PSYCHOMETRIKA—VOL. 80, NO. 2, 428–449
JUNE 2015
DOI: 10.1007/S11336-013-9399-0

# ON LATENT TRAIT ESTIMATION IN MULTIDIMENSIONAL COMPENSATORY ITEM RESPONSE MODELS

CHUN WANG

UNIVERSITY OF MINNESOTA

Making inferences from IRT-based test scores requires accurate and reliable methods of person parameter estimation. Given an already calibrated set of item parameters, the latent trait could be estimated either via maximum likelihood estimation (MLE) or using Bayesian methods such as maximum a posteriori (MAP) estimation or expected a posteriori (EAP) estimation. In addition, Warm's (Psychometrika 54:427–450, 1989) weighted likelihood estimation method was proposed to reduce the bias of the latent trait estimate in unidimensional models. In this paper, we extend the weighted MLE method to multidimensional models. This new method, denoted as multivariate weighted MLE (MWLE), is proposed to reduce the bias of the MLE even for short tests. MWLE is compared to alternative estimators (i.e., MLE, MAP and EAP) and shown, both analytically and through simulations studies, to be more accurate in terms of bias than MLE while maintaining a similar variance. In contrast, Bayesian estimators (i.e., MAP and EAP) result in biased estimates with smaller variability.

Key words: maximum likelihood estimation (MLE), weighted maximum likelihood estimation (WLE), multivariate weighted maximum likelihood estimation (MWLE), Bayesian estimation.

## 1. Introduction

Item response theory (IRT) has been widely applied to analyze psychological and educational test data. Accurately estimating the latent trait measured by IRT, $\theta$, is integral to maintaining the reliability of the test. Many of the common methods of trait estimation in IRT models, such as maximum likelihood estimation (MLE; Lord & Novick, 1968), Bayesian modal estimation (BME; also known as maximum a posteriori estimation, MAP) and expected a posteriori estimation (EAP), result in bias on the order of $O(n^{-1})$, where $n$ denotes test length. Many measurement contexts require the use of short tests, but standard estimation methods typically result in marked estimation bias until the test length is sufficiently long. For example, certification and licensure testing make decisions by comparing estimated $\theta$ to a cut-point, so that bias in estimating $\theta$ can yield invalid decisions. In a related context, bias in estimating patient $\theta$ for clinical diagnosis could lead to inaccurate classifications and thus, patient harm. Bias in $\theta$ would also make it difficult to equate paper-and-pencil and computerized adaptive versions of the same tests (Eignor & Schaeffer, 1995; Segall, 1996). Given these examples, psychometricians must provide essentially unbiased ability estimates in many areas of educational and psychological testing.

Firth (1993) identified two general approaches to reduce the bias of an estimate: correction and prevention. In a corrective approach, the first-order bias of the MLE is directly subtracted from $\hat{\theta}^{\text{mle}}$, and this method requires $\hat{\theta}^{\text{mle}}$ to be bounded. Anderson and Richardson (1979) and Schaefer (1983) used a corrective approach to successfully reduce the bias in estimating discrimination and location parameters of the linear logistic model. However, for models in which bias is a cubic-shaped function of the parameter being estimated, as is the case for $\hat{\theta}^{\text{mle}}$ in IRT, corrective approaches to bias reduction might inadvertently increase, rather than decrease, the

Requests for reprints should be sent to Chun Wang, University of Minnesota, 75 East River Road, Elliott Hall, N658, Minneapolis, MN, 55455, USA. E-mail: wang4066@umn.edu

estimation error (Warm, 1989). Conversely, in a preventive approach, the first derivative of the log-likelihood function (also known as the score function) is modified before finding its root. Warm (1989) proposed a preventive-based weighted likelihood estimation (WLE) method for estimating person parameters of the unidimensional three-parameter logistic IRT model. Warm found that his method reduced the magnitude of bias in the original MLE. Samejima (1993) further generalized the WLE to general IRT models for dichotomous and polytomous items.

Currently, Warm's (1989) bias reduction method has only been applied to unidimensional IRT models. However, as identified by Ackerman, Gierl, and Walker (2003), many educational and psychological tests are inherently multidimensional. For instance, a typical mathematics proficiency test measures both algebra and geometry. If an educator wants to classify examinees into mastery and non-mastery groups based upon proficiency on both dimensions, bias in estimating a latent trait vector may lead to incorrect classifications. Alternatively, personality-based measures also require unbiased multidimensional estimation. For instance, an emotional distress inventory developed by the Patient Report Outcome Information System (PROMIS) measures both the anxiety and depression level of an individual. In this case, bias in latent trait estimation yields incorrect diagnosis of a patient. To reduce the bias of a multidimensional latent trait estimate, Tseng and Hsu (2001) extended Warm's WLE method to multidimensional IRT models. However, the method proposed by these researchers only applies to tests in which each item loads on only one dimension (i.e., the test exhibits simple or clustered structure). In this paper, we extend the multidimensional analog of WLE to tests displaying complex factor structure.

When evaluating the performance of an estimator, one often finds a trade-off between an estimator's bias and its variance in repeated samples. Typically, a reduction in bias corresponds to an increase in variance. For instance, the variance of Bayesian estimators have been shown to usually be much smaller than those of MLE (see, e.g., Kim & Nicewander, 1993; Warm, 1989). In this paper, we compare the multivariate weighted MLE with the standard MLE, Bayesian expected a posteriori (EAP) estimator, and maximum a posteriori (MAP) estimator. We also discuss the variance of each estimator by means of analytical derivations and simulation studies. Not surprisingly, when the prior distribution is different from the generating distribution, both Bayesian methods (EAP and MAP) tend to engender a larger bias but also tend to have much smaller estimation variance and, therefore, slightly smaller mean squared error (MSE) as well. However, when the prior distribution is identical to the generating distribution, Bayesian methods should be preferred due to a resulting slightly smaller estimation bias and a much smaller estimation variance than alternative methods.

The rest of the paper is organized as follows. First, we introduce the multidimensional compensatory item response model as well as describe existing methods of estimating its latent trait vector. Second, we propose the multivariate weighted MLE and derive its variance. We then explain two simulation studies designed to compare the performance of all four estimation methods. One of the simulation studies assumes that the item parameters are identical to their true values, and the second simulation study allows the item parameters to contain certain levels of measurement error. We conclude by presenting results of our simulation and discussing the implications of our study for estimating multidimensional latent trait vectors in IRT models.

## 2. Model Definition and Estimation Methods

### *2.1. Multidimensional Compensatory Item Response Model*

The multidimensional three-parameter logistic IRT model (M3PL) defines the probability of an examinee with multidimensional ability vector, $\boldsymbol{\theta}_i$, correctly responding to item $j$ as follows

(Hattie, 1981):

$$P(u_{ij}=1|\boldsymbol{\theta}_i)=c_j+\frac{1-c_j}{1+\exp(-\boldsymbol{a}'_j\boldsymbol{\theta}_i-b_j)}, \tag{1}$$

where $\boldsymbol{\theta}_i=(\theta_{i1},\theta_{i2},\dots,\theta_{iK})'$ denotes a set of $K$ latent traits, $\boldsymbol{a}_j$ is a vector of $K$ discrimination parameters reflecting the relative importance of each ability in correctly answering the item, $b_j$ is the threshold which directly relates to item difficulty, and $c_j$ is an item-specific guessing parameter. The M3PL model is a direct generalization of the popular (unidimensional) three parameter logistic model (see Reckase, 2009 for a detailed description of the model; also see Mulder & van der Linden, 2009; Segall, 1996; Veldkamp & van der Linden, 2002; and Wang, Chang, & Boughton, 2011 for examples of widely used applications).

MIRT model estimation (when both item and person parameters are unknown) is challenging because there are three types of indeterminacies that need to be considered. They are: (1) the location of the origin of the multidimensional space, (2) the units of measurement for each coordinate axis, and (3) the orientation of the coordinate axes relative to the locations of the persons. The first two indeterminacies are usually dealt with by assuming a multivariate normal distribution for $\boldsymbol{\theta}$ with a mean vector containing all 0's and an identity covariance matrix (Reckase, 2009). The third indeterminacy is tackled differently by various computer programs. For instance, TESTFACT (Bock, Gibons, Schilling, Muraki, Wilson, & Wood, 2003), employing Expectation/Maximization algorithm, deals with rotational indeterminacy by setting the discrimination parameter, $a_{ik}$, to 0 for all $k>i$ to obtain an initial solution, which is then rotated to the principal factor solution. In contrast, NOHARM (Fraser, 1998: Normal-Ogive Harmonic Analysis Robust Method) fixes the orientation of coordinate axes by fixing certain $a$-parameters to be 0 (for details, please see Reckase, 2009). In this study, however, we only focus on $\boldsymbol{\theta}$ estimation assuming item parameters are pre-calibrated, so the model indeterminacy is not a concern.

### 2.2. *Current Latent Trait Estimation Methods*

Many unidimensional methods for estimating the latent person parameter have also been extended to multiple dimensions. Existing methods include maximum likelihood estimation (MLE), maximum a posteriori (MAP) estimation, and expected a posteriori (EAP) estimation. Each of these methods will now be briefly introduced. In the next section, we propose an extension of another popular unidimensional estimation method, weighted maximum likelihood estimation (WLE), to multiple dimensions for dealing with deficiencies of alternative methods.

