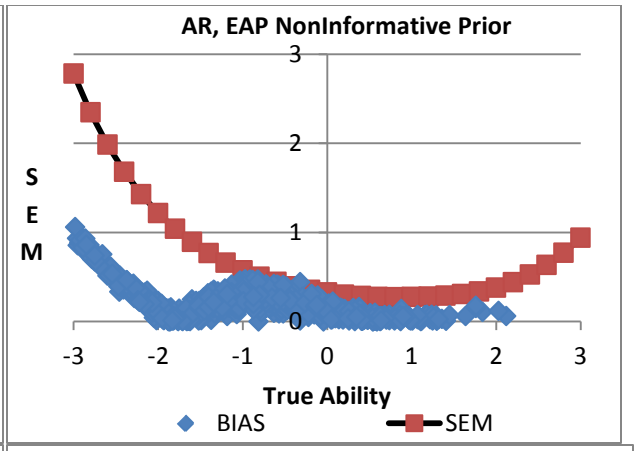
Yao, L., & Li, F. (2010, May). *A DIF detection procedure in multidimensional framework and its applications.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

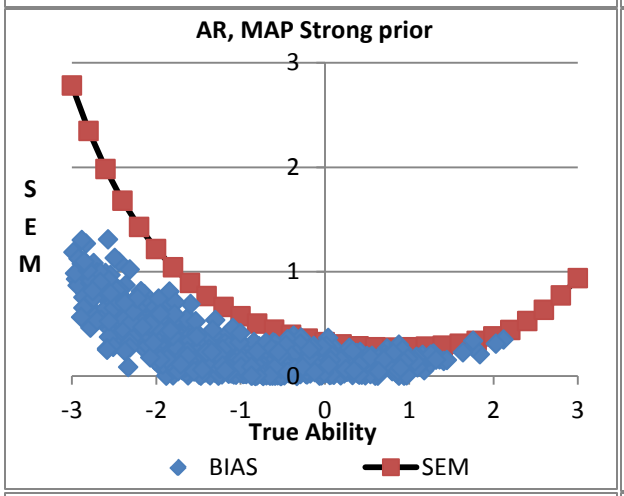
Yao, L., & Schwarz R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement. 30*, 469–492.

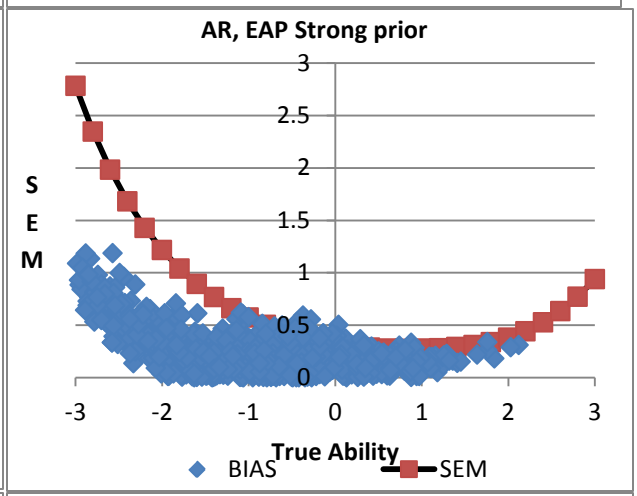
Yen, W. M. (1987, June). A Bayesian/IRT index of objective performance. Paper presented at the annual meeting of the Psychometric Society, Montreal, Qu´ebec, Canada.

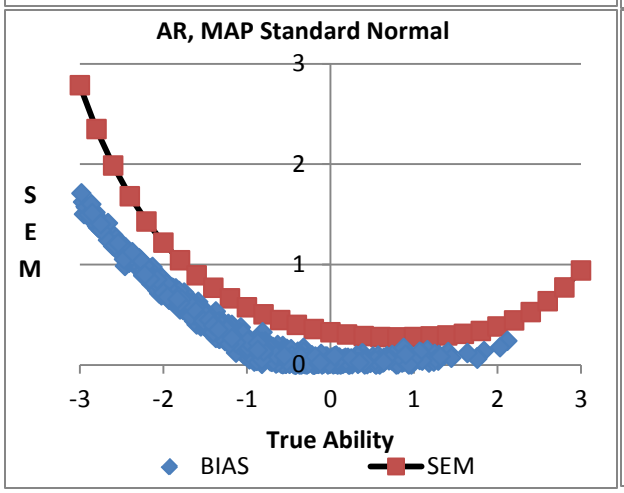
Zhang, S., Pang, X., Xu, Y., Radwan, N., & Madera, E., (2011). *Multidimensional Item Response Theory (MIRT) for Subscale Scoring.* Unpublished research report, Educational Quality and Accountability Office (EQAO), Toronto, ON, Canada

