Educational and Psychological Measurement
72(2) 264–290

DOI: 10.1177/0013164411410056
http://epm.sagepub.com

Characterizing Sources of Uncertainty in Item Response Theory Scale Scores

Ji Seung Yang[1], Mark Hansen[1], and Li Cai[1]

Abstract

Traditional estimators of item response theory scale scores ignore uncertainty carried over from the item calibration process, which can lead to incorrect estimates of the standard errors of measurement (SEMs). Here, the authors review a variety of approaches that have been applied to this problem and compare them on the basis of their statistical methods and goals. They then elaborate on the particular flexibility and usefulness of a multiple imputation–based approach, which can be easily applied to tests with mixed item types and multiple underlying dimensions. This proposed method obtains corrected estimates of individual scale scores, as well as their SEMs. Furthermore, this approach enables a more complete characterization of the impact of parameter uncertainty by generating confidence envelopes (intervals) for item trace lines, test information functions, conditional SEM curves, and the marginal reliability coefficient. The multiple imputation–based approach is illustrated through the analysis of an artificial data set, then applied to data from a large educational assessment. A simulation study was also conducted to examine the relative contribution of item parameter uncertainty to the variability in score estimates under various conditions. The authors found that the impact of item parameter uncertainty is generally quite small, though there are some conditions under which the uncertainty carried over from item calibration contributes substantially to variability in the scores. This may be the case when the calibration sample is small relative to the number of item parameters to be estimated or when the item response theory model fit to the data is multidimensional.

Keywords

IRT, IRT scoring, calibration, standard error of measurement, multiple imputation, confidence envelope

[1]University of California, Los Angeles, Los Angeles, CA, USA

Corresponding Author:
Li Cai, Education Department, University of California, Los Angeles, 2022A Moore Hall, 405 Hilgard Avenue, Los Angeles, CA 90095-1521, USA
Email: lcai@ucla.edu

Introduction

In applications of item response theory (IRT), it is common to score individuals using maximum marginal likelihood (MML) estimates of item parameters, which are ideally obtained using a large and independent calibration sample. In this process, the standard errors of measurement (SEMs) are estimated as either the reciprocal of the square root of the Fisher information function evaluated at the posterior mode for maximum a posteriori (MAP) scoring (or maximum likelihood scoring if there is no prior) or the standard deviation of the posterior under expected a posteriori (EAP) scoring. The use of MML estimates in conjunction with MAP or EAP scoring is a form of empirical Bayes (EB; Carlin & Louis, 2000).

However, both the scale score and its standard error—the latter in particular—are affected by the uncertainty in the item parameter estimates, which is carried over from the calibration process (Tsutakawa & Johnson, 1990). For frequentists, the item parameter estimates differ from true parameter values because of sampling error, which is represented in the standard errors of the parameter estimates. From a likelihood perspective, the uncertainty is reflected by the curvature of the log-likelihood, often represented as the information matrix of the item parameters. For Bayesians, point estimates of item parameters alone do not convey information about the width of the posterior distributions, which is nonnegligible unless the calibration sample size tends to infinity so that the posteriors become peaked. From any statistical point of view, then, uncertainty in the item parameter estimates is concerning.

Unfortunately, traditional scoring approaches (e.g., MAP or EAP) fail to acknowledge this uncertainty. Chief among the problems is an incorrect statement of measurement error. The scale score's SEM is often underestimated and the scale's marginal reliability overestimated, which means that the scores may be taken as having greater precision than is justified. As Cheng and Yuan (2010) pointed out, an understated SEM can also lead to premature termination of a computerized adaptive test. The problem is most pronounced in situations where a small calibration sample is used or when item parameters are estimated using the sample to be scored (Tsutakawa & Johnson, 1990).

Researchers have proposed a number of approaches to address this problem and related issues (e.g., Cheng & Yuan, 2010; Embretson, 1999; Hoshino & Shigemasu, 2008; Lewis, 1985, 2001; Mislevy, Wingersky, & Sheehan, 1993; Mislevy & Yan, 1991; Patz & Junker, 1999; Thissen & Wainer, 1990; Tsutakawa & Johnson, 1990; Tsutakawa & Soltys, 1988; Zhang, Xie, Song, & Lu, 2011). In the research reported here, we first provide a brief review of existing approaches, comparing them with respect to both their underlying methods and intended objectives. In broad terms, three types of methods have been applied to this problem: analytic approximations, fully Bayesian sampling based approaches, and multiple imputation (MI)–based approaches. These approaches have been used to accomplish two related but distinct goals: (a) to obtain corrected SEMs that take uncertainty in the item parameters into account and (b) to characterize the nature and impact of item parameter uncertainty on subsequent estimation and inference.

After reviewing these approaches, we provide a formal justification for the MI-based strategy and illustrate its use with a three-item artificial data set. Extending Mislevy et al.'s (1993) results, we use MI to obtain corrected estimates of individual scale scores and their standard errors of measurement. Using MI, we also characterize the specific ways in which uncertainty in the item parameters can affect scoring. Thissen and Wainer (1990) proposed that it may be helpful to visualize the uncertainty by generating confidence envelopes for item characteristic curves. We carry this idea further by constructing confidence envelopes for test information functions and conditional SEM curves. Such depictions make it possible to observe the ways in which the effects of parameter uncertainty vary across the values of the latent trait. The MI-based approach may also be used to obtain confidence intervals for the marginal reliability coefficient. After illustrating the approach with the artificial data set, we analyze data from a large educational assessment. Finally, through simulation studies, we examine the ways in which model complexity, calibration sample size, and test length contribute to these effects.

Our approach improves on the existing alternatives in several important ways:

- First, in contrast to analytical approximation methods (e.g., Cheng & Yuan, 2010; Zhang et al., 2011) that explicitly require the calculation of a number of non–standard derivative matrices that are model specific, the MI-based approach is easily applied to any IRT model (e.g., unidimensional or multidimensional, dichotomous or polytomous) and scoring method (e.g., MAP, EAP, or even summed score EAP), provided that the asymptotic covariance matrix of the item parameter estimates is available. This flexibility is an important feature, as we note that previous studies have not considered item parameter uncertainty in the context of polytomous or multidimensional IRT models. Here, we illustrate the proposed approach for tests of mixed-item types (including two-parameter logistic [2PL] and three-parameter logistic [3PL] models for dichotomous data and a logistic graded-response model for ordered responses) analyzed using either a unidimensional or a multidimensional model, specifically the item bifactor model (e.g., Cai, Yang, & Hansen, in press; Gibbons & Hedeker, 1992).
- Second, in contrast to fully Bayesian methods (e.g., Patz & Junker, 1999) that must rely on Markov chain Monte Carlo (MCMC) to sample the intractable posterior distributions, the MI approach is computationally simple and efficient. It relies on information that is routinely printed in the output of most standard IRT software programs, in conjunction with the easily accomplished task of random sampling from the multivariate normal distribution.
- Third, in contrast to previous MI-based methods, in which analyses were either limited to single items or the between-item parameter error covariances were not estimated (and, thus, treated as zero), our approach makes use of a modern estimation algorithm (supplemented EM; Cai, 2008) for computing the asymptotic covariance matrix of the item parameters that is applicable to any IRT model. This covariance matrix is not an automatic by-product of the

current gold standard Bock and Aitkin (1981) EM algorithm of item parameter estimation.
- Finally, this approach systematizes a number of seemingly disparate methods under a single framework, including Thissen and Wainer's (1990) confidence envelopes and Lewis's (1985) expected response functions. Even approximation methods based on pseudo maximum likelihood (e.g., Hoshino & Shigemasu, 2008) can be reinterpreted in light of the MI framework.

A Brief Review of Existing Approaches

Two-Stage Versus Single-Stage

With the exception of fully Bayesian sampling–based methods, nearly all existing proposals assume a two-stage process for estimating the IRT scale scores. The items are first calibrated, preferably using MML or Bayesian methods. In this stage, the item parameters are estimated, as well as their error covariance matrix. In the second stage, the item parameters are used to produce the IRT scale scores. Corrections are generally made in the second stage, using (a) the point estimates, (b) the error covariance matrix, and (c) either additional derivatives and linearization arguments for analytic approximations or, in the case of MI methods, random sampling from an approximation to the posterior distributions of the item parameters. Two-stage methods are popular not only because they generally lead to sound estimates but also because they are consistent with the standard operating procedures in applied educational and psychological testing situations.

Fully Bayesian sampling–based methods (e.g., Patz & Junker, 1999), on the other hand, involve a single stage. They rely on MCMC to produce random draws from a Markov chain having the full joint posterior of the item parameters and the individual latent variables as its invariant distribution. Under the ergodicity of the Markov chain, dependent samples from the chain can be used to approximate the full posterior. When inferences regarding individual latent traits are desired, one simply ''marginalizes'' over the other dimensions of the posterior—that is, by integrating out the item parameters—and the IRT scale scores thus obtained (e.g., as posterior means) would already have taken the uncertainty in item parameters into account. Given the MCMC output, marginalization amounts to ignoring that part of the MCMC output related to the item parameters. The single-stage approach is appealing conceptually. However, a significant barrier to its widespread adoption is its complexity, computational intensiveness, and, in some cases, inflexibility. Even as MCMC gains popularity, its proper usage still requires considerably more effort and statistical expertise when compared with more deterministic and better understood methods such as the EM algorithm. As Edwards (2005) noted, a number of competing MCMC samplers have been proposed for IRT, and the question of their relative algorithmic efficiency has not been entirely settled in the methodological literature. Furthermore, we point out that in contrast to two-stage IRT scoring methods, it is difficult to conceive how the single-stage

approach can easily accommodate certain nonstandard but nevertheless popular and useful IRT scoring algorithms such as summed-score to EAP translations (Thissen & Wainer, 2001).