*2.2.1. Maximum Likelihood Estimation (MLE)* Maximum likelihood estimation begins with the likelihood function given a vector of responses for a single person to $n$ items, $\boldsymbol{u}$, $L(\boldsymbol{\theta}|\boldsymbol{u})=\prod_{j=1}^{n}P_j(\boldsymbol{\theta})^{u_j}Q_j(\boldsymbol{\theta})^{1-u_j}$. Then $\hat{\boldsymbol{\theta}}^{\text{mle}}$, the multidimensional vector, is the solution to the set of $K$ simultaneous score equations,

$$\frac{\partial\ln L(\boldsymbol{u}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=\begin{bmatrix}\frac{\partial}{\partial\theta_1}\ln L(\boldsymbol{u}|\boldsymbol{\theta})\\ \frac{\partial}{\partial\theta_2}\ln L(\boldsymbol{u}|\boldsymbol{\theta})\\ \vdots\\ \frac{\partial}{\partial\theta_K}\ln L(\boldsymbol{u}|\boldsymbol{\theta})\end{bmatrix}=\begin{bmatrix}\sum_{j=1}^{n}(\frac{\partial}{\partial\theta_1}\ln\frac{P_j(\boldsymbol{\theta})}{Q_j(\boldsymbol{\theta})})[u_j-P_j(\boldsymbol{\theta})]\\ \sum_{j=1}^{n}(\frac{\partial}{\partial\theta_2}\ln\frac{P_j(\boldsymbol{\theta})}{Q_j(\boldsymbol{\theta})})[u_j-P_j(\boldsymbol{\theta})]\\ \vdots\\ \sum_{j=1}^{n}(\frac{\partial}{\partial\theta_K}\ln\frac{P_j(\boldsymbol{\theta})}{Q_j(\boldsymbol{\theta})})[u_j-P_j(\boldsymbol{\theta})]\end{bmatrix}=0. \tag{2}$$

Because one cannot derive a closed form solution to Equation (2), the Newton–Raphson procedure is often used to obtain a numerical solution (Segall, 1996). If the latent trait is unidimensional, then the bias of the MLE is equal to (see Lord, 1983):

$$B\big(\hat{\theta}^{\mathrm{mle}}\big) \approx \frac{1}{I(\hat{\theta}^{\mathrm{mle}})^2} \sum_{j=1}^{n} a_j I_j\big(\hat{\theta}^{\mathrm{mle}}\big)\big[\phi_j\big(\hat{\theta}^{\mathrm{mle}}\big) - 0.5\big], \tag{3}$$

with $\phi_j(\hat{\theta}^{\mathrm{mle}}) = \frac{P(\hat{\theta}^{\mathrm{mle}})-c_j}{1-c_j}$. Equation (3) indicates that when all of the items in a test are moderately difficult for a particular examinee, then $\phi_j(\hat{\theta}^{\mathrm{mle}})$ will be close to 0.5, so that the bias will be close to 0. If the test is too easy, then many $\phi_j(\hat{\theta}^{\mathrm{mle}})$'s will exceed 0.5 and the bias will be positive; and if the test is too difficult, then the bias will be negative. This bias term is of order $O(\frac{1}{n})$, indicating that the bias approaches 0 as test length, $n$, goes to infinity (for rigorous definition of notation used, $o(\cdot)$ and $O(\cdot)$, see Serfling, 1980). Moreover, because the bias is proportional to inverse test length, when administering short tests, MLE estimates yield large bias. The bias of the multidimensional MLE, $\hat{\boldsymbol{\theta}}^{\mathrm{mle}}$, will be introduced in a subsequent section.

It is well known that the asymptotic variance of the MLE can be approximated by the inverse of test information,

$$\mathrm{var}\big(\hat{\theta}^{\mathrm{mle}}|\theta\big) \approx \frac{1}{I(\hat{\theta}^{\mathrm{mle}})} = \left[\sum_{j=1}^{n} \frac{(1-P_j(\hat{\theta}^{\mathrm{mle}}))(P_j(\hat{\theta}^{\mathrm{mle}})-c_j)^2}{P_j(\hat{\theta}^{\mathrm{mle}})(1-c_j)^2} a_j^2\right]^{-1}, \tag{4}$$

where $\theta$ is the true ability of an individual. Therefore, if a test contains items with high discriminations, and if those items' difficulty levels are close to an examinee's true ability, $\theta$, then the test information will be large and the MLE will have small sample-to-sample variability. In multidimensional models, one finds similar relationships in that the covariance matrix of $\hat{\boldsymbol{\theta}}^{\mathrm{mle}}$ is inversely related to the Fisher information matrix, $\boldsymbol{I}(\hat{\boldsymbol{\theta}}^{\mathrm{mle}})$, so that the variance of each ability estimate, $\hat{\theta}_1^{\mathrm{mle}}, \dots, \hat{\theta}_K^{\mathrm{mle}}$, can be obtained by taking the diagonal element of $\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}^{\mathrm{mle}})$. When the M3PL model is employed, the covariance matrix of $\hat{\boldsymbol{\theta}}^{\mathrm{mle}}$ can be written as

$$\mathrm{var}\big(\hat{\boldsymbol{\theta}}^{\mathrm{mle}}|\boldsymbol{\theta}\big) \approx \left[\sum_{j=1}^{n} \frac{(1-P_j(\hat{\boldsymbol{\theta}}^{\mathrm{mle}}))(P_j(\hat{\boldsymbol{\theta}}^{\mathrm{mle}})-c_j)^2}{P_j(\hat{\boldsymbol{\theta}}^{\mathrm{mle}})(1-c_j)^2} \boldsymbol{a}_j'\boldsymbol{a}_j\right]^{-1}.$$

*2.2.2. Maximum a Posteriori (MAP) Estimation* The only difference between MLE and MAP estimation is that MAP estimation imposes a prior distribution on $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$, such that the score function needed to solve for $\hat{\boldsymbol{\theta}}^{\mathrm{MAP}}$ becomes $U(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\pi(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{u})$. MAP estimation is often employed for short tests because MLE often fails to yield a reasonable estimate of $\boldsymbol{\theta}$ for examinees who answer all items correctly or incorrectly, and adding an informative prior helps resolve such an issue. In the unidimensional case, when a standard normal prior is imposed, the MAP estimator is known to have a bias equal to Lord (1986):

$$\mathrm{Bias}\big(\hat{\theta}^{\mathrm{MAP}}\big) = \mathrm{Bias}\big(\hat{\theta}^{\mathrm{mle}}\big) - \frac{\theta}{\sum_{j=1}^{n} I_j(\theta)}.$$

Given this equation, high ability examinees will typically be underestimated whereas low ability examinees will typically be overestimated when employing MAP estimation. Therefore, one could thus think of MAP ability estimates as being 'shrunk' toward 0. These conclusions will be generalized to the multidimensional case in a subsequent section.

Lehmann and Casella (1998) provided a general theorem showing that under certain regularity conditions, the variance of the MAP estimator is the same as that of the MLE estimator.

Because the two-parameter logistic model belongs to the exponential family, the first four regularity conditions in Lehmann and Casella (1998) are automatically satisfied. Whether the fifth assumption, $\int |\theta| \pi(\theta)\, d\theta < \infty$, will be satisfied depends on the prior density of $\theta$, $\pi(\theta)$. If a standard normal prior is chosen, the fifth assumption is satisfied and, thus, $\mathrm{var}(\hat{\theta}^{\mathrm{MAP}}) \approx \mathrm{var}(\hat{\theta}^{\mathrm{mle}})$.

In multidimensional case, the asymptotic posterior covariance matrix of $\hat{\boldsymbol{\theta}}^{\mathrm{MAP}}$ can be expressed as (Segall, 1996):

$$\mathrm{var}\big(\hat{\boldsymbol{\theta}}^{\mathrm{MAP}} | \boldsymbol{\theta}\big) = \big[\boldsymbol{I}\big(\hat{\boldsymbol{\theta}}^{\mathrm{MAP}}\big) + \boldsymbol{\Sigma}_0^{-1}\big]^{-1} \tag{5}$$

assuming a multivariate normal prior on $\boldsymbol{\theta}$ with a prior covariance $\boldsymbol{\Sigma}_0$. It is easy to see that as the test length goes to infinity, the Fisher test information matrix, part of Equation (5), dominates the prior covariance matrix part. Yet most practical applications use finite-length tests, so that the MAP estimate will often be drastically less variable than the MLE due to the imposition of a prior covariance matrix, $\boldsymbol{\Sigma}_0$. In fact, this claim is supported by a well-known fact of Bayesian statistics, which is that if the likelihood and prior are both normally distributed, then the posterior precision (i.e., the inverse of the variance of an estimator) is equal to data precision plus prior precision (see Lee, 1989). Therefore, selecting an appropriate prior can greatly improve the precision of the posterior distribution (see van der Linden, 1999a, 1999b, 2008, for examples of research finding collateral information to help identify an informative prior).

*2.2.3. Expected a Posteriori (EAP) Estimation* Unlike the previous two methods, which maximize a function of the likelihood, EAP estimation takes as its point estimate of $\theta$ the mean of the posterior distribution. The $k$th element of $\hat{\boldsymbol{\theta}}^{\mathrm{EAP}}$ is therefore obtained by

$$\hat{\theta}_k^{\mathrm{EAP}} = \frac{\int \theta_k \int \cdots \int L(\boldsymbol{u}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta})\, \partial \boldsymbol{\theta}}{\int \cdots \int L(\boldsymbol{u}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta})\, \partial \boldsymbol{\theta}}. \tag{6}$$

The integrations involved in Equation (6) can be approximated by using Gauss–Hermite quadrature (Stroud & Sechrest, 1966) or simple Monte Carlo (MC) integration. If $K \geq 3$, then MC integration is generally preferred because Gauss–Hermite quadrature becomes prohibitively computationally intensive as the number of dimensions becomes large. Suppose $\boldsymbol{\theta}^r$ is the $r$th draw (out of $R$ draws in total) from $\pi(\boldsymbol{\theta})$. Then the MC estimate of $\hat{\theta}_k^{\mathrm{EAP}}$ is computed as $\hat{\theta}_k^{\mathrm{EAP}} = \sum_{r=1}^{R} \frac{\theta_k^r L(\boldsymbol{u}|\boldsymbol{\theta}^r)}{\sum_{r=1}^{R} L(\boldsymbol{u}|\boldsymbol{\theta}^r)}$. EAP estimation always yields a finite estimate of $\boldsymbol{\theta}$ regardless of the shape of the likelihood function/surface. Unfortunately, the bias and variance of the EAP estimator do not have closed analytical forms, but one can roughly infer the magnitude of its bias and variance by using the Gibbs sampler argument. For instance, assume the normal ogive (rather than the logistic) model is employed. Suppose that the prior distribution of $\theta$ is normal with a mean of $\mu_0$ and a variance of $\sigma_0^2$. Then the posterior distribution of $\theta$ will also be normal such that

$$\theta_i | \boldsymbol{z}, \mu_0, \sigma_0^2 \sim \mathcal{N}\left(\frac{\sigma_0^{-2}\mu_0 + \sum_{j=1}^{n} a_j(z_{ij} + b_j)}{\sigma_0^{-2} + \sum_{j=1}^{J} a_j^2}, \left[\sigma_0^{-2} + \sum_{j=1}^{n} a_j^2\right]^{-1}\right), \tag{7}$$

where $z_{ij}$ is a realization of the latent continuous variable $Z_{ij}$. $Z_{ij}$ follows a truncated normal distribution with a mean of $a_j\theta_i - b_j$. $Z_{ij}$ will be truncated on the left by 0 if $u_{ij} = 1$, and truncated on the right by 0 if $u_{ij} = 0$ (for details, see van der Linden, 2007).