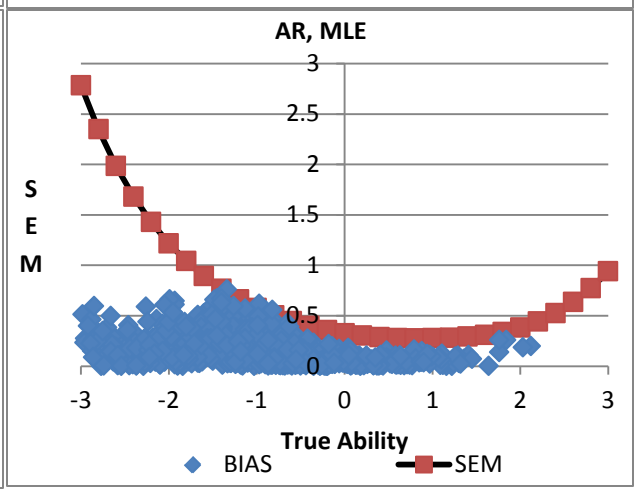

AR, MAP NonInformative prior
SEM
True Ability
BIAS
SEM

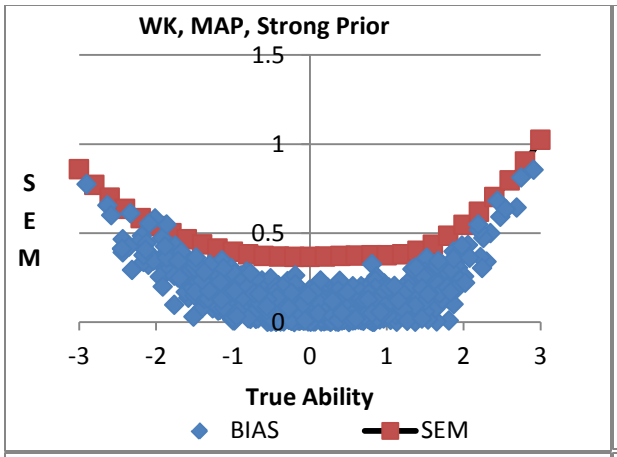

AR, EAP NonInformative Prior
SEM
True Ability
BIAS
SEM

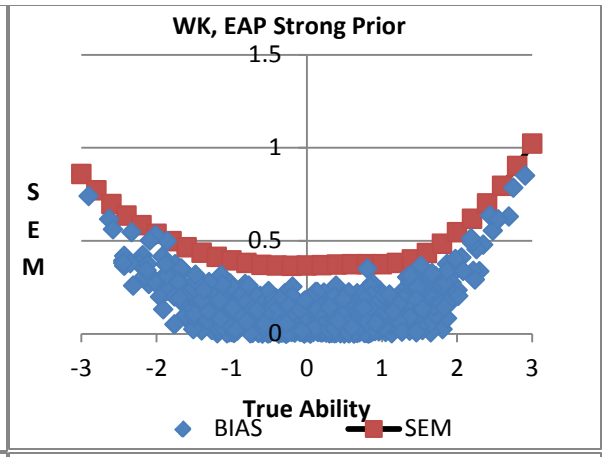

AR, MAP Strong prior
SEM
True Ability
BIAS
SEM

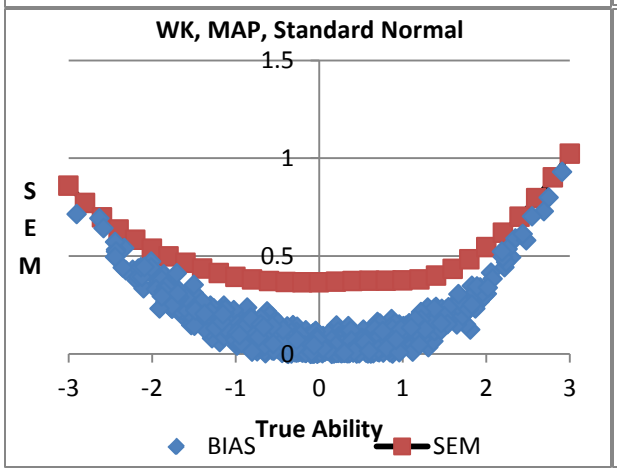

AR, EAP Strong prior
SEM
True Ability
BIAS
SEM