Two Related Goals: Correction Versus Characterization

Most existing approaches have sought solely to obtain corrected estimates of SEM for IRT scale scores that take into account the uncertainty in the item parameters. These include analytic approximations based on Bayesian calculations (Tsutakawa & Johnson, 1990; Tsutakawa & Soltys, 1988), analytic approximations based on pseudo maximum likelihood (Cheng & Yuan, 2010; Hoshino & Shigemasu, 2008), as well as MI-based expected response functions (Lewis, 1985, 2001; Mislevy et al., 1993). MCMC methods (e.g., Patz & Junker, 1999) may also be regarded as belonging to this category.

For analytic approximations (Bayesian or likelihood), the asymptotic argument is typically based on Taylor series expansions of the nonlinear estimating equations of the IRT scale scores. From the first-order (sometimes second-order) approximation, corrected standard error formulae are obtained. In contrast, the expected response functions are motivated by an explicit analogy to MI for treating missing data (Rubin, 1987). Using the point estimates of item parameters to represent the item response functions amounts to a single round of imputation, replacing the unknown (hence ''missing'') item parameter values by the modes of their marginal log-likelihood. In contrast, an MI approach uses more than one random imputation to recover the uncertainty caused by not knowing the item parameters exactly. Averaged response functions under multiple random imputations are expected response functions.

Thissen and Wainer (1990) demonstrated an approach with an entirely different goal. Rather than seeking corrected estimates, they used confidence envelopes around the item response functions to reflect the uncertainty in the item parameters. This approach can be viewed as complimentary to the expected response function approach of Mislevy et al. (1993). Unlike Mislevy et al., however, the focus of Thissen and Wainer (1990) is not on obtaining corrected SEM for the scale scores. The resulting graphical displays (referred to as the *M*-line plots) are simply used to characterized how errors in the item parameter estimates affect the plausible shapes of item characteristic curves. The confluence of Thissen and Wainer's (1990) confidence envelopes and Mislevy et al.'s (1993) expected response functions leads to the dominating insight of this research. That is, MI provides a natural framework to integrate the two goals, simultaneously providing corrected SEMs and visual representations of the uncertainty.

The Proposed Approach

Some Notation

Without loss of generality, let $\mathbf{y}_i$ be a $J \times 1$ vector of observed item responses for an individual randomly sampled from an ability distribution with density $f(\theta)$, where θ is

the latent trait. Suppose the item parameters for the J items are contained in $\boldsymbol{\gamma}$. The IRT model postulates a conditional distribution for $\mathbf{y}$ on θ and $\boldsymbol{\gamma}$: $f(\mathbf{y}|\theta, \boldsymbol{\gamma})$. We assume that θ and $\boldsymbol{\gamma}$ are a priori independent. Integrating out the incidental parameter θ, we have $f(\mathbf{y}|\boldsymbol{\gamma}) = \int f(\mathbf{y}|\theta, \boldsymbol{\gamma}) f(\theta) d\theta$. For a sample of N independent respondents, the distribution of response patterns is $f(\mathbf{Y}|\boldsymbol{\gamma}) = \prod_{i=1}^{N} f(\mathbf{y}_i|\boldsymbol{\gamma})$, where $\mathbf{Y}$ is an $N \times J$ matrix of item responses. Given prior distribution $\pi(\boldsymbol{\gamma})$, the posterior of the item parameters is

$$f(\boldsymbol{\gamma}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})}{\int f(\mathbf{Y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma}) d\boldsymbol{\gamma}}.$$

If the prior is taken to be uniform, then the posterior is proportional to the marginal log-likelihood $\log L(\boldsymbol{\gamma}|\mathbf{Y})$ based on an observed matrix of item responses $\mathbf{Y}$ (the calibration sample). Numerical optimization of the log-likelihood leads to the MML estimator $\hat{\boldsymbol{\gamma}}_N$, and Bock and Aitkin's (1981) EM algorithm is often used. The log-likelihood curvature around the mode is usually presented as the asymptotic covariance matrix $\mathcal{I}_N^{-1}$, where $\mathcal{I}_N$ is the information matrix based on a calibration sample of size N. Recently Cai (2008) proposed the use of a supplemented EM algorithm (Meng & Rubin, 1991) to compute $\mathcal{I}_N$ for IRT models.

Because of the Bernstein–von Mises phenomenon (see, e.g., van der Vaart, 1998), it is well known that a multivariate normal with mean $\hat{\boldsymbol{\gamma}}_N$ and a covariance matrix equal to $\mathcal{I}_N^{-1}$ provides a reasonable approximation to the posterior $f(\boldsymbol{\gamma}|\mathbf{Y})$. With an abuse of notation, we may write

$$f(\boldsymbol{\gamma}|\mathbf{Y}) \overset{a}{=} \phi(\boldsymbol{\gamma}; \hat{\boldsymbol{\gamma}}_N, \mathcal{I}_N^{-1}), \tag{1}$$

where $\phi(\cdot; \hat{\boldsymbol{\gamma}}_N, \mathcal{I}_N^{-1})$ represents a normal distribution with mean $\hat{\boldsymbol{\gamma}}_N$ and covariance matrix $\mathcal{I}_N^{-1}$, and $\overset{a}{=}$ indicates asymptotic equivalence (in N). In other words, the (analytically intractable) true posterior $f(\boldsymbol{\gamma}|\mathbf{Y})$ becomes the gold standard here, and $\phi(\cdot; \hat{\boldsymbol{\gamma}}_N, \mathcal{I}_N^{-1})$ provides a convenient approximation. If $\phi(\cdot; \hat{\boldsymbol{\gamma}}_N, \mathcal{I}_N^{-1})$ provides a good enough approximation to $f(\boldsymbol{\gamma}|\mathbf{Y})$, estimands (e.g., IRT scores, information, and SEMs) that are based on $f(\boldsymbol{\gamma}|\mathbf{Y})$ should also be approximated well.

Illustration

To illustrate the proposed MI-based procedure, we created a simple data set consisting of 500 simulated responses to three items. Each of the items was of a different type: 2PL, 3PL, and three-category graded response (GRM3). Although this illustration is somewhat unrealistic, the very short test length offers some advantages. Most important, it allows us to easily present the generating and estimated values of all eight item parameters, their error covariance matrix, and examples of the randomly imputed parameter sets. In addition, the numbers of possible response patterns (12) and summed scores (5) for this three-item test are small, allowing us to demonstrate the MI-based scoring for each possibility. Finally, despite the simplicity of this example, it nonetheless features two kinds of complexity not previously considered in studies

Table 1. Item Parameter Estimates and the Asymptotic Covariance Matrix for the Three-Item Artificial Data Set

				Parameter error covariance matrix ($\mathcal{I}_N^{-1}$)							
				Item 1 (2PL)		Item 2 (3PL)			Item 3 (GRM3)		
Item	Parameter	True value	$\hat{\gamma}_N$ (*SE*)	*c*	*a*	logit(*g*)	*c*	*a*	c_0	c_1	*a*
1 (2PL)	Intercept, *c*	−0.54	−0.50 (0.11)	0.01							
	Slope, *a*	0.62	0.67 (0.29)	−0.01	0.08						
2 (3PL)	Guessing, logit(*g*)	−1.39	−1.41 (0.50)	0.00	0.00	0.25					
	Intercept, *c*	1.34	1.35 (0.24)	0.00	0.00	−0.07	0.06				
	Slope, *a*	0.88	0.90 (0.40)	0.00	0.00	0.02	0.05	0.16			
3 (GRM3)	Intercept, c_0	0.66	0.59 (0.13)	0.00	−0.01	0.00	0.00	−0.01	0.02		
	Intercept, c_1	−0.54	−0.39 (0.12)	0.00	0.01	0.00	0.00	0.01	0.00	0.01	
	Slope, *a*	1.01	0.91 (0.40)	0.01	−0.05	0.00	−0.03	−0.07	0.03	−0.02	.16

Note. 2PL = two-parameter logistic model; 3PL = three-parameter logistic model; GRM3 = three-category graded response model.

addressing parameter uncertainty: polytomous items (as represented by Item 3) and tests of mixed-item types. These features will also be present in the real data set considered later.

The generating item slopes and intercepts for this illustration were randomly drawn from distributions chosen to resemble values commonly observed in educational or psychological testing. Both data generation and parameter estimation were conducted using IRTPRO (Cai, du Toit, & Thissen, in press).

Table 1 shows the generating values, along with the MML estimates and their error covariance matrix. The standard errors of the item parameter estimates are the square roots of the diagonal elements of this matrix. The off-diagonal elements of the matrix represent covariances between the parameters. The fact that some of these elements are nonzero is notable, as some prior methods seeking to account for parameter uncertainty have ignored these covariances. In the case of the 3PL model (used for Item 2), the guessing parameter is expressed as the logit of the guessing probability *g*; the generating value of −1.39 corresponds to a probability of about .2, as might be expected for a multiple choice item with five options. The MML parameter estimates and the error covariance matrix will be used in the MI-based approach, as we describe in the following sections.

Multiple Imputation Inference for EAP Scores

Preliminaries. We now consider EAP scoring under item parameter uncertainty. The logic developed in this section, however, is quite general. Though we will assume a unidimensional θ for simplicity of notation, we note that the methods apply directly to multidimensional IRT models, where θ is a vector. In the ideal case where item parameters are known, inference for θ for an individual with $J \times 1$ response pattern $\mathbf{x}$ should be based on the following posterior:

$$f(\theta|\mathbf{x},\boldsymbol{\gamma}) = \frac{f(\mathbf{x}|\theta,\boldsymbol{\gamma})f(\theta)}{\int f(\mathbf{x}|\theta,\boldsymbol{\gamma})f(\theta)d\theta} = \frac{f(\mathbf{x}|\theta,\boldsymbol{\gamma})f(\theta)}{f(\mathbf{x}|\boldsymbol{\gamma})}. \quad (2)$$

The EAP estimator with given $\boldsymbol{\gamma}$ is an expectation over the posterior distribution in Equation (2)

$$\hat{\theta}_{\boldsymbol{\gamma}} = \int \theta f(\theta|\mathbf{x},\boldsymbol{\gamma})d\theta = \frac{1}{f(\mathbf{x}|\boldsymbol{\gamma})}\int \theta f(\mathbf{x}|\theta,\boldsymbol{\gamma})f(\theta)d\theta, \quad (3)$$

with SEM given by the square root of posterior variance

$$V(\hat{\theta}_{\boldsymbol{\gamma}}) = \int (\theta - \hat{\theta}_{\boldsymbol{\gamma}})^2 f(\theta|\mathbf{x},\boldsymbol{\gamma})d\theta = \frac{1}{f(\mathbf{x}|\boldsymbol{\gamma})}\int \theta^2 f(\mathbf{x}|\theta,\boldsymbol{\gamma})f(\theta)d\theta - \hat{\theta}^2_{\boldsymbol{\gamma}}. \quad (4)$$

If the item parameters are unknown, standard practice is to use the ''plug-in'' estimator, in which the MML estimates $\hat{\boldsymbol{\gamma}}_N$ are used in place of $\boldsymbol{\gamma}$. Notice that unless N tends to infinity, the traditional estimator

$$\hat{\theta}_{\hat{\boldsymbol{\gamma}}_N} = \int \theta f(\theta|\mathbf{x},\hat{\boldsymbol{\gamma}}_N)d\theta$$

ignores the inherent variability of $\hat{\boldsymbol{\gamma}}_N$ as reflected by $\mathcal{I}_N^{-1}$. The traditional SEM estimator $V(\hat{\theta}_{\hat{\boldsymbol{\gamma}}_N})$ also ignores this variability.