Equation (7) is a posterior distribution conditioning on the unobserved variable $Z_{ij}$, so that the actual variance of $\hat{\theta}^{\mathrm{EAP}}$ will be greater than the variance in the posterior distribution in (7), due to the variability in $Z_{ij}$. In this regard, the variance term, $[\sigma_0^{-2} + \sum_{j=1}^{J} a_j^2]^{-1}$, provides a lower bound of the variance of $\hat{\theta}^{\text{EAP}}$. It is clear that the variance in (7) depends on the prior variance, test length, the magnitude of the item discrimination parameters, and does not depend on item difficulties. Therefore, consistent with intuition, if a test contains highly discriminative items, then the resulting EAP ability estimate will be more accurate. Moreover, examinees with different true abilities will tend to have similar var$(\hat{\theta}^{\text{EAP}})$, unlike the (bowl-shaped) variance obtained with MLE. To show the bias of $\hat{\theta}^{\text{EAP}}$, one can take an expectation of the conditional posterior mean with respect to $Z_{ij}$. Because $E(Z_{ij}) = a_j\theta_i - b_j$, we have $E(\hat{\theta}_k^{\text{EAP}}) = \frac{\sigma_\theta^{-2}\mu_\theta + \sum_{j=1}^{n} a_j\theta_j}{\sigma_\theta^{-2} + \sum_{j=1}^{J} a_j^2} \neq \theta_i$, indicating that the EAP estimate is essentially inwardly biased (which is also the case with MAP).

Similarly, in the multidimensional case, the variance of $\hat{\boldsymbol{\theta}}^{\text{EAP}}$ is the variance of the posterior distribution of $\boldsymbol{\theta}$. As before, assume the normal ogive form of the item response function (instead of the logistic form). Then $Z_{ij} + b_j = \boldsymbol{a}_j\boldsymbol{\theta}_i + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 1)$ as in the unidimensional case, so that the conditional posterior mean of $\boldsymbol{\theta}$ is

$$\left(\sum_{j=1}^{n} \boldsymbol{a}'_j \boldsymbol{a}_j + \Sigma_0^{-1}\right)^{-1} \left(\sum_{j=1}^{n} \boldsymbol{a}'_j (z_{ij} + b_j) + \boldsymbol{\mu}_0 \boldsymbol{\Sigma}_0^{-1}\right),$$

and the conditional posterior covariance matrix is $[\sum_{j=1}^{n} \boldsymbol{a}'_j\boldsymbol{a}_j + \boldsymbol{\Sigma}_0^{-1}]^{-1}$ (which is, again, the lower bound of the actual variance of the EAP estimates). Note that $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ represent the prior mean and covariance matrix, respectively. Therefore, minimizing the variance of the ability estimates requires a test with items load highly on all dimensions with little regard to the specific item difficulty values.

### 2.3. *Multivariate Weighted MLE (MWLE)*

Warm (1989) proposed a bias reduction method for unidimensional models by including the first-order bias term in the score function. For instance, let $B(\theta)$ be the bias of the MLE. Then the score function, $S(\theta)$, can be modified by $I(\theta)B(\theta)$ to become

$$S^*(\theta) = S(\theta) - I(\theta)B(\theta). \tag{8}$$

The root of $S^*(\theta)$, $\hat{\theta}^*$, is similar to $\hat{\theta}^{\text{mle}}$ but without the first order bias $O(n^{-1})$ (Warm, 1989; Firth, 1993; Wang & Wang, 2001). Moreover, the variance of $\hat{\theta}^*$ is asymptotically equivalent to that of $\hat{\theta}^{\text{mle}}$. Warm's (1989) treatment can be viewed as imposing a weight function, $w(\theta)$, on the likelihood so that $\hat{\theta}^*$ maximizes this weighted likelihood function, $L(\boldsymbol{u}|\theta)w(\theta)$.

Tseng and Hsu (2001) extended Warm's weighted maximum likelihood estimation method to a specific multidimensional case. They proposed to find a class of solutions, $\hat{\boldsymbol{\theta}}^*$, to the following estimation equations:

$$\begin{bmatrix} \frac{\partial}{\partial\theta_1} \ln L(\boldsymbol{u}|\boldsymbol{\theta}) + \frac{\partial}{\partial\theta_1} \ln w(\boldsymbol{\theta}) \\ \frac{\partial}{\partial\theta_2} \ln L(\boldsymbol{u}|\boldsymbol{\theta}) + \frac{\partial}{\partial\theta_2} \ln w(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial\theta_K} \ln L(\boldsymbol{u}|\boldsymbol{\theta}) + \frac{\partial}{\partial\theta_K} \ln w(\boldsymbol{\theta}) \end{bmatrix} = 0 \tag{9}$$

where

$$\frac{\partial}{\partial\theta_k} \ln w(\boldsymbol{\theta}) = \frac{J_k(\boldsymbol{\theta})}{I_k(\boldsymbol{\theta})}, \tag{10}$$

with $I_k(\boldsymbol{\theta}) = \sum_{j=1}^{n} \frac{[\frac{\partial}{\partial\theta_k} P_j(\boldsymbol{\theta})]^2}{P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})}$ and $J_k(\boldsymbol{\theta}) = -2\sum_{j=1}^{n} a_{jk} \frac{[\frac{\partial}{\partial\theta_k} P_j(\boldsymbol{\theta})]^2}{P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})}(P_j(\boldsymbol{\theta}) - 0.5)$. Note that (10) includes the assumption that all off-diagonal elements of the Fisher information matrix are zero. Thus, the method described by Tseng and Hsu (2001) is only effective when a test displays simple structure. However, most practicable tests have items that load on multiple dimensions, so that this simple WLE method would not apply and a more general WLE method would need to be derived.

*2.3.1. Bias of $\hat{\boldsymbol{\theta}}$: Multivariate Version* Anderson and Richardson (1979) provided a general form for the bias of the multivariate maximum likelihood estimator, $\boldsymbol{B}(\hat{\boldsymbol{\theta}}^{\text{mle}}) = [B(\theta_1), B(\theta_2), \ldots, B(\theta_K)]'$. According to these authors, the $t$th element of $\boldsymbol{B}(\hat{\boldsymbol{\theta}}^{\text{mle}})$ is

$$B\big(\hat{\theta}_t^{\text{mle}}\big) = \frac{1}{2}\sum_{k,p,q=1}^{K} I^{pt} I^{qk} \left\{ 2E\left(\frac{\partial \ln L}{\partial\theta_q}\frac{\partial^2 \ln L}{\partial\theta_p\theta_k}\right) + E\left(\frac{\partial^3 \ln L}{\partial\theta_p\,\partial\theta_q\,\partial\theta_k}\right)\right\}, \tag{11}$$

where $t = 1, 2, \ldots, K$, and $L$ denotes the short form of the likelihood function, $L(\boldsymbol{\theta}|\boldsymbol{u})$. $I^{pq}$ is the reciprocal of the $(p,q)$th term of the Fisher information matrix (i.e., $I_{pq} = -E(\partial^2 \ln L/\partial\theta_p\,\partial\theta_q)$). To simplify the computation, note that

$$\frac{\partial \ln L}{\partial\theta_q}\frac{\partial^2 \ln L}{\partial\theta_p\theta_k} = \left(\sum_{j=1}^{n}\frac{\partial \ln l_j}{\partial\theta_q}\right)\left(\sum_{j=1}^{n}\frac{\partial^2 \ln l_j}{\partial\theta_p\,\partial\theta_k}\right) = \sum_{j=1}^{n}\frac{\partial \ln l_j}{\partial\theta_q}\frac{\partial^2 \ln l_j}{\partial\theta_p\theta_k} + \sum_{j\neq m;\, j,m=1}^{n}\frac{\partial \ln l_j}{\partial\theta_q}\frac{\partial^2 \ln l_m}{\partial\theta_p\theta_k}, \tag{12}$$

where $\ln L = \sum_{j=1}^{n} \ln l_j$. $l_j \equiv l_j(\boldsymbol{\theta}|u_j)$ denotes the likelihood for a given examinee responding to item $j$. In Equation (12), the expectation of both terms always equals 0 because $E[\sum_{j\neq m;\, j,m=1}^{n} \frac{\partial \ln l_j}{\partial\theta_q}\frac{\partial^2 \ln l_m}{\partial\theta_p\theta_k}] = \sum_{j\neq m;\, j,m=1}^{n} E[\frac{\partial \ln l_j}{\partial\theta_q}]E[\frac{\partial^2 \ln l_m}{\partial\theta_p\theta_k}] = 0$ and $E[\frac{\partial \ln L}{\partial\theta_q}\frac{\partial^2 \ln L}{\partial\theta_p\theta_k}] = \sum_{j=1}^{n} E[\frac{\partial \ln l_j}{\partial\theta_q}\frac{\partial^2 \ln l_j}{\partial\theta_p\theta_k}] = 0$. Therefore, the bias in Equation (11) reduces to

$$B\big(\hat{\theta}_t^{\text{mle}}\big) = \frac{1}{2}\sum_{k,p,q=1}^{K} I^{pt} I^{qk} E\left(\frac{\partial^3 \ln L}{\partial\theta_p\,\partial\theta_q\,\partial\theta_k}\right). \tag{13}$$

For the multidimensional 2PL model,

$$E\left(\frac{\partial^3 \ln L}{\partial\theta_p\,\partial\theta_q\,\partial\theta_k}\right) = -\sum_{j=1}^{n} a_{j_p} a_{j_q} a_{j_k} P_j(\boldsymbol{\theta})\big(1 - P_j(\boldsymbol{\theta})\big)\big(1 - 2P_j(\boldsymbol{\theta})\big), \tag{14}$$

where $a_{j_p}$ indicates the $p$th discrimination parameter for the $j$th item.

Following this argument, one can also identify the bias of $\hat{\boldsymbol{\theta}}^{\text{MAP}}$. Instead of plugging the likelihood into Equation (11), one should use $L(\boldsymbol{\theta}|\boldsymbol{u})\pi(\boldsymbol{\theta})$. As is easily shown, if the prior mean is 0 and the prior covariance is an identity matrix, then the bias of the $t$th element in $\hat{\boldsymbol{\theta}}^{\text{MAP}}$ is

$$B\big(\hat{\theta}_t^{\text{MAP}}\big) = B\big(\hat{\theta}_t^{\text{mle}}\big) - \sum_{p,t,k=1}^{K} I^{pt} I^{tk}\big(I_{pk}\hat{\theta}_t^{\text{mle}}\big),$$

where $I_{pk}$ represents the $(p,k)$th element of the Fisher information matrix. In the more general case of no restrictions on the prior mean or covariance matrix, then the bias of the MAP estimator can be shown to be

$$\mathrm{Bias}\big(\hat{\theta}_t^{\mathrm{MAP}}\big) = \mathrm{Bias}\big(\hat{\theta}_t^{\mathrm{mle}}\big) - \sum_{p,t,k=1}^{K} I^{pt} I^{tk}\left( - I_{pk}\big(\hat{\boldsymbol{\theta}}^{\mathrm{mle}} - \boldsymbol{\mu}_0\big)_t \sum_{k=1}^{K} \big(\boldsymbol{\Sigma}_0^{-1}\big)_{tk}\right),$$

where $(\hat{\boldsymbol{\theta}}^{\mathrm{mle}} - \boldsymbol{\mu}_0)_t$ represents the $t$th component of the vector $(\hat{\boldsymbol{\theta}}^{\mathrm{mle}} - \boldsymbol{\mu}_0)$.