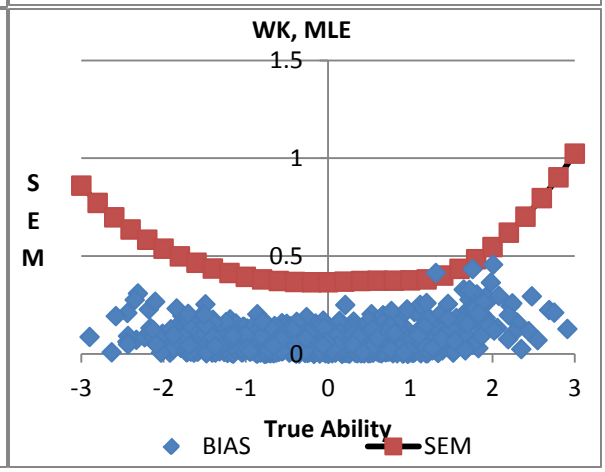

AR, MAP Standard Normal
SEM
True Ability
BIAS
SEM

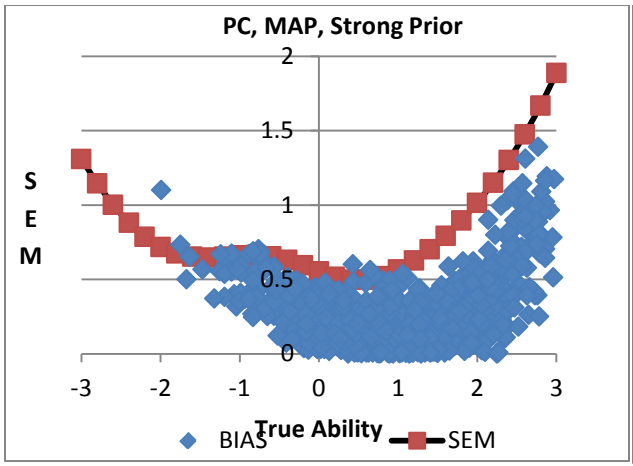

AR, MLE
SEM
True Ability
BIAS
SEM

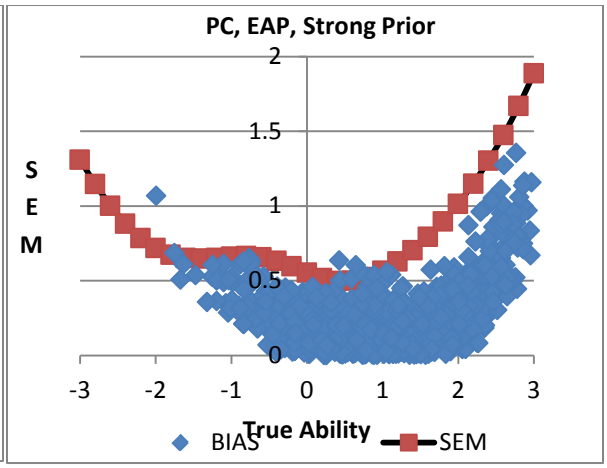

WK, MAP, Strong Prior
SEM
1.5
1
0.5
0
-3 -2 -1 0 1 2 3
True Ability
BIAS
SEM


WK, EAP Strong Prior
SEM
1.5
1
0.5
0
-3 -2 -1 0 1 2 3
True Ability
BIAS
SEM


WK, MAP, Standard Normal
SEM
1.5
1
0.5
0