A formal justification for the proposed multiple imputation procedure. Tsutakawa and Johnson (1990) demonstrated that to properly account for the uncertainty in $\hat{\boldsymbol{\gamma}}_N$, one must base the inference for θ on the posterior distribution of θ given $\mathbf{x}$ and $\mathbf{Y}$, which they represented as (their Equation 14, presented here with slight notational change)

$$f(\theta|\mathbf{x},\mathbf{Y}) = \frac{f(\theta)}{f(\mathbf{x}|\mathbf{Y})}\int f(\mathbf{x}|\theta,\boldsymbol{\gamma})f(\boldsymbol{\gamma}|\mathbf{Y})d\boldsymbol{\gamma}. \quad (5)$$

Note that a critical feature of Equation (5) is that the posterior of the item parameters $f(\boldsymbol{\gamma}|\mathbf{Y})$ from calibration now serves as the prior. It can be shown (see the appendix) that the EAP estimator of θ can be represented as

$$\hat{\theta} = \int \hat{\theta}_{\boldsymbol{\gamma}} f(\boldsymbol{\gamma}|\mathbf{x},\mathbf{Y})d\boldsymbol{\gamma}. \quad (6)$$

This estimator does not depend on any particular values of $\hat{\boldsymbol{\gamma}}$ because it averaged over all plausible values of $\boldsymbol{\gamma}$. The posterior variance (see the appendix)

$$V(\hat{\theta}) = \int V(\hat{\theta}_{\boldsymbol{\gamma}})f(\boldsymbol{\gamma}|\mathbf{x},\mathbf{Y})d\boldsymbol{\gamma} + \int (\hat{\theta}_{\boldsymbol{\gamma}} - \hat{\theta})^2 f(\boldsymbol{\gamma}|\mathbf{x},\mathbf{Y})d\boldsymbol{\gamma} \quad (7)$$

also automatically takes the uncertainty in $\boldsymbol{\gamma}$ into account. Equation (6) shows that $\hat{\theta}$ is an expectation of $\hat{\theta}_{\boldsymbol{\gamma}}$ with respect to $f(\boldsymbol{\gamma}|\mathbf{x},\mathbf{Y})$. Equation (7) reveals a familiar variance

decomposition. The posterior variance is equal to the sum of two components, the expectation of a variance and the variance of an expectation. The first component may be conceived of as the ''within'' variance, and the second is the ''between'' variance.

It turns out that $f(\boldsymbol{\gamma}|\mathbf{x},\mathbf{Y})$ can be treated as $f(\boldsymbol{\gamma}|\mathbf{Y})$ for a number of reasons. If $\mathbf{x}$ is actually a component of $\mathbf{Y}$ (i.e., calibration and scoring for the same sample), then the appropriate notation should replace $\mathbf{Y}$ with $\mathbf{Y}_{(\mathbf{x})}$, where $\mathbf{Y}_{(\mathbf{x})}$ denotes the response pattern matrix without observation $\mathbf{x}$. However, $f(\boldsymbol{\gamma}|\mathbf{x},\mathbf{Y}_{(\mathbf{x})})$ coincides with $f(\boldsymbol{\gamma}|\mathbf{Y})$. If $\mathbf{x}$ is an independent observation (i.e., the scoring sample), IRT scoring requires that $f(\boldsymbol{\gamma}|\mathbf{x},\mathbf{Y})$ be the same as $f(\boldsymbol{\gamma}|\mathbf{Y})$, as we do not wish to change the scoring criterion once the items have been calibrated. Furthermore, because of Equation (1), we can use a normal approximation of $f(\boldsymbol{\gamma}|\mathbf{Y})$ as $\phi(\boldsymbol{\gamma};\hat{\boldsymbol{\gamma}}_N,\mathcal{I}_N^{-1})$. Consequently, we may replace Equations (6) and (7) with the following approximations:

$$\hat{\theta} \stackrel{a}{=} \int \hat{\theta}_{\boldsymbol{\gamma}}\phi(\boldsymbol{\gamma};\hat{\boldsymbol{\gamma}}_N,\mathcal{I}_N^{-1})d\boldsymbol{\gamma}, \tag{8}$$

$$V(\hat{\theta}) \stackrel{a}{=} \int V(\hat{\theta}_{\boldsymbol{\gamma}})\phi(\boldsymbol{\gamma};\hat{\boldsymbol{\gamma}}_N,\mathcal{I}_N^{-1})d\boldsymbol{\gamma} + \int (\hat{\theta}_{\boldsymbol{\gamma}}-\hat{\theta})^2\phi(\boldsymbol{\gamma};\hat{\boldsymbol{\gamma}}_N,\mathcal{I}_N^{-1})d\boldsymbol{\gamma}. \tag{9}$$

Now, we have an MI algorithm to approximate $\hat{\theta}$ and $V(\hat{\theta})$:

1. Draw $M > 1$ sets of values from a multivariate normal distribution with mean $\hat{\boldsymbol{\gamma}}_N$ and covariance matrix $\mathcal{I}_N^{-1}$. Denote them as $\boldsymbol{\gamma}_j$ for $j = 1, \ldots, M$.
2. Plug each $\boldsymbol{\gamma}_j$ into Equations (3) and (4) and compute $\hat{\theta}_{\boldsymbol{\gamma}_j}$ as well as $V(\hat{\theta}_{\boldsymbol{\gamma}_j})$. Write $\hat{\theta}_j$ as shorthand for $\hat{\theta}_{\boldsymbol{\gamma}_j}$ and V_j for $V(\hat{\theta}_{\boldsymbol{\gamma}_j})$.
3. The MI EAP approximation to $\hat{\theta}$ is the empirical average,

$$\bar{\theta} \simeq \frac{1}{M}\sum_{j=1}^{M}\hat{\theta}_j, \tag{10}$$

which approximates the right-hand side expression of Equation (8).

4. The MI variance approximation is

$$V(\bar{\theta}) \simeq \bar{V} + (1+M^{-1})B, \tag{11}$$

where $\bar{V} = M^{-1}\sum_{j=1}^{M} V_j$ is an estimate of the ''within''-imputation variance (expectation of a variance), and $B = (M-1)^{-1}\sum_{j=1}^{M}(\hat{\theta}_j - \bar{\theta})^2$ is an estimate of the ''between''-imputation variance (variance of an expectation). These correspond to the two parts on the right-hand side of Equation (9).

The ratio

$$r = \frac{(1+M^{-1})B}{\bar{V}} \tag{12}$$

Table 2. MML Parameter Estimates and Random Imputations for the Three-Item Artificial Data Set

			Imputed parameters ($M = 20$)				
Item	Parameter	γ_N (*SE*)	γ_1	γ_2	γ_3	...	γ_{20}
1 (2PL)	Intercept, *c*	−0.50 (0.11)	−0.50	−0.68	−0.60	...	−0.38
	Slope, *a*	0.67 (.29)	0.61	0.75	0.71	...	0.32
2 (3PL)	Guessing, logit(*g*)	−1.41 (0.50)	−2.09	−0.86	−0.95	...	−1.84
	Intercept, *c*	1.35 (0.24)	1.41	1.33	1.31	...	1.00
	Slope, *a*	0.90 (0.40)	0.76	1.10	0.91	...	0.35
3 (GRM3)	Intercept, c_0	0.59 (0.13)	0.64	0.67	0.32	...	0.79
	Intercept, c_1	−0.39 (0.12)	−0.50	−0.26	−0.56	...	−0.51
	Slope, *a*	0.91 (0.40)	1.16	0.81	0.23	...	1.36

Note. MML = maximum marginal likelihood; 2PL = two-parameter logistic model; 3PL = three-parameter logistic model; GRM3 = three-category graded response model.

is known as the *relative increase in variance* due to missing data (Schafer, 1997). It can be taken as a crude measure of the impact of item parameter uncertainty. We will use this measure in the simulation studies we describe subsequently.

As previously mentioned, the MI-based approach may be applied to any number of other scoring methods. Thus, in our analyses, we present results not only for pattern EAP scoring but also for summed score to EAP translations (Lord & Wingersky, 1984). As will be demonstrated, the impact of item parameter uncertainty may differ according to scoring method.

Illustration for the proposed procedure. The MI-based procedure was applied to the three-item data set, following the steps described above.