*2.3.2. Multivariate Extension of Warm's Weighted MLE* To apply Warm's correction to multidimensional IRT models, the score function must be modified by $\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{B}(\boldsymbol{\theta})$, where each element in the vector $\boldsymbol{B}(\boldsymbol{\theta})$ takes the form of (11). Note that the $k$th element of $\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{B}(\boldsymbol{\theta})$, $[\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{B}(\boldsymbol{\theta})]_k$, equals $\frac{J_k(\boldsymbol{\theta})}{I_k(\boldsymbol{\theta})}$ in (10) only when $a_i a_j = 0$ for all $i \neq j$. Therefore, as suggested earlier, Tseng and Hsu's (2001) method is a special case for dealing with simple structure tests. As in the MLE estimation, one cannot derive a closed form solution to $\boldsymbol{S}^*(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{\theta}) - \boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{B}(\boldsymbol{\theta}) = 0$, and numerical methods, such as Newton–Raphson method, must be applied.

Interestingly, note that the multidimensional 2PL model belongs to the exponential family in that the log-likelihood of $\boldsymbol{\theta}$ can be written in a canonical form as $\ln L(\boldsymbol{\theta}|\boldsymbol{u}) = \sum_{j=1}^{n}[u_j(\boldsymbol{a}_j'\boldsymbol{\theta}) - u_j b_j - \ln(1 + \exp(\boldsymbol{a}_j'\boldsymbol{\theta} - b_j))]$. Let $-[\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{B}(\boldsymbol{\theta})]_k = A_k(\boldsymbol{\theta})$. Firth (1993) showed that

$$A_k(\boldsymbol{\theta}) = \frac{1}{2}\mathrm{tr}\left[\boldsymbol{I}^{-1}(\boldsymbol{\theta})\left(\frac{\partial \boldsymbol{I}(\boldsymbol{\theta})}{\partial \theta_k}\right)\right] = \frac{\partial}{\partial \theta_k}\left[\frac{1}{2}\log\big|I(\boldsymbol{\theta})\big|\right],$$

so that the solution to $S_k^*(\boldsymbol{\theta}) \equiv S_k(\boldsymbol{\theta}) + A_k(\boldsymbol{\theta})$ locates a stationary point, $\hat{\boldsymbol{\theta}}^*$, that maximizes the modified log-likelihood,

$$\ln L^*(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) + \frac{1}{2}\log\big|\boldsymbol{I}(\boldsymbol{\theta})\big|. \tag{15}$$

The treatment by Firth (1993) is similar to imposing a Jeffrey prior, $|\boldsymbol{I}(\boldsymbol{\theta})|^{\frac{1}{2}}$, on the likelihood function, so that the first order bias of MLE is removed in calculating the posterior mode.

Warm (1989) proved, via a Taylor expansion, that the variance of $\hat{\theta}^*$ is approximately equal to $\mathrm{var}(\hat{\theta}^*) \approx \frac{\sum_{j=1}^{n} I_j(\hat{\theta}^*) + (\frac{d\ln w(\theta)}{d\theta})^2|\theta=\hat{\theta}^*}{(\sum_{j=1}^{n} I_j(\hat{\theta}^*))^2}$, where $I_j(\hat{\theta}^*)$ is the Fisher information for item $j$, and $w(\theta)$ is the weight imposed on the likelihood function. This approximation was found after discarding the higher-order terms, $o(n^{-1})$. Note that $\frac{(\frac{d\ln w(\theta)}{d\theta})^2|_{\theta=\hat{\theta}^*}}{(\sum_{j=1}^{n} I_j(\hat{\theta}^*))^2} = O(n^{-2})$, so that $\mathrm{var}(\hat{\theta}^*) \approx \mathrm{var}(\hat{\theta}^{\mathrm{mle}})$. The approximate variance of the multivariate weighted likelihood estimator $\hat{\boldsymbol{\theta}}^*$ can also be derived in a similar way. After using a Taylor expansion and dropping all terms of $o(n^{-1})$, one finds that

$$\mathrm{var}\big(\hat{\boldsymbol{\theta}}^*\big) \approx \big[\boldsymbol{I}\big(\hat{\boldsymbol{\theta}}^*\big)\big]^{-1}. \tag{16}$$

Equation (16) is derived with details provided in the Appendix. As was shown by Warm (1989) in the unidimensional case, the variance of this multivariate weighted MLE estimator is roughly the same as that of MLE.

## 3. Simulation and Results

To evaluate the performance of the weighted maximum likelihood estimator in multidimensional models, two simulation studies were conducted using the M2PL model. In each study, we compared the weighted maximum likelihood estimator, $\hat{\boldsymbol{\theta}}^*$, to: (1) $\hat{\boldsymbol{\theta}}^{\mathrm{mle}}$, (2) $\hat{\boldsymbol{\theta}}^{\mathrm{EAP}}$, and (3) $\hat{\boldsymbol{\theta}}^{\mathrm{MAP}}$.

Both EAP and MAP estimators were obtained assuming a multivariate normal prior with zero mean vector and a diagonal covariance matrix having either 1's on the off diagonals (i.e., an informative prior) or 10's on the off-diagonals (i.e., a less informative prior, denoted as $\hat{\boldsymbol{\theta}}^{\text{EAP}}$(flat) and $\hat{\boldsymbol{\theta}}^{\text{MAP}}$ (flat) hereafter).[1] Employing a less informative prior should produce smaller bias and larger variance (e.g., Wang, Hanson, & Lau, 1999). In the first simulation study, item parameters were either obtained from a real, two-dimensional test, or constructed to closely mimic a real, three-dimensional test, and in both cases, item parameters are assumed to be calibrated without error. For each condition, we constructed a sample of examinees to have specific, true ability vectors. When a two-dimensional test was considered, examinees were generated to be at 25 discrete ability points with vertices at $(-2, -2)$ and $(2, 2)$ with 1 increment. When a three-dimensional test was considered, examinees were generated to be at 27 discrete ability points with $\theta_1$, $\theta_2$, and $\theta_3 \in (-1, 0, 1)$. At each true ability point, we simulated 1000 response vectors. We then repeated the simulation by randomly generating 2000 ability vectors from a multivariate normal distribution with a zero mean vector and an identity covariance matrix. The first sample was designed to assess conditional estimation accuracy, whereas the second sample was designed to determine aggregate accuracy across a distribution. In the second simulation study, item parameters were assumed calibrated with certain measurement errors.

### *3.1. Item Parameters Without Error*

For the first simulation study, we assumed that the item parameters were calibrated without error. This assumption is commonplace in the literature (e.g., Warm, 1989; Wang et al., 1999) so as to eliminate possible contamination in estimating examinees' abilities.

*3.1.1. Two-Dimensional Test* As described above, we adopted item parameters from a real, two-dimensional test with 40 items. The test items purportedly measured two broad mathematical abilities: numerical reasoning skills (including numbers and numeration; operations and computation; patterns, functions, and algebra; data, probability, and statistics) and spatial reasoning skills (including geometry and measurement). This test was essentially simple structure, in that 39 of the 40 items only loaded on one of the two dimensions. Due to the large sample (over 2000) used to calibrate the corresponding item parameters, we assume them to be free of calibration error. The $a_1$ parameter had a mean of 1.00 and a standard deviation of 1.10; the $a_2$ parameter had a mean of 0.30 and a standard deviation of 0.64; and the $b$ parameter had a mean of 0.76 and a standard deviation of 1.78. Note that this mathematics test measures the ability along the first dimension (numerical reasoning skill) with more discrimination than the second dimension.

We first simulated multiple responses to each item conditional on specific ability vectors. Results from this simulation are presented in Figures 1 and 2. One can draw several interesting conclusions from the results. First, the MWLE resulted in a considerable decrease in the estimation bias as compared to MLE. As expected, when assuming an informative prior, both MAP and EAP (Bayesian) estimators yielded much larger absolute bias than either MLE or WMLE. These results are partly due to the informative multivariate normal prior being inappropriate for fixed true ability vectors. Conversely, when assuming a less informative prior, Bayesian estimators had a much smaller bias but a larger sampling variability. Second, consistent with Equation (3),

[1] The simulation study was conducted in MATLAB R2012a (The MathWorks, Inc., 2007). The numerical solution of MWLE was obtained via a combination of a grid search method and a nonlinear optimization algorithm. In a nutshell, we supply the modified score function $S^*(\boldsymbol{\theta})$ to 'fsolve' (a nonlinear equation solver available in MATLAB) first, and if the algorithm (the default algorithm is 'trust-region-dogleg') fails to converge, we use grid search instead to find $\hat{\boldsymbol{\theta}}^*$ that maximizes the multivariate weighted likelihood. The source codes for all estimation methods are available from the author upon request.

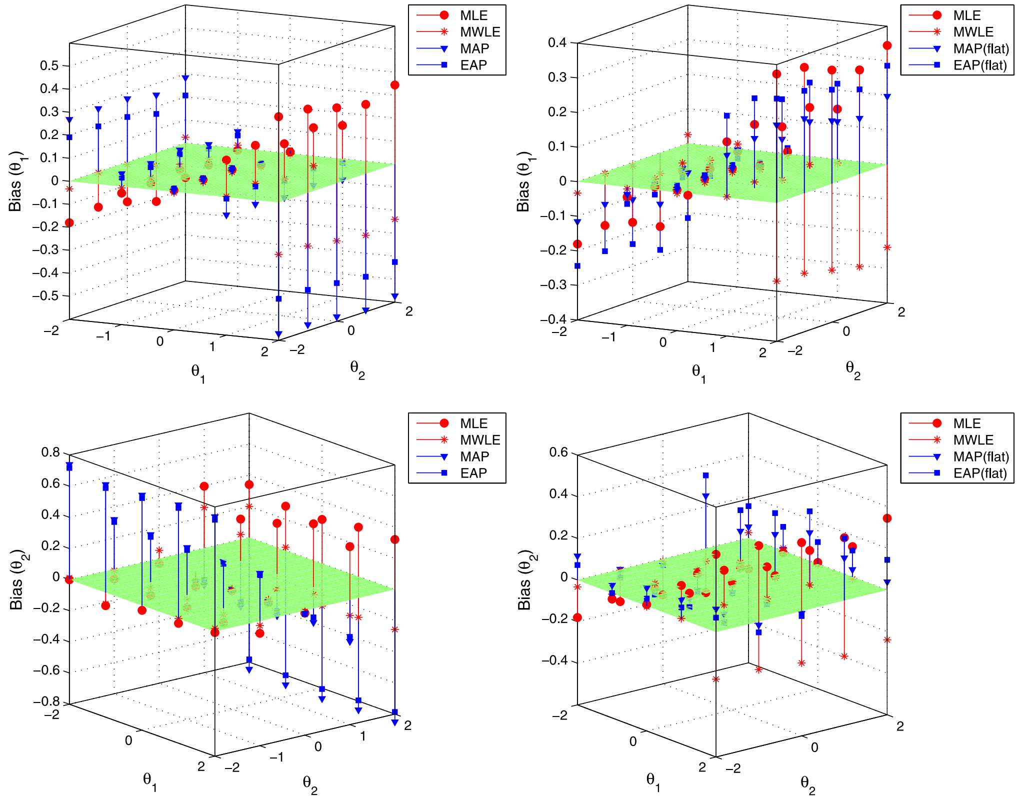

FIGURE 1.
Average bias for $\theta_1$ and $\theta_2$.