-3 -2 -1 0 1 2 3
True Ability
BIAS
SEM


WK, MLE
SEM
1.5
1
0.5
0
-3 -2 -1 0 1 2 3
True Ability
BIAS
SEM


PC, MAP, Strong Prior
SEM
2
1.5
1
0.5
0
-3 -2 -1 0 1 2 3
True Ability
BIAS
SEM


PC, EAP, Strong Prior
SEM
2
1.5
1
0.5
0
-3 -2 -1 0 1 2 3
True Ability
BIAS
SEM

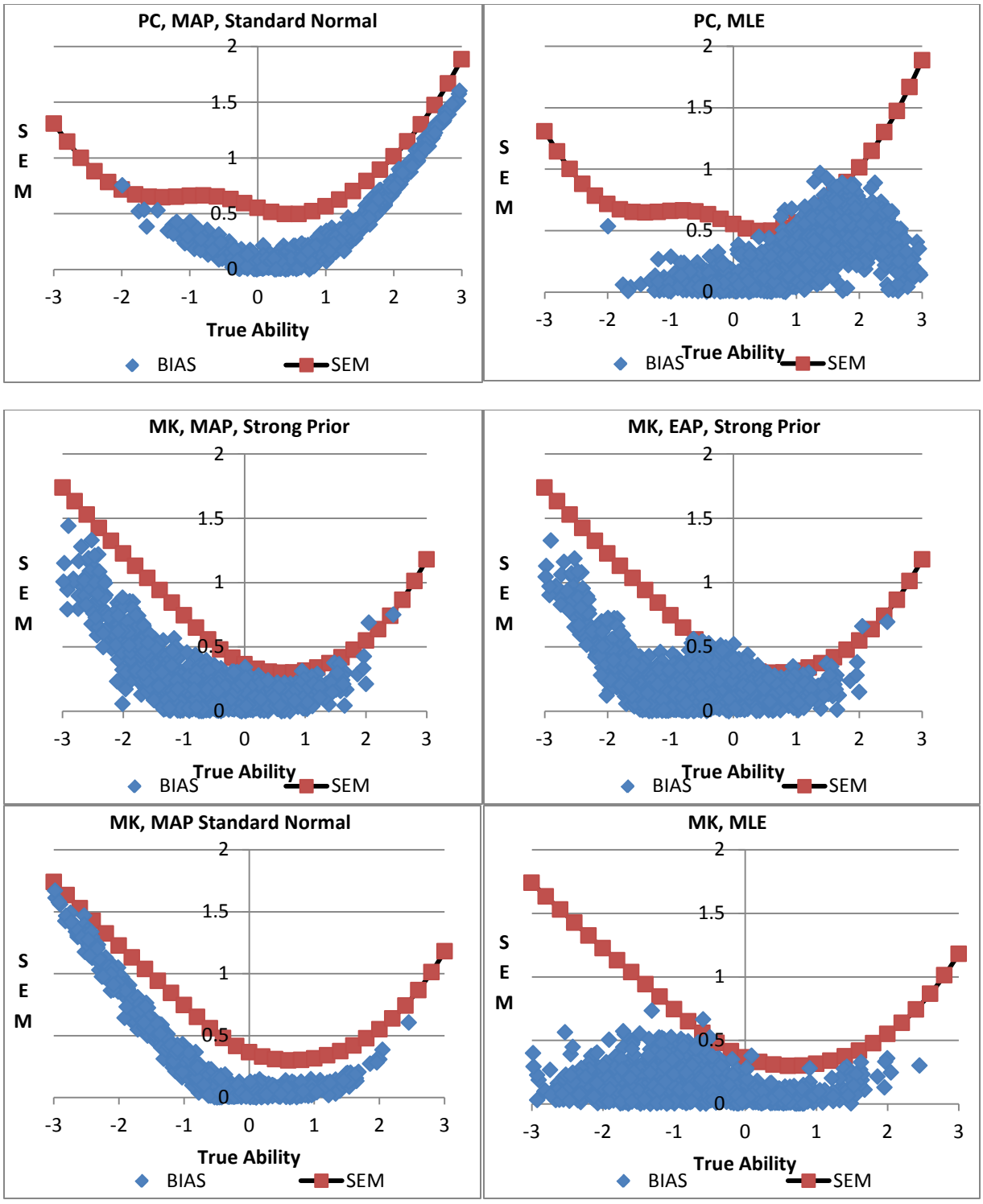
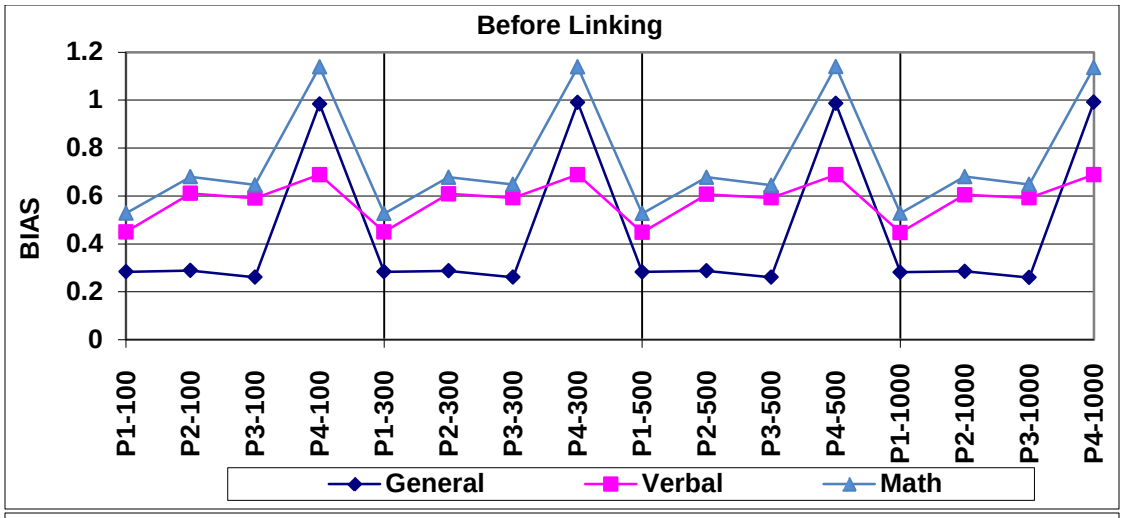
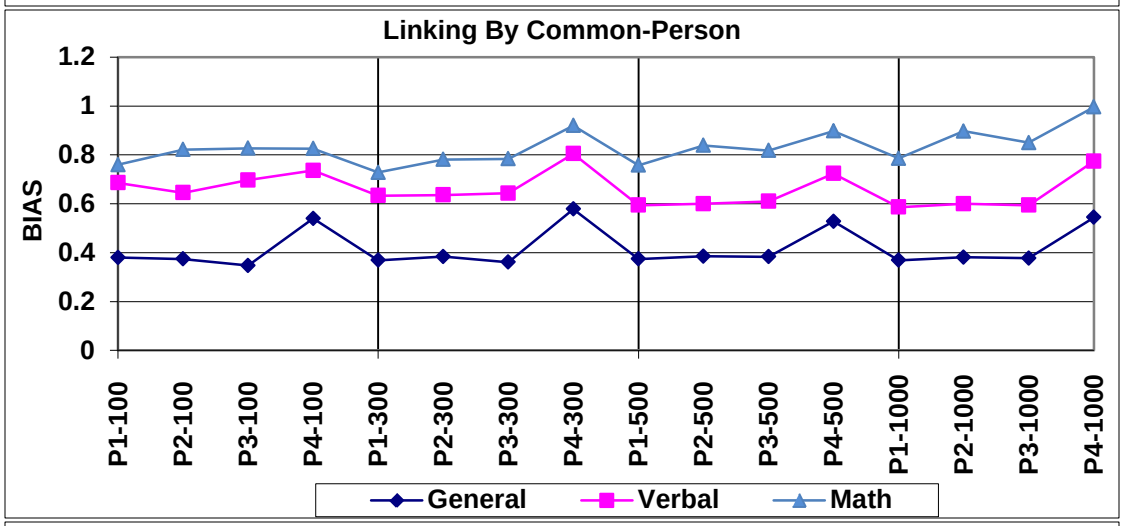