1. We drew $M = 20$ sets of plausible item parameter values from a multivariate normal distribution with mean $\hat{\boldsymbol{\gamma}}_N$ and covariance matrix $\mathcal{I}_N^{-1}$, both given in Table 1. We denote the plausible parameter values as $\boldsymbol{\gamma}_j$ for $j = 1, \ldots, 20$. The imputed sets of $\boldsymbol{\gamma}_j$ are presented in Table 2.
2. We now plug each $\boldsymbol{\gamma}_j$ into Equations (3) and (4). We obtain a different posterior distribution for each vector $\boldsymbol{\gamma}_j$. This is illustrated in Figure 1, in which the prior distribution, item trace lines, and posterior distributions are shown for three response patterns. The posteriors have different means and standard deviations.
3. The MI-based approximation to $\hat{\theta}$ is calculated by taking the average of $\hat{\theta}_{\boldsymbol{\gamma}_j}$ across all imputations.
4. The MI-based variance approximation $V(\bar{\theta})$ is calculated using Equation (11), and the relative increase in variance r is calculated using Equation (12). Table 3 shows $\bar{V}$, B, $V(\bar{\theta})$, and r for both full-pattern EAP scoring and the summed-score EAP translations. For the full-pattern EAP, the relative increase in variance r ranged from 0.2% to 10.8%. For the summed-score

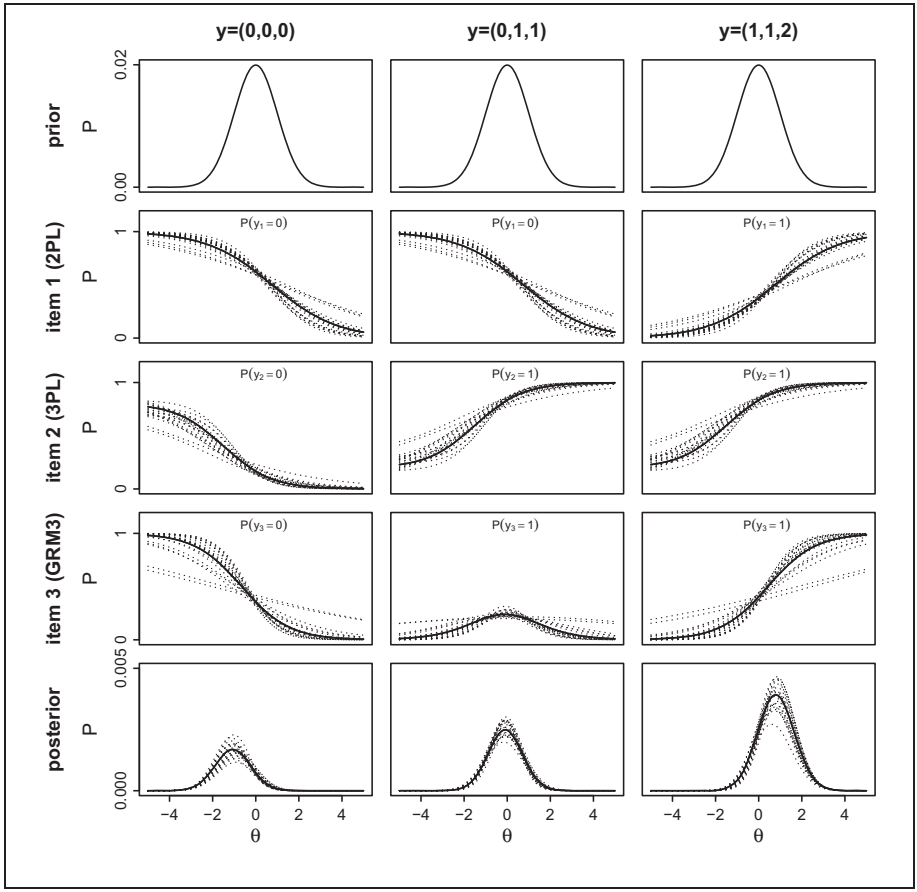

Figure 1. Item trace lines and posterior distributions for three response patterns
Note. Solid lines represent the trace lines and posteriors based on the maximum marginal likelihood parameter estimates; those based on M = 20 randomly imputed parameter sets appear as dotted lines.

EAP translations, the observed values of *r* had a much more narrow range, from 1.4% and 1.6%.

The scores obtained using the MML parameter estimates are compared with MI-based scores in Figure 2. These two sets of scores are almost perfectly correlated, for both the response pattern EAPs and summed-score EAPs. This result is consistent with previous studies; although item parameter uncertainty may result in overestimates of score precision, it does not necessarily result in biased scores.

Table 3. EAP Scores Based on Imputed Item Parameters

Response pattern			Estimates based on imputed parameters ($M = 20$)					MI summary statistics			
y_1	y_2	y_3	$\hat{\theta}_{y_1}$ (*SE*)	$\hat{\theta}_{y_2}$ (*SE*)	$\hat{\theta}_{y_3}$ (*SE*)	...	$\hat{\theta}_{y_{20}}$ (*SE*)	$\bar{\theta}$	$\bar{V}$	*B*	$V(\bar{\theta})$
0	0	0	−1.07 (.83)	−1.10 (.82)	−0.86 (.89)	...	−0.90 (.85)	−1.03	0.69	0.01	0.70
0	0	1	−0.51 (.77)	−0.72 (.78)	−0.74 (.88)	...	−0.26 (.76)	−0.58	0.61	0.02	0.64
0	1	0	−0.58 (.83)	−0.47 (.86)	−0.21 (.92)	...	−0.67 (.84)	−0.51	0.72	0.02	0.74
1	0	0	−0.65 (.83)	−0.59 (.82)	−0.30 (.89)	...	−0.68 (.84)	−0.55	0.68	0.03	0.71
0	0	2	−0.10 (.83)	−0.44 (.82)	−0.64 (.89)	...	0.30 (.84)	−0.26	0.69	0.07	0.76
0	1	1	−0.08 (.77)	−0.12 (.82)	−0.09 (.91)	...	−0.07 (.76)	−0.10	0.63	0.00	0.63
1	0	1	−0.15 (.76)	−0.26 (.79)	−0.18 (.89)	...	−0.08 (.76)	−0.15	0.61	0.01	0.62
1	1	0	−0.15 (.84)	0.09 (.86)	0.38 (.92)	...	−0.44 (.84)	0.00	0.72	0.05	0.77
0	1	2	0.42 (.84)	0.26 (.86)	0.02 (.92)	...	0.54 (.85)	0.30	0.72	0.03	0.74
1	0	2	0.33 (.83)	0.08 (.83)	−0.08 (.89)	...	0.53 (.84)	0.23	0.69	0.04	0.74
1	1	1	0.29 (.77)	0.39 (.82)	0.50 (.91)	...	0.11 (.76)	0.35	0.64	0.01	0.65
1	1	2	0.85 (.84)	0.81 (.86)	0.61 (.92)	...	0.77 (.85)	0.81	0.73	0.01	0.74

Note. EAP = expected a posteriori; MI = multiple imputation.

Table 4. Response Pattern and Summed-Score EAPs

Response pattern EAP						Summed-score EAP			
y_1	y_2	y_3	$\hat{\theta}_{\hat{y}_N}$ (*SE*)	$\bar{\theta}$ (*SE*)	*r* (%)	Sum	$\hat{\theta}_{\hat{y}_N}$ (*SE*)	$\bar{\theta}$ (*SE*)	*r* (%)
0	0	0	−1.07 (.84)	−1.03 (.84)	1.6	0	−1.07 (.84)	−1.03 (.84)	1.6
0	0	1	−0.62 (.79)	−0.58 (.80)	4.2	1	−0.52 (.85)	−0.52 (.85)	1.5
0	1	0	−0.50 (.86)	−0.51 (.86)	2.5				
1	0	0	−0.60 (.84)	−0.55 (.85)	4.3				
0	0	2	−0.30 (.84)	−0.26 (.88)	10.8	2	−0.10 (.84)	−0.09 (.85)	1.5
0	1	1	−0.09 (.81)	−0.10 (.80)	0.2				
1	0	1	−0.21 (.79)	−0.15 (.79)	2.5				
1	1	0	−0.01 (.86)	0.00 (.88)	7.6				
0	1	2	0.32 (.86)	0.30 (.86)	3.7	3	0.31 (.85)	0.31 (.85)	1.5
1	0	2	0.18 (.84)	0.23 (.86)	6.6				
1	1	1	0.34 (.81)	0.35 (.81)	2.1				
1	1	2	0.81 (.86)	0.81 (.86)	1.4	4	0.81 (.86)	0.81 (.86)	1.4

Note. EAP = expected a posteriori. $\hat{\theta}_{\hat{y}_N}$ is the "plug-in" estimator that uses maximum likelihood estimates of item parameters, and $\bar{\theta}$ is the multiple imputation estimator.

In Figure 3, the SEM are plotted against the score estimates. There are differences in the SEM. As expected, the SEMs are generally larger for the MI scores. The magnitude of the correction, however, is substantially larger for the full-pattern EAP scores than for the summed-score EAPs (consistent with the values of *r* reported in Table 4).

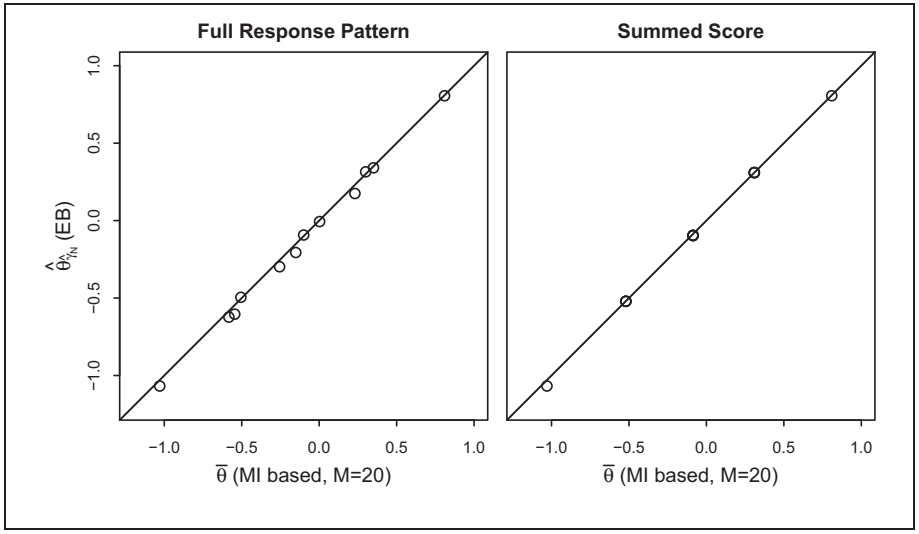

Figure 2. Traditional and MI-based EAP estimates
Note. MI = multiple imputation; EAP = expected a posteriori.