Figures 1 and 2 show that the MLE was biased outward, whereas Bayesian estimates with informative priors were biased inwards. The direction of bias in Bayesian estimators reversed when a less informative prior was used. Therefore, when using maximum likelihood estimation, examinees with low ability on a dimension tend to have an underestimated ability on that dimension, whereas examinees with high ability on a dimension tend to have overestimated ability. In contrast to alternative estimators, MWLE results in the smallest estimation bias. Third, both MLE and MWLE typically generate larger variance than the Bayesian estimators. Moreover, in line with the analytical results, the estimated variance for Bayesian estimates is nearly uniform across true ability points whereas $\hat{\boldsymbol{\theta}}^*$ and $\hat{\boldsymbol{\theta}}^{\text{mle}}$ have larger variance for examinees with more extreme true abilities.

Table 1 presents the summary of the estimation accuracy for both the conditional results (as in Figures 1 and 2) and the results when randomly sampling ability from a multivariate distribution. With the sample of fixed 25 ability points, these results include the average bias ($\frac{1}{25}\sum_{j=1}^{25}(\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\theta}_{ij}-\theta_{ij})))$, conditional mean absolute bias (MAB, $\frac{1}{25}\sum_{j=1}^{25}|\frac{1}{1000}\times\sum_{i=1}^{1000}(\hat{\theta}_{ij}-\theta_{ij})|$), conditional variance ($\frac{1}{25}\sum_{j=1}^{25}(\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\theta}_{ij}-\bar{\theta}_j)^2)$), and mean squared error (MSE, $\frac{1}{25}\sum_{j=1}^{25}(\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\theta}_{ij}-\theta_{ij})^2)$), where $j$ denotes the $j$th ability point, and $\hat{\theta}_{ij}$ is the ability estimate for $i$th examinee with true $j$th ability vector on a certain dimension. Here, only scalar notation is used for simplicity. But such summary statistics were computed for each dimension separately. With the general bivariate normal sample, we present the average bias ($\frac{1}{2000}\sum_{i=1}^{2000}(\hat{\theta}_i-\theta_i)$), mean absolute bias ($\frac{1}{2000}\sum_{i=1}^{2000}|\hat{\theta}_i-\theta_i|$), and mean squared error

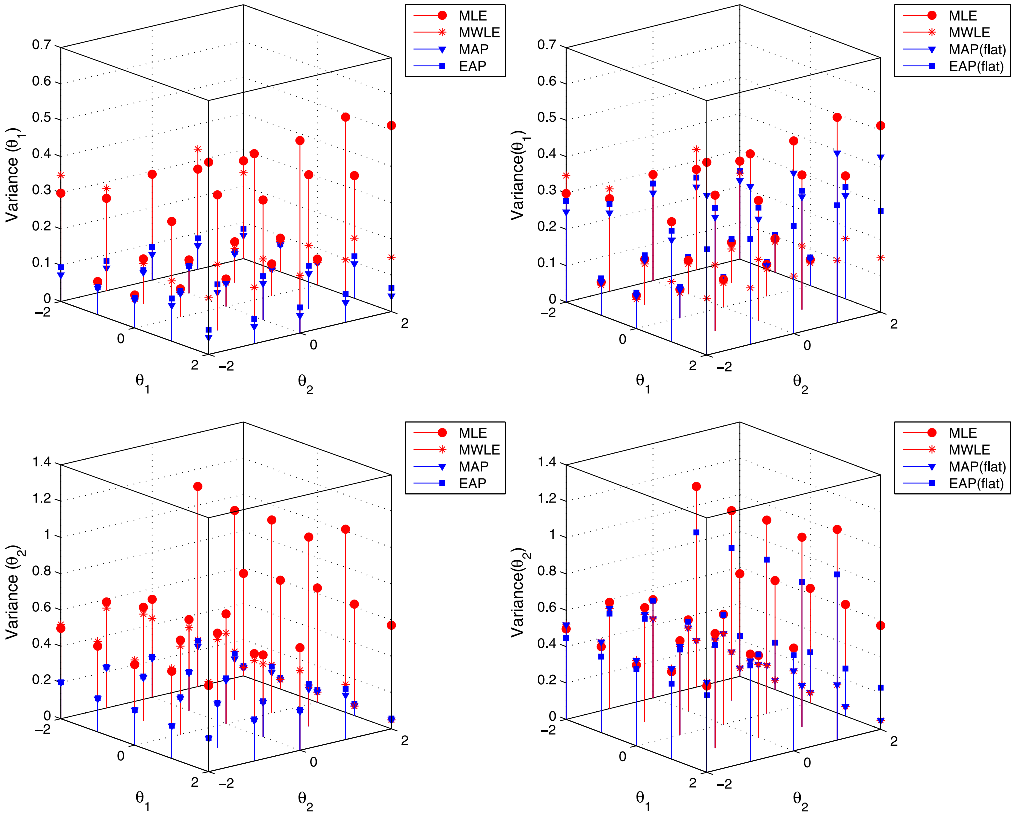

FIGURE 2.
Variance of $\hat{\theta}_1$ and $\hat{\theta}_2$.

$(\frac{1}{2000}\sum_{i=1}^{2000}(\hat{\theta}_i-\theta_i)^2)$. The variance cannot be computed in this sample because such a variance is contaminated by the actual variability of the true $\boldsymbol{\theta}$'s in the sample.

Consistent with Figures 1 and 2,[2] WMLE reduced the estimation bias and mean absolute bias of latent trait estimation as compared to MLE. If sampling ability vectors from a multivariate normal distribution (i.e., second part of the table), then EAP and MAP estimation with informative priors tended to result in the smallest mean absolute bias among all estimators. This result is an artifact of the prior ability distribution being identical to the generating distribution. Yet, when assuming a less informative priors, then WMLE resulted in a smaller absolute bias and MSE than Bayesian estimators.

*3.1.2. Three-Dimensional Test* As an extension of the previous simulation, we conducted a follow-up simulation study designed to evaluate the performance of the aforementioned latent trait estimators when a test measures three dimensions. The three-dimensional test was constructed from 30 dichotomously-scored items obtained from Reckase (2009) (p. 153, Table 6.1). The items adopted for this simulation are similar to a typical set of multidimensional ability items, in that every item loads on all three dimensions but each item prominently measures only one trait. We adopted the same conditions as in the previous simulation: two sets of examinees, four ability estimators, and two prior distributions.

[2] In fact, the first part of the table is a different way of presenting the results in Figures 1 and 2.

TABLE 1.
Average bias, mean absolute bias, variance and MSE for $\hat{\boldsymbol{\theta}}$ estimation.

| Sample | | $\hat{\boldsymbol{\theta}}^{\text{mle}}$ | | $\hat{\boldsymbol{\theta}}^{*}$ | | $\hat{\boldsymbol{\theta}}^{\text{MAP}}$ | | $\hat{\boldsymbol{\theta}}^{\text{MAP}}$ (flat) | | $\hat{\boldsymbol{\theta}}^{\text{EAP}}$ | | $\hat{\boldsymbol{\theta}}^{\text{EAP}}$ (flat) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ |
| Fixed ability points | Bias | 0.063 | 0.101 | −0.052 | −0.026 | 0.041 | 0.055 | −0.064 | −0.024 | −0.037 | −0.013 | 0.085 | 0.051 |
| | Conditional mean absolute bias | 0.142 | 0.163 | 0.059 | 0.149 | 0.200 | 0.435 | 0.091 | 0.137 | 0.138 | 0.410 | 0.125 | 0.164 |
| | Variance | 0.209 | 0.556 | 0.158 | 0.363 | 0.072 | 0.180 | 0.136 | 0.363 | 0.086 | 0.192 | 0.214 | 0.516 |
| | Conditional MSE | 0.273 | 0.601 | 0.170 | 0.451 | 0.155 | 0.463 | 0.242 | 0.536 | 0.131 | 0.444 | 0.253 | 0.540 |
| Bivariate normal | Bias | 0.040 | 0.110 | −0.021 | 0.019 | −0.031 | 0.014 | 0.032 | 0.076 | 0.011 | 0.019 | 0.050 | 0.065 |
| | Mean absolute bias | 0.307 | 0.669 | 0.261 | 0.529 | 0.259 | 0.463 | 0.305 | 0.617 | 0.255 | 0.465 | 0.311 | 0.606 |
| | MSE | 0.195 | 0.753 | 0.118 | 0.441 | 0.113 | 0.341 | 0.171 | 0.618 | 0.109 | 0.342 | 0.187 | 0.590 |

Figures 3 and 4 present the average bias and variance of each estimator conditioning on various true ability values. The relative performance of each ability estimator was similar to that of the previous section. Namely, MWLE resulted in smaller bias than MLE, and Bayesian estimation either yielded estimates with the largest bias and smallest variance (if using an informative prior) or estimates similar to those obtained via maximum likelihood estimation (if using a less informative prior). Similarly to Table 1, Table 2 presents the summary statistics of $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ estimates for both samples of examinees. Consistent with the previous section, EAP and MAP estimators outperformed MLE and MWLE only when the prior and true distributions coincide. In all other cases, MWLE estimation resulted in the best trade-off between bias and variance.

As demonstrated in this section, given a short test loading on either two or three dimensions, MWLE results in the most preferable balance between small bias and acceptable variance in almost all cases. The only cases in which MWLE performed worse than Bayesian estimators were those in which the prior distribution aligned perfectly with the generating distribution. However, one limitation with the simulation is we assume all item parameters were calibrated without measurement errors.