Figure 1: BIAS against the true values and the true SEM for the four contents for methods MLE, MAP, and EAP with noninformative, standard normal and strong priors.





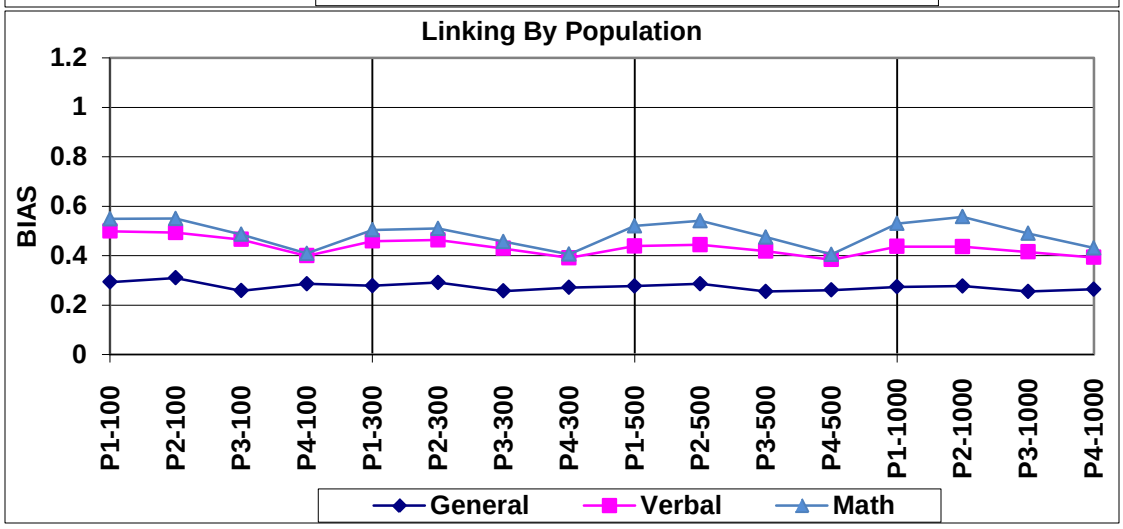


Figure 2: BIAS for the three-dimensional abilities between the estimates and the true before linking and after linking for the four populations and four sample sizes.