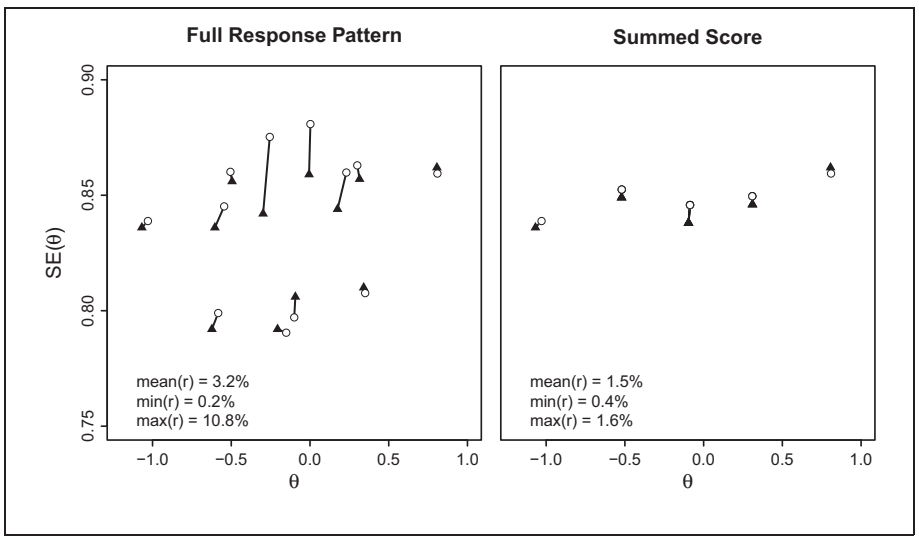

Figure 3. MI-based scale score and SEM corrections
Note. MI = multiple imputation; SEM = standard error of measurement. Traditional estimates are represented by the closed triangles. MI-based estimates are shown as open circles.

Multiple Imputation Confidence Envelopes

Confidence envelopes for item characteristic curves. Thissen and Wainer (1990) considered confidence envelopes for the item characteristic curves. Let $T_k(\theta|\boldsymbol{\gamma})$ denote a generic item category response curve for category k. For example, for the 2PL IRT model, the characteristic curve for the ''correct'' response is

$$T_1(\theta|\boldsymbol{\gamma}) = \frac{1}{1 + \exp[-(c + a\theta)]},$$

where c is the item intercept, a is the item slope, and both are components of $\boldsymbol{\gamma}$. When $\boldsymbol{\gamma}$ is known without error, the curves are simply mathematical functions of the item parameters. When estimates of item parameters are used, $T_k(\theta|\hat{\boldsymbol{\gamma}}_N)$ contains uncertainty. Thissen and Wainer (1990) reasoned that since the posterior distribution of $\boldsymbol{\gamma}$ is reasonably well approximated by a normal with mean $\hat{\boldsymbol{\gamma}}_N$ and covariance matrix $\mathcal{I}_N^{-1}$ (i.e., Equation 1), if one can produce multiple random samples of $\boldsymbol{\gamma}$ from this approximate posterior, approximate confidence limits of the item characteristic curves can be found by plotting the randomly varying item characteristic curves over repeated imputations. These so-called ''M-line'' plots were generated by Thissen and Wainer (1990) to illustrate the shape of confidence envelopes for a variety of dichotomous IRT models. However, Thissen and Wainer's method has the same limitations as Mislevy et al.'s (1993) expected response function method. The random sampling of the item parameters was done in an item-by-item manner because they did not have access to the full error covariance matrix of the item parameter estimates.

We present here a slight variation of Thissen and Wainer's (1990) basic idea:

1. Draw $M > 1$ sets of values from a multivariate normal distribution with mean $\hat{\boldsymbol{\gamma}}_N$ and covariance matrix $\mathcal{I}_N^{-1}$. Denote them as $\boldsymbol{\gamma}_j$ for $j = 1, \ldots, M$.
2. Choose a reasonably fine grid to numerically represent θ, for example, from -3 to $+3$ in step sizes of .01.
3. Plug each $\boldsymbol{\gamma}_j$ into $T_k(\theta|\boldsymbol{\gamma})$, and at a chosen θ level, empirically locate the upper and lower $1 - \alpha/2$ quantile from the M values of $T_k(\theta|\boldsymbol{\gamma}_j)$.
4. Repeat the last step for all θ levels to find a $(1 - \alpha) \times 100\%$ confidence envelope.

To ensure that the boundaries of the confidence envelopes are well characterized, a large M is necessary. We generally use $M = 1{,}000$ random draws.

Figure 4 presents confidence envelopes for the three items in our illustration. The upper, middle, and lower panels correspond to Items 1 (2PL), 2 (3PL), and 3 (GRM) in order. The solid curve is the usual item characteristic curve based on the MML item parameter estimates. The dotted curves represent the upper and lower 95% confidence limits, based on $M = 1{,}000$ random samples drawn from the multivariate normal approximation to the posterior distribution of the item parameters. Incidentally, we also computed the expected response functions (dashed curves), obtained by averaging the response functions across all the imputed parameter sets. As Mislevy et al. (1993) noted, the expected response functions are not logistic and tend to have slightly

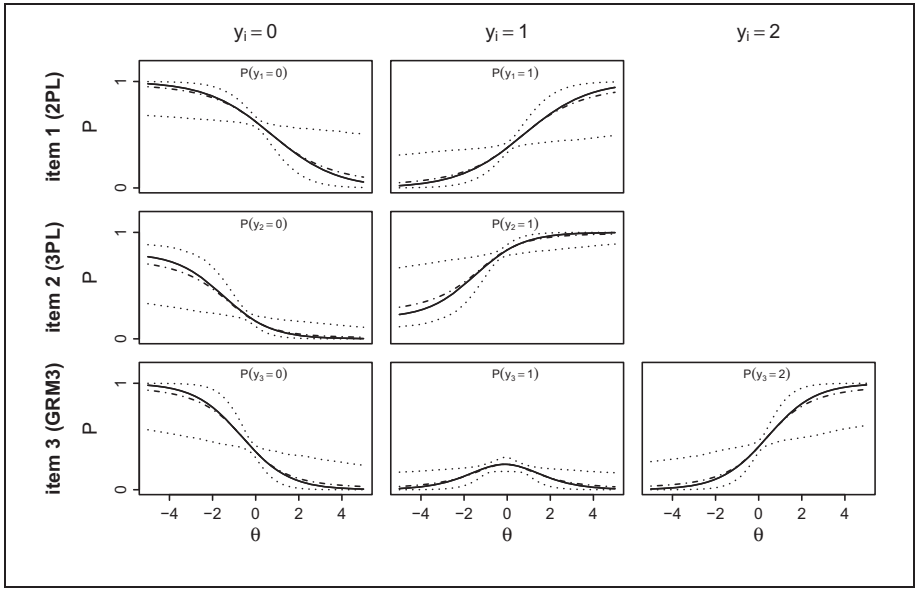

Figure 4. Confidence envelopes for item trace lines

Note. The solid curves represent the item characteristic curves based on the maximum marginal likelihood item parameter estimates. The dotted curves represent the 95% confidence limits. The dashed curves show the expected response functions. The confidence limits and expected functions are generated based on $M = 1{,}000$ imputations.

lower slopes than the item response function based on the MML item parameter estimates.

Confidence envelopes for information and SEM curves. For general multiple-categorical IRT models, item i's Fisher information function is given by the following expression:

$$F_i(\theta|\boldsymbol{\gamma}) = -\sum_{k=0}^{K-1} T_k(\theta|\boldsymbol{\gamma}) \frac{\partial^2}{\partial \theta^2} \log T_k(\theta|\boldsymbol{\gamma}),$$

where K is the number of categories (see Baker & Kim, 2004). The Fisher information functions are additive. For n items, the test information function is the sum of item information functions:

$$F(\theta|\boldsymbol{\gamma}) = \sum_{i=1}^{n} F_i(\theta|\boldsymbol{\gamma}). \tag{13}$$

When a prior for the ability distribution is used in scoring (e.g., MAP scoring), test information must also include the contribution from the prior. The standard error of

measurement curve is found as a one-to-one transformation of the test information curve:

$$\text{sem}(\theta|\boldsymbol{\gamma}) = \frac{1}{\sqrt{F(\theta|\boldsymbol{\gamma})}}. \tag{14}$$

Recognizing that $F(\theta|\boldsymbol{\gamma})$ is a nonlinear transformation of $\boldsymbol{\gamma}$, we follow essentially the same strategy used in the previous sections to create confidence envelopes for the test information function:

1. Draw $M > 1$ sets of values from a multivariate normal distribution with mean $\hat{\boldsymbol{\gamma}}_N$ and covariance matrix $\mathcal{I}_N^{-1}$. Denote them as $\boldsymbol{\gamma}_j$ for $j = 1, \ldots, M$.
2. Choose a reasonably fine grid to numerically represent θ, for example, from -3 to $+3$ in step sizes of .01.
3. Plug each $\boldsymbol{\gamma}_j$ into $F(\theta|\boldsymbol{\gamma})$, and at a chosen θ level, empirically locate the upper and lower $1 - \alpha/2$ quantile from the M values of $F(\theta|\boldsymbol{\gamma}_j)$.
4. Repeat the last step for all θ levels to find a $(1 - \alpha) \times 100\%$ confidence envelope.

Because $F(\theta|\boldsymbol{\gamma})$ and $\text{sem}(\theta|\boldsymbol{\gamma})$ enjoy one-to-one relation, the confidence limits for the SEM curve are found by transforming the confidence limits of the test information function.

As an illustration, consider the confidence envelopes for the test information function and the conditional SEM for the three-item data set, presented in Figure 5.