### *3.2. Item Parameters with Error*

We followed up the initial simulation with a second simulation that includes estimation error in the item parameters. Specifically, examinee responses were generated given the true item parameters, but abilities were estimated with knowledge only from the estimated item parameters. Unlike the previous simulation, in which items mostly loaded on only one dimension, we constructed item banks for this simulation that conformed to both simple structure (with items loading on only one dimension and half of the items loading on each dimension) and complex structure (with items loading similarly on both dimensions). Similarly to van der Linden (1999a, 1999b), all item discrimination parameters were generated from $\mathcal{U}(0, 3)$[3] (a uniform distribution with a minimum of 0 and a maximum of 3), and all $b$-parameters were generated from $\mathcal{N}(0, 1)$. These item parameters were then estimated given a separate set of responses from two normally distributed calibration samples, both with a mean of $(0, 0)$ and a covariance matrix of $[1, 0.5; 0.5, 1]$. The first sample was of size 250, which should result in a high degree of calibration error, whereas the second sample was of size 1000, which should result in less calibration error. Item parameters were calibrated using the 'MIRT' package in R (Chalmers, 2012). This package calibrates multidimensional IRT item parameters with the Metropolis–Hastings Robbins–Monro (Cai, 2008, 2010a, 2010b) algorithm. Performance of each ability estimation algorithm was determined from a separate set of 1000 response vectors (simulated using the true item parameters) at each of 25 ability points on the $[-2, -2]$ to $[2, 2]$ square. The average bias, conditional mean absolute bias, conditional variance, and MSE are presented in Table 3.

Note that the same pattern is shown in Table 3 as presented in the previous simulations, that is, MWLE yielded considerably reduced estimation bias and slightly decreased estimation variance compared to MLE. Both MLE and MWLE resulted in much smaller bias but much larger variance than Bayesian estimators with informative priors. Bayesian estimators with less informative priors led to dramatically decreased bias but also increased variance. Due to the observed similar average bias, variance, and MSE for both small ($N = 250$) or relatively large ($N = 1000$) calibration samples, the results indicated that the measurement error associated with using estimated item parameters had a small effect on the latent trait estimation accuracy.

[3] In van der Linden (1999a, 1999b), the discrimination parameters were generated from $U(0, 1.3)$ whereas we used a higher upper bound due to the relatively short test length and non-adaptive tests.

TABLE 2.
Average bias, mean absolute bias, variance and MSE for $\hat{\boldsymbol{\theta}}$ estimation.

| | | Fixed true ability point | | | | Multivariate normal sample | | |
|---|---|---|---|---|---|---|---|---|
| | | Bias | Conditional MAB | Conditional variance | MSE | Bias | MAB | MSE |
| $\hat{\boldsymbol{\theta}}^{\text{mle}}$ | $\hat{\theta}_1$ | −0.028 | 0.063 | 0.641 | 0.645 | −0.016 | 0.631 | 0.622 |
| | $\hat{\theta}_2$ | −0.020 | 0.068 | 0.571 | 0.576 | −0.040 | 0.560 | 0.493 |
| | $\hat{\theta}_3$ | −0.024 | 0.076 | 0.538 | 0.544 | −0.022 | 0.596 | 0.559 |
| $\hat{\boldsymbol{\theta}}^{*}$ | $\hat{\theta}_1$ | 0.008 | 0.046 | 0.505 | 0.507 | 0.011 | 0.566 | 0.501 |
| | $\hat{\theta}_2$ | 0.010 | 0.037 | 0.451 | 0.452 | −0.005 | 0.512 | 0.419 |
| | $\hat{\theta}_3$ | −0.002 | 0.035 | 0.438 | 0.439 | 0.000 | 0.544 | 0.474 |
| $\hat{\boldsymbol{\theta}}^{\text{MAP}}$ | $\hat{\theta}_1$ | 0.004 | 0.254 | 0.213 | 0.298 | 0.017 | 0.472 | 0.347 |
| | $\hat{\theta}_2$ | 0.010 | 0.228 | 0.203 | 0.273 | −0.007 | 0.456 | 0.338 |
| | $\hat{\theta}_3$ | 0.017 | 0.220 | 0.196 | 0.263 | 0.004 | 0.463 | 0.350 |
| $\hat{\boldsymbol{\theta}}^{\text{MAP}}$ (flat) | $\hat{\theta}_1$ | −0.032 | 0.056 | 0.691 | 0.694 | −0.011 | 0.655 | 0.727 |
| | $\hat{\theta}_2$ | −0.021 | 0.059 | 0.621 | 0.625 | −0.027 | 0.583 | 0.577 |
| | $\hat{\theta}_3$ | −0.010 | 0.063 | 0.575 | 0.579 | −0.011 | 0.597 | 0.582 |
| $\hat{\boldsymbol{\theta}}^{\text{EAP}}$ | $\hat{\theta}_1$ | −0.011 | 0.238 | 0.229 | 0.305 | 0.004 | 0.487 | 0.375 |
| | $\hat{\theta}_2$ | −0.003 | 0.200 | 0.225 | 0.279 | −0.006 | 0.474 | 0.361 |
| | $\hat{\theta}_3$ | 0.001 | 0.203 | 0.223 | 0.277 | −0.003 | 0.452 | 0.323 |
| $\hat{\boldsymbol{\theta}}^{\text{EAP}}$ (flat) | $\hat{\theta}_1$ | −0.073 | 0.159 | 0.891 | 0.920 | −0.008 | 0.610 | 0.714 |
| | $\hat{\theta}_2$ | −0.048 | 0.185 | 0.851 | 0.894 | −0.066 | 0.596 | 0.759 |
| | $\hat{\theta}_3$ | −0.025 | 0.178 | 0.788 | 0.824 | −0.057 | 0.601 | 0.748 |

TABLE 3.
Average bias, conditional mean absolute bias, conditional variance, and MSE for $\hat{\boldsymbol{\theta}}$ estimation.

| Test structure | Sample size | | $\hat{\boldsymbol{\theta}}^{\text{mle}}$ | | $\hat{\boldsymbol{\theta}}^{*}$ | | $\hat{\boldsymbol{\theta}}^{\text{EAP}}$ | | $\hat{\boldsymbol{\theta}}^{\text{MAP}}$ | | $\hat{\boldsymbol{\theta}}^{\text{EAP}}$ (flat) | | $\hat{\boldsymbol{\theta}}^{\text{MAP}}$ (flat) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ |
| Simple | 250 | Bias | 0.045 | −0.048 | −0.013 | −0.042 | 0.078 | 0.084 | −0.041 | −0.080 | 0.024 | −0.042 | 0.026 | −0.052 |
| | | Conditional MAB | 0.192 | 0.117 | 0.150 | 0.101 | 0.317 | 0.354 | 0.371 | 0.412 | 0.157 | 0.116 | 0.174 | 0.111 |
| | | Conditional variance | 0.442 | 0.433 | 0.346 | 0.337 | 0.112 | 0.136 | 0.094 | 0.106 | 0.429 | 0.485 | 0.376 | 0.437 |
| | | MSE | 0.489 | 0.435 | 0.377 | 0.354 | 0.305 | 0.285 | 0.262 | 0.335 | 0.445 | 0.506 | 0.413 | 0.448 |
| | 1000 | Bias | 0.035 | 0.023 | −0.041 | 0.009 | −0.023 | −0.001 | −0.034 | −0.007 | 0.004 | 0.027 | −0.019 | 0.022 |
| | | Conditional MAB | 0.185 | 0.117 | 0.110 | 0.070 | 0.312 | 0.326 | 0.321 | 0.382 | 0.140 | 0.096 | 0.204 | 0.115 |
| | | Conditional variance | 0.421 | 0.375 | 0.228 | 0.305 | 0.110 | 0.126 | 0.090 | 0.106 | 0.426 | 0.404 | 0.364 | 0.415 |
| | | MSE | 0.470 | 0.414 | 0.241 | 0.311 | 0.233 | 0.298 | 0.254 | 0.290 | 0.450 | 0.482 | 0.415 | 0.432 |
| Complex | 250 | Bias | 0.017 | 0.041 | 0.005 | −0.008 | −0.007 | 0.027 | −0.016 | 0.030 | 0.031 | 0.023 | 0.025 | 0.027 |
| | | Conditional MAB | 0.223 | 0.145 | 0.129 | 0.131 | 0.366 | 0.457 | 0.394 | 0.467 | 0.145 | 0.117 | 0.211 | 0.142 |
| | | Conditional variance | 0.518 | 0.615 | 0.416 | 0.474 | 0.133 | 0.155 | 0.112 | 0.137 | 0.433 | 0.392 | 0.384 | 0.477 |
| | | MSE | 0.589 | 0.651 | 0.457 | 0.413 | 0.326 | 0.445 | 0.345 | 0.447 | 0.445 | 0.510 | 0.464 | 0.523 |
| | 1000 | Bias | 0.013 | 0.043 | 0.019 | 0.004 | 0.068 | 0.092 | −0.017 | 0.032 | 0.029 | 0.028 | 0.023 | 0.031 |
| | | Conditional MAB | 0.214 | 0.189 | 0.086 | 0.135 | 0.366 | 0.470 | 0.379 | 0.468 | 0.147 | 0.119 | 0.215 | 0.147 |
| | | Conditional variance | 0.487 | 0.551 | 0.337 | 0.394 | 0.131 | 0.149 | 0.114 | 0.133 | 0.428 | 0.399 | 0.382 | 0.484 |
| | | MSE | 0.553 | 0.614 | 0.348 | 0.432 | 0.323 | 0.459 | 0.327 | 0.442 | 0.459 | 0.528 | 0.442 | 0.518 |