Note: Pi-j=Population I where i=1, 2, 3,and 4; j=100, 300, 500, and 1000.

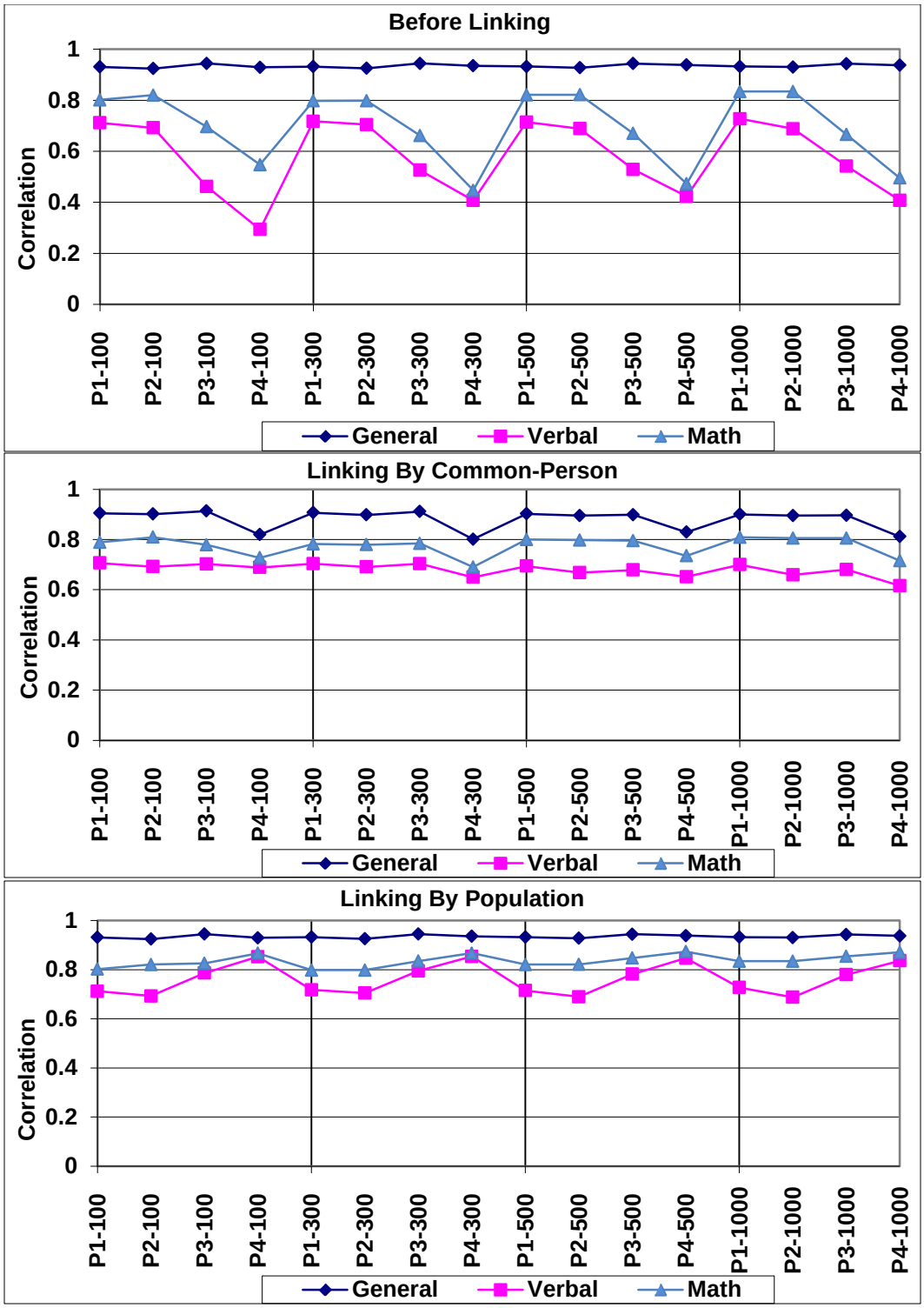
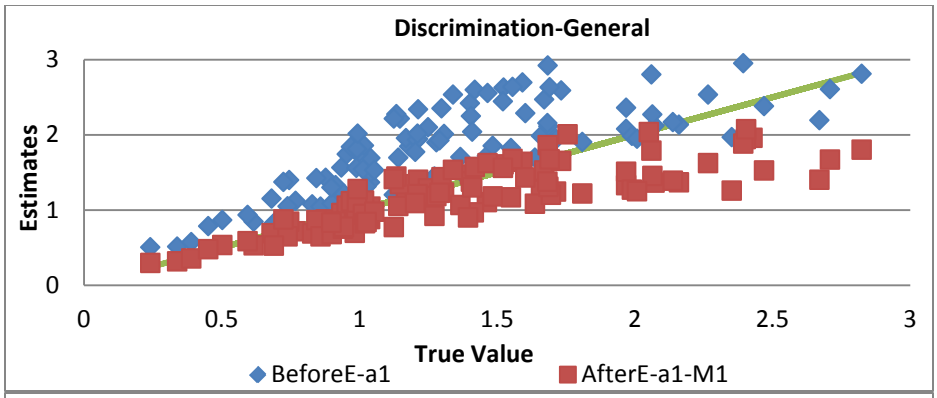
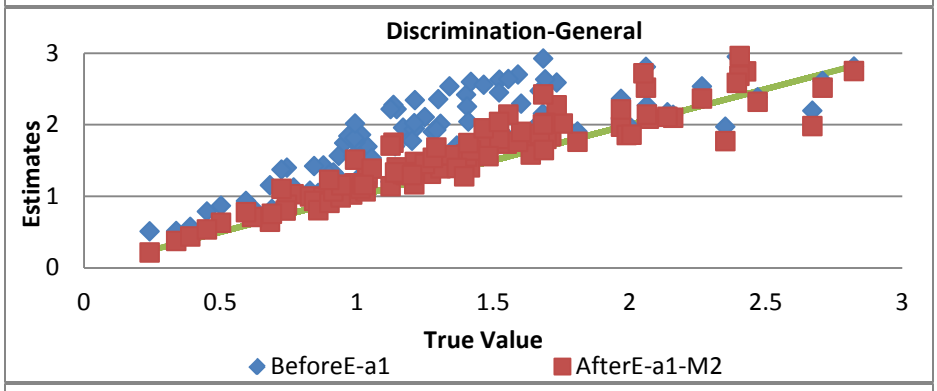
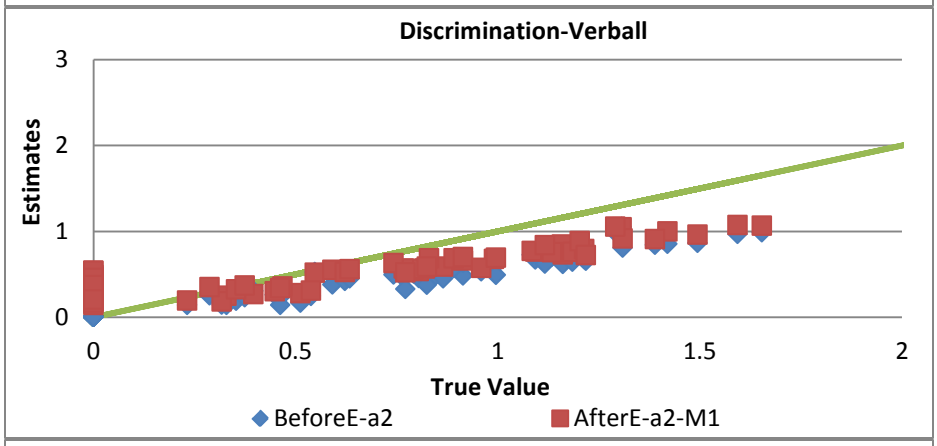