These curves highlight the extent to which our characterizations of test information or the conditional SEM can be affected by item parameter uncertainty. This may have implications for test assembly, in which particular combinations of items may be selected to produce a particular test information or conditional SEM curve. The width of the confidence envelopes we observe for these functions suggests that assembling tests to closely match target information of SEM curves may be misguided. The confidence envelopes generated through the MI-based approach allow the imprecision because of uncertainty in item parameters to be visualized.

Confidence intervals for marginal reliability. As a final application, we consider confidence intervals for the marginal reliability coefficient, which is an important IRT-based measure of overall scale reliability. Let the prior $f(\theta)$ for the ability distribution have variance σ^2. Marginal reliability is defined as

$$\bar{\rho}(\boldsymbol{\gamma}) = \frac{\sigma^2 - \int [\text{sem}(\theta|\boldsymbol{\gamma})]^2 f(\theta) d\theta}{\sigma^2}, \tag{15}$$

(see, e.g., Thissen & Wainer, 2001), where the integral in the numerator returns the average error variance, and $\text{sem}(\theta|\boldsymbol{\gamma})$ is as defined in Equation (14). Of course, $\bar{\rho}$ is also a nonlinear transformation of $\boldsymbol{\gamma}$, suggesting the following algorithm:

1. Draw $M > 1$ sets of values from a multivariate normal distribution with mean $\hat{\boldsymbol{\gamma}}_N$ and covariance matrix $\mathcal{I}_N^{-1}$. Denote them as $\boldsymbol{\gamma}_j$ for $j = 1, \ldots, M$.

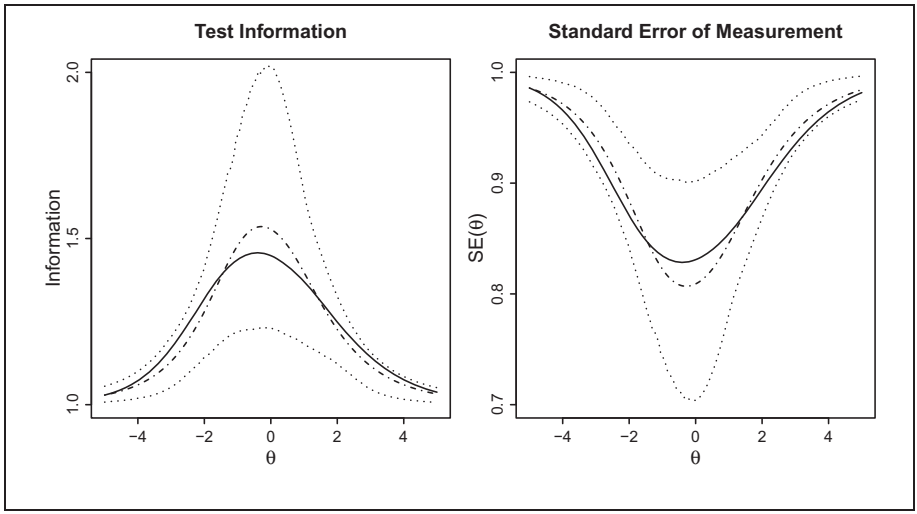

Figure 5. Confidence envelopes for test information and the conditional SEM
Note. SEM = standard error of measurement. Solid curves show test information and the conditional SEM based on the maximum marginal likelihood item parameter estimates. The dotted curves represent the 95% confidence limits. The expected information and conditional SEM curves are shown as dashed lines. The confidence limits and expected functions are based on $M = 1{,}000$ imputations.

2. Plug each $\boldsymbol{\gamma}_j$ into $\bar{\rho}(\boldsymbol{\gamma})$ and empirically locate the upper and lower $1 - \alpha/2$ quantile from the M values of $\bar{\rho}(\boldsymbol{\gamma}_j)$. The result is a $(1 - \alpha) \times 100\%$ confidence interval for marginal reliability.

For the three-item data set, the MML item parameter estimates yield an estimated marginal reliability of .29 (which is low, as expected, given the very short test length). However, the MI-based approach offers a more complete characterization—that the marginal reliability has a 95% confidence interval between .17 and .43 based on $M = 1{,}000$ imputations.

Summation

The proposed MI approach provides not only corrected scale scores and standard errors of measurement that take parameter uncertainty into account but also confidence envelopes or intervals for other important quantities such as the item characteristic curve, the test information function, the conditional SEM curve, and the marginal reliability coefficient. As demonstrated with the three-item data set, the approach is flexible enough to be applied to not only full-pattern EAP scores but also other scoring methods (e.g., summed-score EAP) and various IRT models (e.g., 2PL, 3PL, or GRM).

Application to Empirical Data

Data and Method

As an empirical demonstration of the proposed MI approach, we analyzed data from the 2000 Program for International Student Assessment (PISA; Adams & Wu, 2002). We extracted a random sample of 1,500 students from the United States with complete responses for Math Booklet 8. For our analyses, we used 500 students as a calibration sample and 1,000 students as a scoring sample. Math Booklet 8 consists of 15 items—8 free response (FR), 5 multiple choice (MC), and 2 complex multiple choice (CMC). Logistic graded-response models were fit to the FR and CMC items with the number of ordered categories determined by the number of different scores assigned (either 2 or 3). A 3PL model was used for the MC items. In estimating item parameters, a log-normal prior was placed on the guessing parameter. The mean of this prior was based on the expected probability of a correct response, assuming blind guessing and given the number of response options. Item parameter estimates and the item parameter asymptotic covariance matrix were obtained using IRTPRO (Cai, du Toit, et al., in press).

Because PISA items are nested within testlets or passages, we obtained estimates for both a unidimensional (ignoring the testlet structure but consistent with how the test data are scored in practice) and a bifactor item response model with specific factors modeling testlet effects. Parameter estimation and scoring for the bifactor model has been described elsewhere (Cai, Yang, et al., 2011). Here, we focus on the effects of item parameter uncertainty on the precision of scores for the general dimension and show the application of the MI-based method in both the unidimensional and multidimensional IRT contexts.

As outlined in previous sections, the MML item parameter estimates and the item parameter error covariance matrix were used to generate multiple sets of plausible parameter values. For the unidimensional model, 39 parameters were estimated. The bifactor model required estimation of additional 10 parameters (specific factor slopes for the 10 items loading on three specific factors).

As before, 20 sets of parameters were imputed to obtain the MI-based score estimates based on the full response patterns and summed scores for individuals in the scoring sample. Confidence envelopes for test information and the conditional SEM and a confidence interval of the marginal reliability coefficient were based on $M = 1{,}000$ imputations.

Results

As shown in Figure 6, the scale score estimates are almost perfectly correlated for both unidimensional model and bifactor model, regardless of the scoring method (full response pattern or summed-score EAP).

Figure 7 shows the scale score and SEM estimates and their corrections based on the MI approach. For most individuals in the scoring sample, the SEM increases under the MI-based approach. Across the scoring sample, the average relative increase in

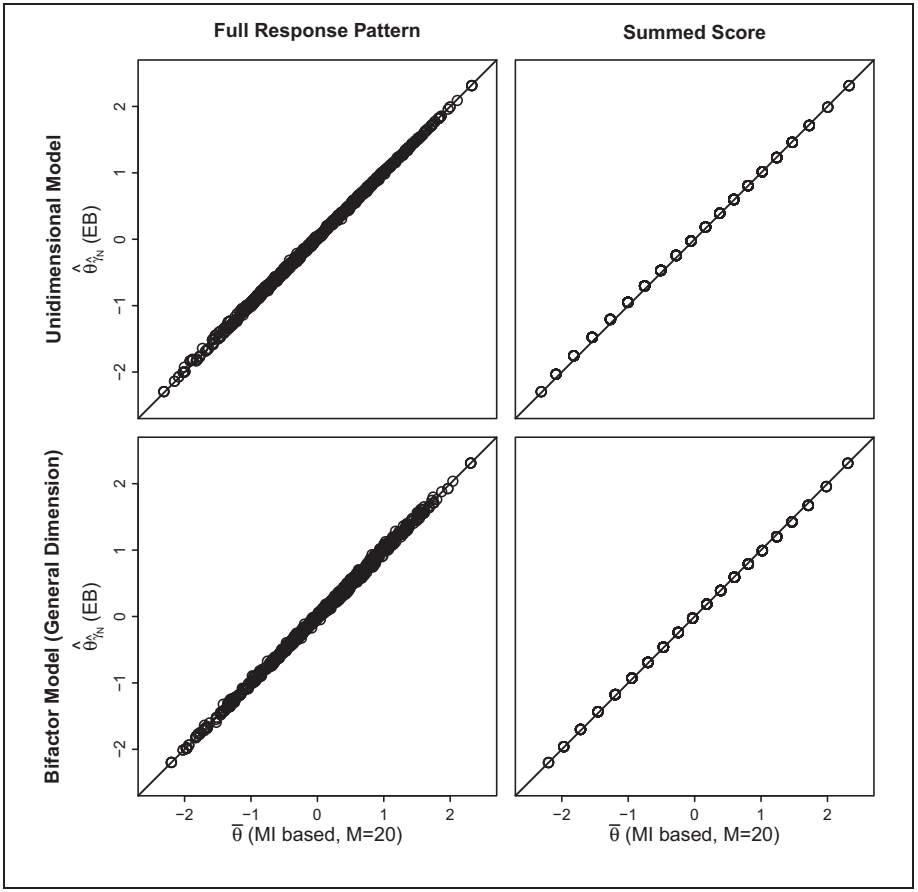

Figure 6. Estimated IRT scale scores for PISA Math Booklet 8
Note. IRT = item response theory; PISA = Program for International Student Assessment.

variance (r) based on Equation (12) was 3.5% for the unidimensional model and 7.6% for the bifactor model under full–response pattern scoring. In contrast, the average relative increase in variance for the summed-score EAP estimates was less than 1% for either scoring method. In addition to having larger average increases, the impact of item parameter uncertainty on the full-pattern scores appears to be more variable than for the summed score translations. Specifically, there are some response patterns with vary large increases in the SEM (up to 32.9% for the unidimensional model and 67.1% for the bifactor model). The increases for the summed score are much more uniform across all response patterns.