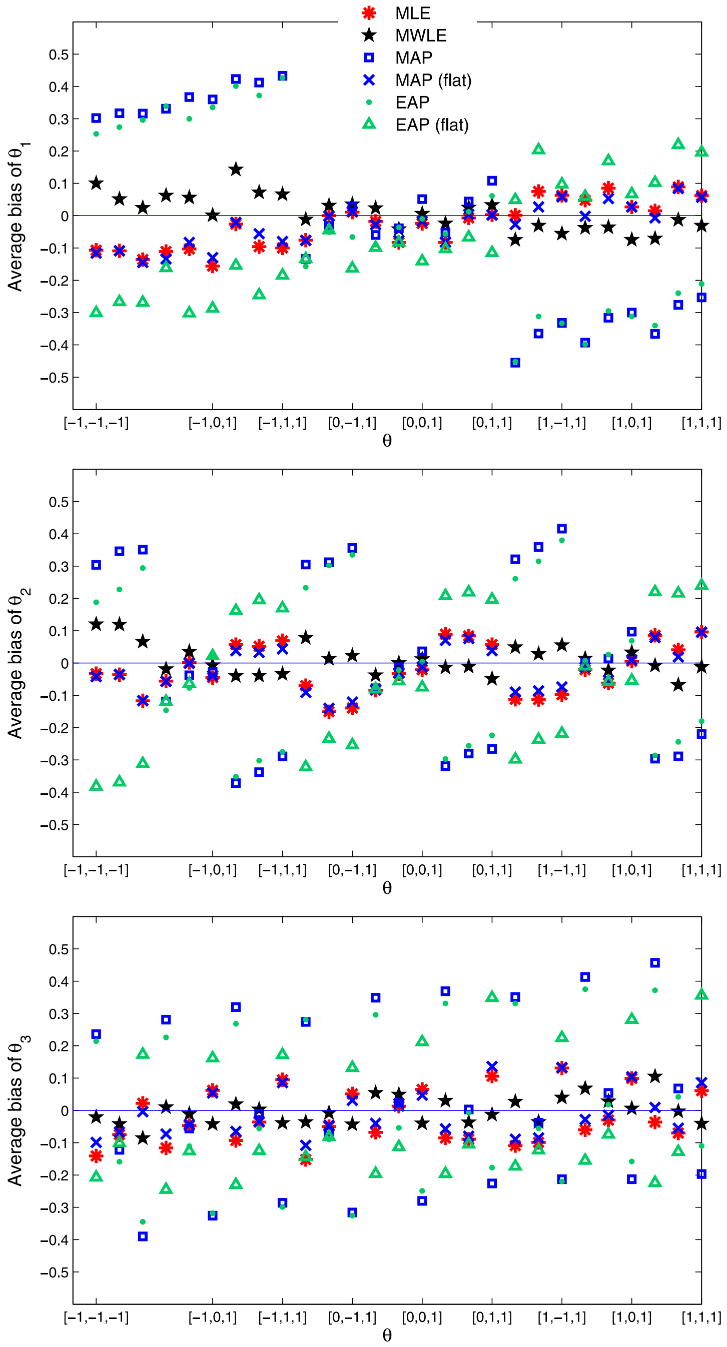

FIGURE 3.
Average bias for $\theta_1$, $\theta_2$, and $\theta_3$ estimates in a three-dimensional case.

Interestingly, Table 3 indicates that examinees administered a simple structure test tend to have ability that is estimated with less bias across all estimators. The reason for the decrease in bias when employing a simple structure test can be demonstrated using properties of the MLE bias function. According to Equation (13), if a two-dimensional test displays simple structure,

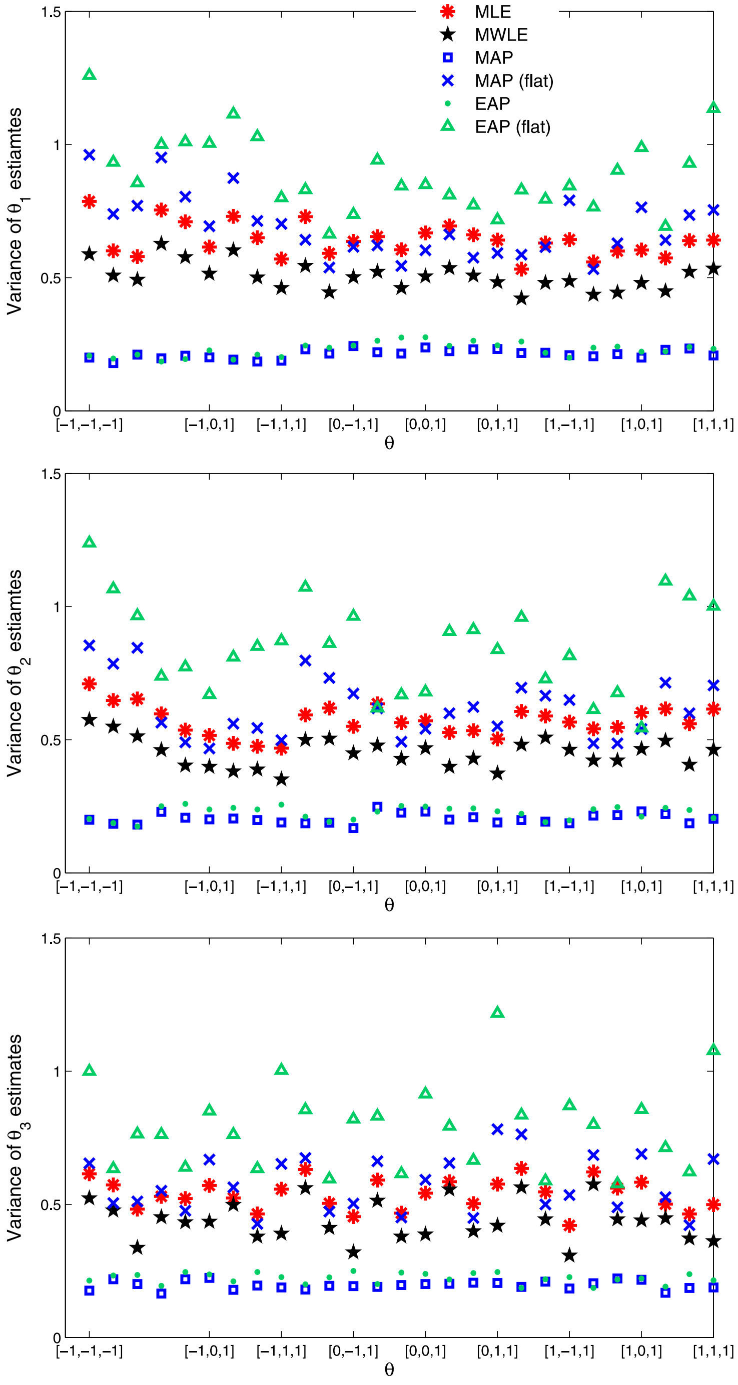

FIGURE 4.
Variance for $\theta_1$, $\theta_2$, and $\theta_3$ estimates in a three-dimensional case.

then the bias for $\hat{\theta}_2^{\text{mle}}$ becomes

$$B(\hat{\theta}_2^{\text{mle}}) = \frac{1}{2}\frac{\sum a_2^3 PQ(P-Q)}{(\sum a_2^2 PQ)^2}, \tag{17}$$

and when a two-dimensional test displays complex structure, then

$$B(\hat{\theta}_2^{\text{mle}}) = \frac{1}{2}\left[\frac{\sum a_2^3 PQ(P-Q)}{(\sum a_2^2 PQ)^2} + \frac{3\sum a_1 a_2^2 PQ(P-Q)}{(\sum a_2^2 PQ)(\sum a_1 a_2 PQ)} + \frac{3\sum a_1^2 a_2 PQ(P-Q)}{(\sum a_1 a_2 PQ)^2} + \frac{\sum a_1^3 PQ(P-Q)}{(\sum a_2^2 PQ)(\sum a_1 a_2 PQ)}\right], \tag{18}$$

with subscript $j$ omitted and $P$ or $Q$ representing $P_j(u_j = 1|\hat{\boldsymbol{\theta}}^{\text{mle}})$ and $Q_j(\hat{\boldsymbol{\theta}}^{\text{mle}}) = 1 - P_j(u_j = 1|\hat{\boldsymbol{\theta}}^{\text{mle}})$ respectively. Obviously, the bias in (18) is typically higher than that in (17).

## 4. Discussion

If a test is constructed from several presumably unidimensional scales, then one could either calibrate item parameters separately for each scale (using multiple, unidimensional IRT models) or concurrently across all scales (using multidimensional IRT model). Wang, Chen, and Cheng (2004) showed that one can more precisely estimate item parameters by considering the correlations between latent traits via MIRT models, especially for short tests. Consequently, MIRT can also provide a vector of latent trait estimates with higher precision, which is extremely useful in educational testing. For instance, evaluating students' achievement has become an increasingly important feature of public education. Rather than reporting a single sum score for each student, new generations of tests must provide feedback about what students know and have achieved to improve teaching and learning. To this end, MIRT is a commonly used psychometric model capable of providing diagnostic profile of each examinee.

If adopting the MIRT-derived ability vector, $\boldsymbol{\theta}$, as a latent profile in formative assessment, one must be confident that the estimate of $\boldsymbol{\theta}$ is close to its true value. Simulation studies have shown that in MIRT, latent trait estimates have larger measurement error (either larger bias or larger variance) when examinees have high ability on one dimension and low ability on the other dimension or have high/low abilities on both dimensions (Finkelman, Nering, & Roussos, 2009; Wang & Chang, 2011). This study proposes a general extension of Warm's WLE method to the multivariate case for reducing bias in the MLE associated with short tests. Unlike the first attempt to generalize WLE to multiple dimensions by Tseng and Hsu (2001), our approach can be used for tests exhibiting complex structure. We derived the MWLE by incorporating the well-established bias of the MLE into the score function according to Firth's (1993) preventive framework. Simulation studies showed that the MWLE estimate, $\hat{\boldsymbol{\theta}}^*$, resulted in decreased bias across various ability vectors with slightly smaller variance as compared with MLE. We also compared MLE/WMLE to Bayesian estimators and found that both EAP and MAP estimators typically had larger inward bias toward the prior mean but smaller variance, especially when responses were generated from $\boldsymbol{\theta}$'s far away from a zero vector.

Based on the improved bias and similar sampling variability of the MWLE relative to alternative ability estimators, we recommend that MWLE be implemented in commercial MIRT calibration software, such as TESTFACT or IRTPRO (Cai, Thissen, & du Toit, 2011). In addition, because the precision of the latent trait estimate can affect the accuracy of equating and linking in large-scale assessments (Tao, Shi, & Chang, 2012), or item selection in an adaptive test, future studies could compare the performance of MWLE to the alternative estimators in additional, practicable, testing conditions. Finally, licensure and classification tests also depend on accurate estimate of ability, Therefore, future researchers could determine the consequences of estimation bias in making accurate classification decisions. One limitation of the current study is that the analytical discussions are built upon the assumptions that item parameters are correctly calibrated. In the future, such discussions need to be generalized to situations in which item parameters contain certain levels of measurement errors (Zhang, Xie, Song, & Lu, 2011).

## Appendix: Derivation of the Variance of Multivariate Weighted Maximum Likelihood in Multidimensional Case

By definition, the weighted maximum likelihood maximizes the weighted likelihood as

$$\hat{\boldsymbol{\theta}}^* = \operatorname{argmax}\big[w(\boldsymbol{\theta})L(\boldsymbol{u}|\boldsymbol{\theta})\big]. \tag{A.1}$$

The following derivation was extended from, and closely parallels, the derivation in unidimensional context by Warm (1989) and Lord (1983). The four regularity conditions are

(1) $\boldsymbol{\theta}$ is bounded on a continuous scale.
(2) $P(\boldsymbol{\theta})$ is continuous and bounded away from 0 and 1 at all values of $\boldsymbol{\theta}$, $i = 1, 2, \dots, n$.
(3) At least the first five derivatives with respect to $\boldsymbol{\theta}$ of $P(\boldsymbol{\theta})$ exist at all values of $\boldsymbol{\theta}$ and are bounded.
(4) For asymptotic considerations, $n$ is considered to be incremented with replications of all of the original $n$ experiments.