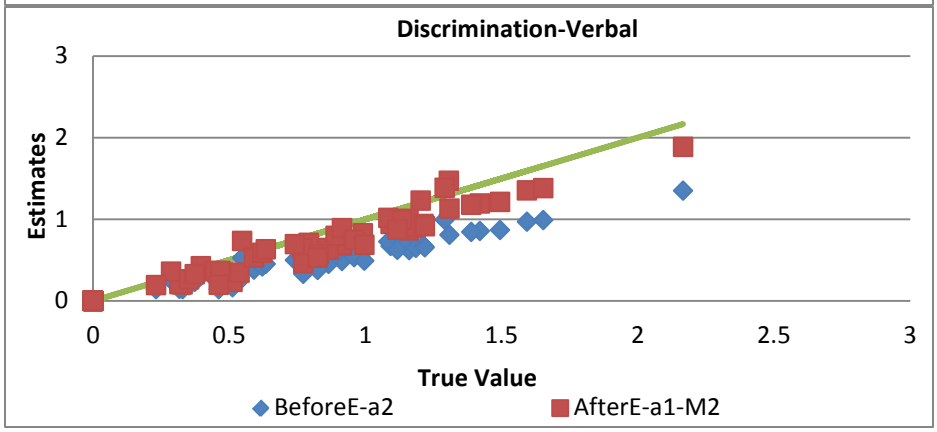
Figure 3: Correlations for the three-dimensional abilities between the estimates and the true before linking and after linking for the four populations and four sample sizes.