The 95% confidence envelopes for test information and the conditional SEM for the unidimensional model are shown in Figure 8. Graphical representations and proper

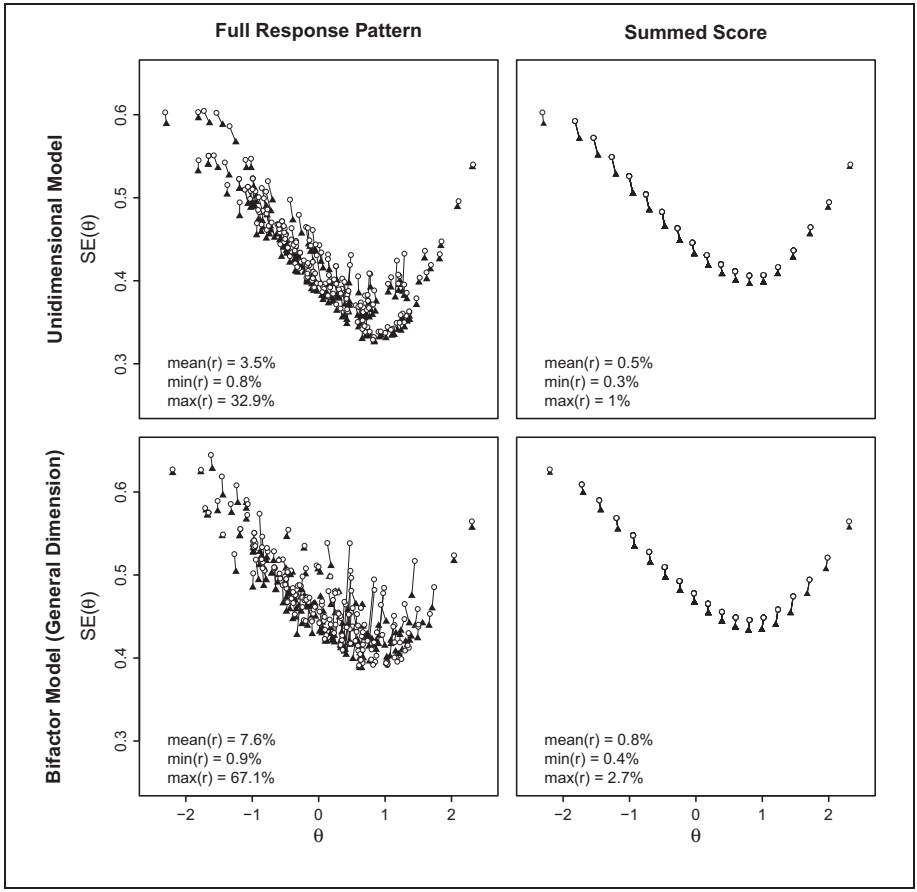

Figure 7. Traditional and MI-based response pattern and summed-score SEM estimates plotted against EAPs for PISA Math Booklet 8

Note. MI = multiple imputation; SEM = standard error of measurement; EAP = expected a posteriori; PISA = Program for International Student Assessment. For clarity, values for 200 randomly selected individuals in the scoring sample are plotted; the mean, minimum, and maximum values of *r* are based on the full scoring sample of 1,000 individuals. Traditional estimates and SEMs based on the maximum marginal likelihood item parameters are shown as triangles. MI-based results are shown as circles.

interpretation of test information in the multidimensional case, though potentially quite interesting, are beyond the scope of this study.

For this 15-item test, the MML parameter estimates produce a marginal reliability of .82. From the MI-based approach, its 95% confidence interval was found to be .78 to .84.

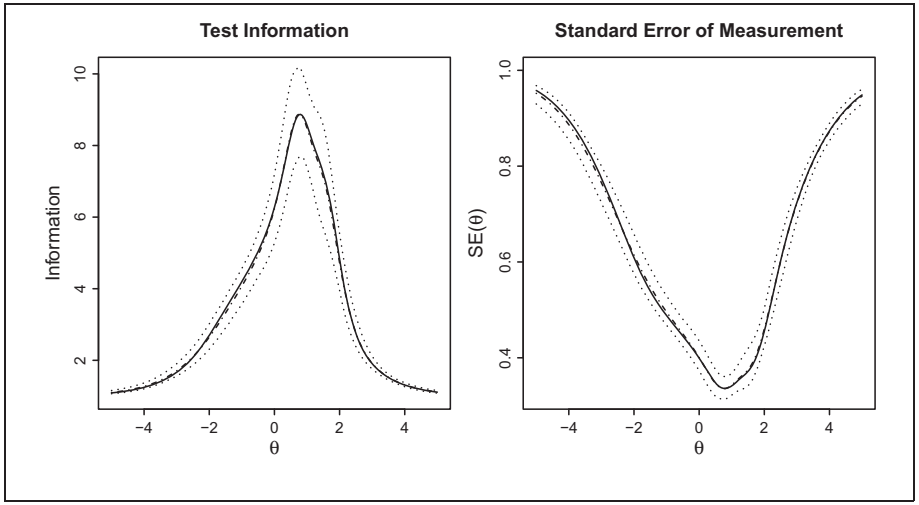

Figure 8. Confidence envelopes for test information and the conditional SEM for PISA Math Booklet 8

Note. SEM = standard error of measurement; PISA = Program for International Student Assessment. The solid curves show test information and the conditional SEM based on the maximum marginal likelihood item parameter estimates. The dotted curves represent the 95% confidence limits. The expected information and conditional SEM curves are shown as dashed lines. The confidence limits and expected functions are based on $M = 1{,}000$ imputations.

A Simulation Study

Previous studies have shown that the magnitude of bias in the SEMs can vary depending on conditions such as item response model complexity, calibration sample size, and test length, all of which influence the test information. For example, Tsutakawa and Johnson (1990) observed that relatively simple Rasch models tend to provide decent estimation of the latent ability levels and their SEMs, even with small calibration samples. In contrast, a 3PL model with a calibration sample size of 400 individuals produced biases of the posterior means of the latent trait and underestimation of the posterior standard deviations by more than 40%, on average.

As a complement to the simple illustration and the empirical analysis of PISA data, a simulation study was conducted. The goal was to further characterize the conditions in which the uncertainty carried over from item calibration leads to uncertainty in the scoring process. Based on the findings of previous studies, we manipulated the following: model complexity, calibration sample size, and test length. For this study, the number of imputations was fixed to $M = 20$, as previous research has already demonstrated that 20 imputations should provide reasonably good correction of the SEMs (Mislevy et al., 1993). The combination of the number of items ($J = 5, 10, 20, 40$), the size of

Table 5. Relative Increase in Variance, *r*(%)

	Calibration sample size (*N*)			
Number of items (*J*)	500	1,000	2,000	5,000
Two-parameter logistic model				
5	2.6	1.2	0.6	0.2
10	2.2	1.1	0.5	0.2
20	2.7	1.3	0.7	0.3
40	14.7	3.5	1.2	0.4
Three-parameter logistic model				
5	4.7	2.1	0.9	0.4
10	2.9	1.4	0.8	0.3
20	3.5	1.8	0.9	0.4
40	8.8	2.4	1.1	0.5
Graded-response model				
5	2.3	1.1	0.5	0.2
10	1.8	0.9	0.5	0.2
20	2.3	1.2	0.6	0.3
40	3.3	1.7	0.9	0.4

calibration sample ($N = 500, 1{,}000, 2{,}000, 5{,}000$), and item type (2PL, 3PL, and logistic GRM with five categories) yielded a total of 48 conditions.

For each condition, 500 calibration samples were generated; in addition, one independent scoring sample of 10,000 cases was produced. The item parameters used for data generation were chosen to resemble estimates obtained from real educational and psychological data sets. Following data generation, we calibrated the items using MML estimation and obtained the parameter error covariance matrix. There were calibrations for which either the EM algorithm (used to obtain the parameter estimated) or supplemented EM (needed for the covariance matrix) failed to converge. These replications (about 9% of the total) were omitted from the subsequent analyses. Twenty item parameter sets were randomly drawn from the multivariate normal approximation to the posterior distribution of the item parameters, as described in previous sections. These parameter sets were then used to obtain the MI-based scores and SEM for the scoring sample. For each of the accepted calibrations (500, for most conditions), r was calculated for each of case in the corresponding scoring sample. The median value of r (across calibrations) was then obtained for each case. These medians were then averaged across the entire scoring sample and within deciles (1,000 individuals) of the ''true'' level of the latent trait for each condition. The results are reported in Table 5 and Figure 9.

For most conditions, the average relative increase in variance is rather small, ranging from 1.8% to 14.7%. However, some trends are evident. For a given test (i.e, for a fixed number of items and item type), the average relative increase in variance decreases as the size of the calibration sample increases. This is to be expected, as larger calibration samples yield more precise estimates of the item parameters. The patterns of relative increase across tests of different length and across item types

are more complex. Test information increases with the number of items. Thus, for longer tests, we expect smaller ''within''-imputation error variance. At the same time, longer tests require estimation of more item parameters, resulting in greater ''between'' variance. In cases where the calibration sample is small, those parameter estimates may not be very precise. Consequently, the highest relative increase in variance was observed for the 40-item tests with the smallest calibration sample examined. Even for these conditions, however, the increases appear rather modest.

Figure 9 presents a more detailed view of the relative increase in variance for the five-category graded items across the simulation conditions of varied test length and calibration sample size. Here, the average relative increase in variance is obtained within each decile of the true scores on the latent trait (which are known for these simulated data but, of course, are never known in practice). As in Table 5, it is clear that the relative increase in variance diminishes with increasing sample size. It is also apparent that the percentages are not uniform across the range of the latent trait. For those conditions where there is substantial relative increase in variance (e.g, $J = 40$, $N = 500$), the percentages are greatest at the highest and lowest deciles of the true θs. This is again consistent with existing results obtained by analytic approximation (e.g., Cheng & Yuan, 2010) for simpler IRT models.