Because maximizing $w(\boldsymbol{\theta})L(\boldsymbol{u}|\boldsymbol{\theta})$ is equivalent to maximizing $[w(\boldsymbol{\theta})L(\boldsymbol{u}|\boldsymbol{\theta})]^{\frac{1}{n}}$, let

$$T_1^j = \frac{\partial \ln[w(\boldsymbol{\theta})L(\boldsymbol{u}|\boldsymbol{\theta}])]^{\frac{1}{n}}}{\partial\theta_j},$$

$$T_2^{ij} = \frac{\partial^2 \ln[w(\boldsymbol{\theta})L(\boldsymbol{u}|\boldsymbol{\theta}])]^{\frac{1}{n}}}{\partial\theta_i\,\partial\theta_j},$$

$$T_3^{ijk} = \frac{\partial^3 \ln[w(\boldsymbol{\theta})L(\boldsymbol{u}|\boldsymbol{\theta}))]^{\frac{1}{n}}}{\partial\theta_i\,\partial\theta_j\theta_k}.$$

Then, by definition, we know that $T_1^j(\boldsymbol{\theta}^*) = 0$ for all $j = 1, \dots, n$. Since

$$T_1^j = \frac{1}{n}\frac{\partial \ln[w(\boldsymbol{\theta})]}{\partial\theta_j} + \frac{1}{n}\frac{\partial \ln[L(\boldsymbol{u}|\boldsymbol{\theta})]}{\partial\theta_j},$$

by letting $l_s = \frac{\partial^s \ln[L(\boldsymbol{u}|\boldsymbol{\theta})]}{\partial\boldsymbol{\theta}^s}$ and $d_s = \frac{\partial^s \ln[w(|\boldsymbol{\theta})]}{\partial\boldsymbol{\theta}^s}$, therefore $d_1^j = \frac{\partial \ln[w(\boldsymbol{\theta})]}{\partial\theta_j}$ and $l_1^j = \frac{\partial \ln[L(\boldsymbol{u}|\boldsymbol{\theta})]}{\partial\theta_j}$. Also, let $g_1^j = E(\frac{l_1^j}{n})$, $e_1^j = \frac{l_1^j}{n} - E(\frac{l_1^j}{n})$. Then,

$$T_1^j = g_1^j + e_1^j + \frac{d_1^j}{n}. \tag{A.2}$$

Let $\boldsymbol{g}_2^j = E(\frac{\boldsymbol{l}_2^j}{n})$, where $\boldsymbol{l}_2^j$ represents the $j$th column of the second derivative matrix, $\tilde{\boldsymbol{l}}_2$. Similarly, $\boldsymbol{e}_2^j = \frac{\boldsymbol{l}_2^j}{n} - \boldsymbol{g}_2^j$. Note that for ease of exposition, we use $\tilde{}$ to represent a matrix, and boldface to represent a vector. Expand $T_1^j(\boldsymbol{\theta}^*)$ at the point of $\boldsymbol{\theta}$ via Taylor series approximation so that

$$0 = T_1^j(\boldsymbol{\theta}^*) \approx T_1^j(\boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\boldsymbol{T}_2^j(\boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\tilde{\boldsymbol{T}}_3^j(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \tag{A.3}$$

$$= T_1^j(\boldsymbol{\theta}) + \sum_i x_i T_2^{ij}(\boldsymbol{\theta}) + \frac{1}{2}\sum_{i,k} x_i T_3^{ikj} x_k \tag{A.4}$$

for the $j$th component in $\boldsymbol{\theta}$. Now, substitute (A.2) into (A.4), so that

$$0 \approx g_1^j + e_1^j + \frac{d_1^j}{n} + \boldsymbol{x}\left(\boldsymbol{g}_2^j + \boldsymbol{e}_2^j + \frac{\boldsymbol{d}_2^j}{n}\right) + \frac{1}{2}\boldsymbol{x}'\left(\tilde{\boldsymbol{g}}_3^j + \tilde{\boldsymbol{e}}_3^j + \frac{\tilde{\boldsymbol{d}}_3^j}{n}\right)\boldsymbol{x},$$

where $\boldsymbol{x} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^*$. We can then write this equation as

$$-\left(e_1^j + \frac{d_1^j}{n}\right) \approx \boldsymbol{x}\left(\boldsymbol{g}_2^j + \boldsymbol{e}_2^j + \frac{\boldsymbol{d}_2^j}{n}\right) + \frac{1}{2}\boldsymbol{x}'\left(\tilde{\boldsymbol{g}}_3^j + \tilde{\boldsymbol{e}}_3^j + \frac{\tilde{\boldsymbol{d}}_3^j}{n}\right)\boldsymbol{x} \tag{A.5}$$

because $g_1^j = 0$. To obtain the variance of $\boldsymbol{\theta}^*$, square (A.5) and take expectations. Now let us explore the orders of each term. First, it can be verified that both $g_s^j$ and $d_s^j$ are of order $O(1)$ because of assumptions 2 and 3, so that $\frac{d_s^j}{n} \sim O(n^{-1})$. According to Warm (1989) and Lord (1983), $E(e_s) \sim O(n^{-\frac{1}{2}})$, $E(x^r) \sim O(n^{-\frac{r}{2}})$, and $E(x^r e_s^t) \sim O(n^{-\frac{(r+t)}{2}})$, where $r$ and $t$ here mean the $r$th and $t$th power. Therefore, after squaring both sides of (A.5), the second term in (A.5) drops out because its order is $o(n^{-1})$. Now, for notational simplicity, let

$$A_j = -\left(e_1^j + \frac{d_1^j}{n}\right), \tag{A.6}$$

$$B_{ij} = g_2^{ij} + e_2^{ij} + \frac{d_2^{ij}}{n}, \tag{A.7}$$

where $\boldsymbol{A}$ is a $K \times 1$ vector, and $\tilde{\boldsymbol{B}}$ is a $K \times K$ matrix. Then,

$$E(\boldsymbol{A}\boldsymbol{A}') = \tilde{\boldsymbol{B}}\text{var}(\hat{\boldsymbol{\theta}}^*)\tilde{\boldsymbol{B}}'. \tag{A.8}$$

Discarding the $o(n^{-1})$ terms on both sides, (i.e., the terms $E(\frac{e_1^j d_1^j}{n})$ and $E(\frac{d_1^j}{n})^2$ on the left side, and $E(e_2^{ij^2} x_i x_j)$, $E(e_2^{ij} x_i x_j \frac{d_2^{ij}}{n})$, and $E((\frac{d_2^{ij}}{n})^2 x_i x_j)$ on the right side), we finally have

$$\text{var}(\hat{\boldsymbol{\theta}}^*) \approx \boldsymbol{g}_2^{-1} E(\boldsymbol{e}_1\boldsymbol{e}_1')\boldsymbol{g}_2^{-1} = [\boldsymbol{I}(\hat{\boldsymbol{\theta}}^*)]^{-1}.$$

## References

Ackerman, T., Gierl, M.J., & Walker, C.M. (2003). Using multidimensional item response theory to evaluate educational psychological tests. *Educational Measurement: Issues and Practices*, *22*, 37–51.

Anderson, J.A., & Richardson, S.C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, *21*, 71–78.

Bock, R.D., Gibbons, R., Schilling, S.G., Muraki, E., Wilson, D.T., & Wood, R. (2003). *TESTFACT 4.0*. Lincolnwood: Scientific Software International. [Computer software and manual].

Cai, L. (2008). *A Metropolis–Hastings Robbins–Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model*. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill, NC.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*, 33–57.

Cai, L. (2010b). Metropolis–Hastings Robbins–Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307–335.

Cai, L., Thissen, D., & du Toit, S.H.C. (2011). *IRTPRO: flexible, multidimensional, multiple categorical IRT modeling [Computer software]*. Lincoln wood: Scientific Software International.

Chalmers, R.P. (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*. www.jstatsoft.org.

Eignor, D. R., & Schaeffer, G. A. (1995). *Comparability studies for the GRE General CAT and the NCLEX using CAT*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April.

Finkelman, M., Nering, M.L., & Roussos, L.A. (2009). A conditional exposure method for multidimensional adaptive testing. *Journal of Educational Measurement*, *46*, 84–103.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*, 27–38.

Fraser, C. (1998). *NOHARM: a Fortran program for fitting unidimensional and multidimensional normal ogive models in latent trait theory*. The University of New England, Center for Behavioral Studies, Armidale, Australia.

Hattie, J. (1981). *Decision criteria for determining unidimensionality*. Unpublished doctoral dissertation, University of Toronto, Canada.

Kim, J.K., & Nicewander, W.A. (1993). Ability estimation for conventional tests. *Psychometrika*, *58*, 587–599.

Lee, P. (1989). *Bayesian statistics: an introduction*. London: Edward Arnold.

Lehmann, E.L., & Casella, G. (1998). *Theory of point estimation*. New York: Springer.

Lord, F.M. (1983). Unbiased estimation of ability parameters, of their variance and of their parallel forms reliability. *Psychometrika*, *48*, 223–245.

Lord, F.M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *2*, 157–162.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Mulder, J., & van der Linden, W.J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*, 273–296.

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

Samejima, F. (1993). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika*, *58*, 119–138.

Schaefer, R.L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine*, *2*, 71–78.

Segall, D.O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354.

Serfling, R.J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.

Stroud, A.H., & Sechrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliffs: Prentice-Hall.

Tao, J., Shi, N., & Chang, H. (2012). Item-weighted likelihood method for ability estimation in tests composed of both dichotomous and polytomous items. *Journal of Educational and Behavioral Statistics*, *37*, 298–315.

Tseng, F.L., & Hsu, T.C. (2001). *Multidimensional adaptive testing using the weighted likelihood estimation: a comparison of estimation methods*. Paper presented at the annual meeting of Seattle, WA.

van der Linden, W.J. (1999a). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, *24*, 398–412.

van der Linden, W.J. (1999b). A procedure for empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement*, *23*, 21–29.

van der Linden, W.J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W.J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20.

Veldkamp, B.P., & van der Linden, W.J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575–588.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Wang, C., & Chang, H. (2011). Item selection in multidimensional computerized adaptive tests—gaining information from different angles. *Psychometrika*, *76*, 363–384.

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement*, *25*, 317–331.

Wang, T., Hanson, B.A., & Lau, C.-M.A. (1999). Reducing bias in CAT trait estimation: a comparison of approaches. *Applied Psychological Measurement*, *23*, 263–278.

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.

Wang, C., Chang, H., & Boughton, K. (2011). Kullback–Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika*, *76*, 13–39.

Zhang, J., Xie, M., Song, X., & Lu, T. (2011). Investigating the impact of uncertainty about item parameters on ability estimation. *Psychometrika*, *76*, 97–118.

*Manuscript Received: 2 JUN 2013*
*Published Online Date: 7 MAR 2014*