Note: Pi-j=Population I where i=1, 2, 3,and 4; j=100, 300, 500, and 1000.


Discrimination-General
Estimates
True Value
BeforeE-a1
AfterE-a1-M1


Discrimination-General
Estimates
True Value
BeforeE-a1
AfterE-a1-M2


Discrimination-Verball
Estimates
True Value
BeforeE-a2
AfterE-a2-M1


Discrimination-Verbal
Estimates
True Value
BeforeE-a2
AfterE-a1-M2

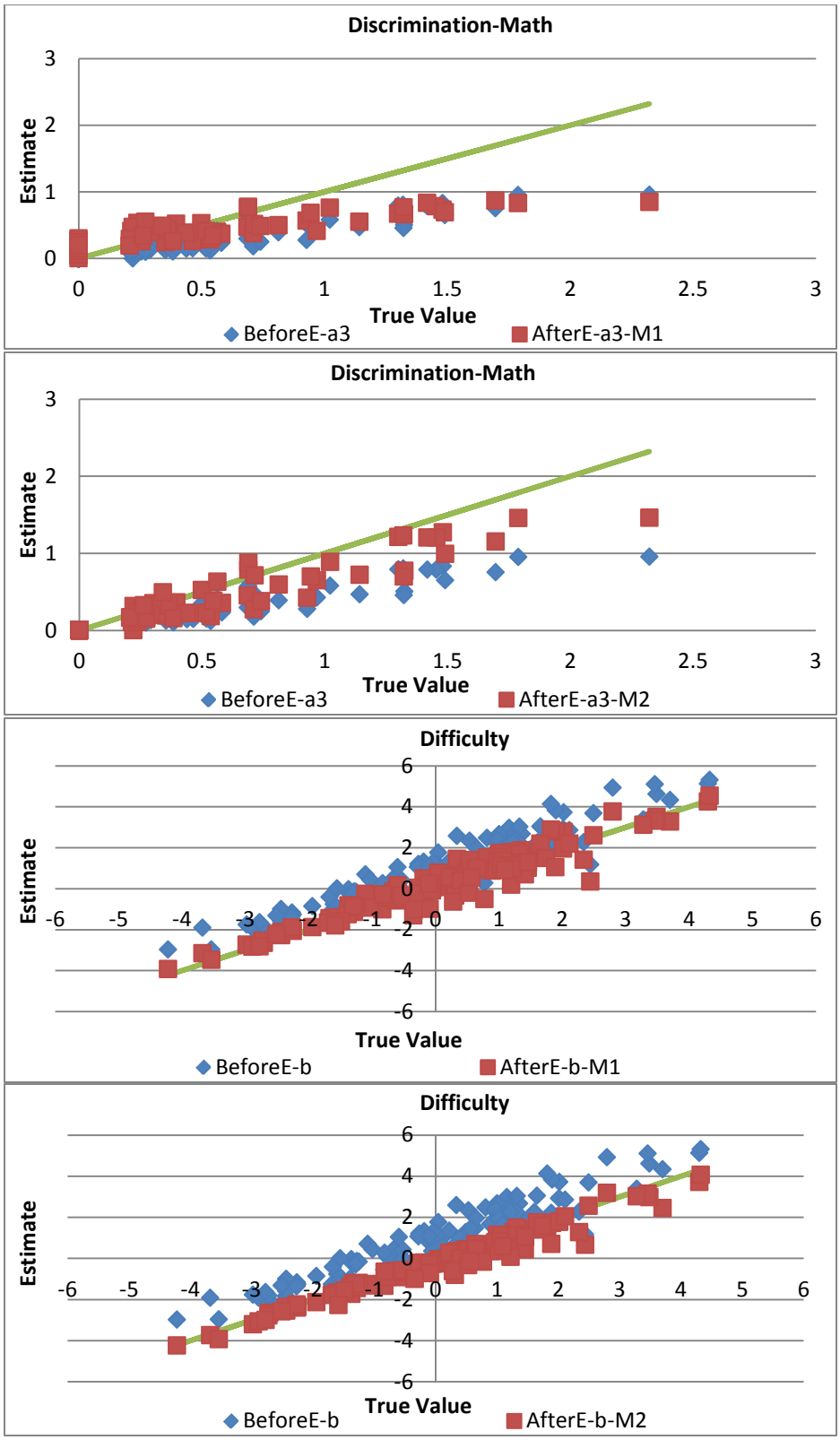


Figure 4: The estimated item parameters before and after linking against their true values for population P4 and sample size 1000.