Discussion

In this research, we have proposed an MI-based approach for characterizing the ways in which uncertainty about item parameters affects item response functions, test information functions, SEMs, and marginal reliability. We argue that the approach presents several advantages over the existing alternatives. First, it can be applied to a variety of IRT models and scoring methods. Second, it is computationally simple and uses information that is readily available in the output of standard IRT software programs. Third, the approach makes use of the asymptotic covariance matrix of the item parameters, obtained through an application of the supplemented EM algorithm (Cai, 2008). This allows us to conduct analyses of entire tests, whereas past efforts have mostly focused on parameter uncertainty with respect to individual items. Finally, our proposed approach connects a variety of seemingly disparate methods that have been used to handle item parameter uncertainty. In so doing, our approach integrates the related goals of providing corrected standard errors of measurement and the (highly visual) characterization of the effects of uncertainty.

To demonstrate the relevance of the proposed approach to the problem of uncertainty in item parameters, we derived approximations for EAP scores and their SEM (Equations 8 and 9, respectively) that use the MML estimates of the item parameters and their covariance matrix. The MI-based EAP approximation is an average of the EAP scores obtained with $M > 1$ sets of plausible values for the item parameters. The error variance for this estimate is a combination of the within- and between-imputation variances. The square root of the total error variance provides a corrected SEM.

Crucial to the MI-based method is the random sampling of plausible item parameter values from a multivariate normal approximation to the item parameter posteriors.

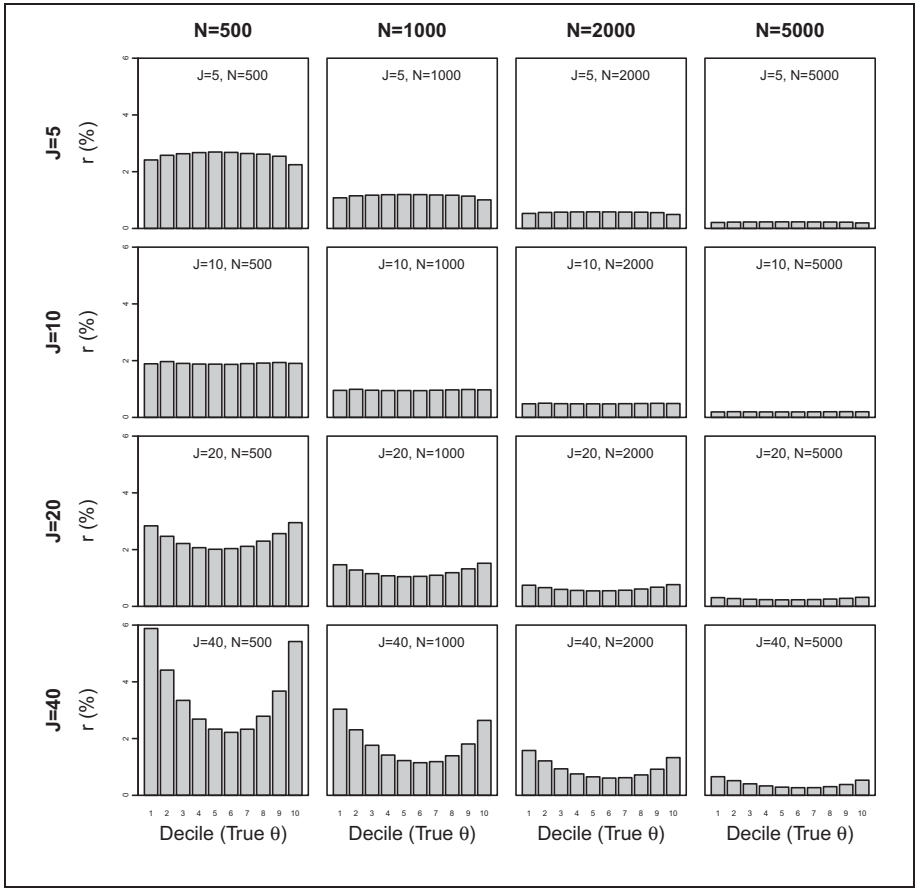

Figure 9. Mean relative increase in variance (by deciles) under various simulated test lengths and calibration sample sizes for the logistic graded-response model with five categories

The MI-based score estimates and SEMs are obtained by scoring with these imputed parameter sets and combining the results. In addition to these corrections, we constructed confidence envelopes for item characteristic curves, building on the work of Thissen and Wainer (1990). A depiction such as this helps to convey the extent to which error in the item parameters leads to uncertainty about the shape of the response functions. Importantly, the amount of uncertainty may vary across levels of the latent trait. A similar approach was taken to generate confidence envelopes for the test information function and the conditional SEM curve. Finally, we used MI to calculate confidence intervals for marginal reliability. This allows us to convey the uncertainty in the marginal reliability that is because of error in the estimation of the item parameters.

After illustrating the proposed approach with an artificial, three-item data set, we examined data from the 2000 PISA math test. We also conducted simulations, which

allowed us to investigate how various factors can influence the uncertainty in scores. In these simulations, we manipulated the size of the calibration sample, the length of the test, and the complexity of the response model. The effect of parameter uncertainty was quantified as the relative increase in error variance. This analysis demonstrated that, on the whole, parameter uncertainty contributes little to total error variance. However, in situations where the calibration sample is small and the number of items is large (and especially in the case of a more complex response model), the error carried over from item calibration may occasionally be nonnegligible.

It is important to know the extent to which the latent trait estimates are uncertain because a number of critical decisions are based on the SEM. In variable-length computerized adaptive testing algorithms, for example, the SEM is often used as a termination criterion. In such cases, underestimation of SEM can result in premature termination of the test. In addition, such underestimation could result in flawed inferences concerning individuals' standing relative to a certain performance standard or to one another. Specifically, scores might be assumed to have greater precision than is warranted, given the known uncertainty in the item parameters. Standard errors have also been incorporated into statistical models such as hierarchical linear models with latent variables (Raudenbush & Bryk, 2002). For such applications, improved SEM estimates will enhance estimation of regression parameters and their associated standard errors. The MI-based approach presented in this article provides a simple and flexible means of obtaining these improved estimates.

Appendix

The EAP estimator of θ is a posterior expectation, that is,

$$\hat{\theta} = \int \theta f(\theta|\mathbf{x}, \mathbf{Y}) d\theta.$$

From Equation (5), the equation above can be rewritten as

$$\hat{\theta} = \frac{1}{f(\mathbf{x}|\mathbf{Y})} \int \left[\int f(\mathbf{x}|\theta, \boldsymbol{\gamma}) f(\boldsymbol{\gamma}|\mathbf{Y}) d\boldsymbol{\gamma} \right] \theta f(\theta) d\theta.$$

Interchanging the order of integration, we see that

$$\begin{aligned}
\hat{\theta} &= \frac{1}{f(\mathbf{x}|\mathbf{Y})} \int \left[\int \theta f(\mathbf{x}|\theta, \boldsymbol{\gamma}) f(\theta) d\theta \right] f(\boldsymbol{\gamma}|\mathbf{Y}) d\boldsymbol{\gamma} \\
&= \frac{1}{f(\mathbf{x}|\mathbf{Y})} \int \left[\int \theta f(\theta|\mathbf{x}, \boldsymbol{\gamma}) d\theta \right] f(\mathbf{x}|\boldsymbol{\gamma}) f(\boldsymbol{\gamma}|\mathbf{Y}) d\boldsymbol{\gamma} \\
&= \int \hat{\theta}_{\boldsymbol{\gamma}} f(\boldsymbol{\gamma}|\mathbf{x}, \mathbf{Y}) d\boldsymbol{\gamma},
\end{aligned}$$

where the last line requires the conditional independence of $\mathbf{x}$ and $\mathbf{Y}$ given $\boldsymbol{\gamma}$. By the same token, the posterior variance can be written as

$$\begin{aligned} V(\hat{\theta}) &= \int \theta^2 f(\theta|\mathbf{x}, \mathbf{Y}) d\theta - \hat{\theta}^2 \\ &= \int [V(\hat{\theta}_{\boldsymbol{\gamma}}) + \hat{\theta}_{\boldsymbol{\gamma}}^2] f(\boldsymbol{\gamma}|\mathbf{x}, \mathbf{Y}) d\boldsymbol{\gamma} - \hat{\theta}^2 \\ &= \int V(\hat{\theta}_{\boldsymbol{\gamma}}) f(\boldsymbol{\gamma}|\mathbf{x}, \mathbf{Y}) d\boldsymbol{\gamma} + \int (\hat{\theta}_{\boldsymbol{\gamma}} - \hat{\theta})^2 f(\boldsymbol{\gamma}|\mathbf{x}, \mathbf{Y}) d\boldsymbol{\gamma}. \end{aligned}$$

Authors' Note

The views expressed in this article belong to the authors and do not reflect the views/policies of the funding agencies or grantees.

Acknowledgment

We thank Dr. David Thissen for comments on an earlier draft.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article:

The authors acknowledge financial support from the following sources: Institute of Education Sciences (R305B080016 and R305D100039) and the National Institute on Drug Abuse (R01DA026943 and R01DA030466).

References

Adams, R., & Wu, M. (2002). *PISA 2000 technical report*. Paris, France: Organization for Economic Cooperation and Development.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*, 309-329.

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

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Cheng, Y., & Yuan, K.-H. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika*, *75*, 280-291.

Edwards, M. C. (2005). *A Markov chain Monte Carlo approach to confirmatory item factor analysis* (Unpublished doctoral dissertation). Department of Psychology, University of North Carolina at Chapel Hill.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*, 407-433.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika*, *57*, 423-436.

Hoshino, T., & Shigemasu, K. (2008). Standard errors of estimated latent variable scores with estimated structural parameters. *Applied Psychological Measurement, 32*, 181-189.

Lewis, C. (1985). *Estimating individual abilities with imperfectly known response functions*. Paper presented at the Annual Meeting of the Psychometric Society, Nashville, TN.

Lewis, C. (2001). Expected response functions. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 163-70). New York, NY: Springer-Verlag.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score ''equatings.'' *Applied Psychological Measurement, 8*, 453-461.

Meng, X.-L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association, 86*, 899-909.

Mislevy, R. J., Wingersky, M. S., & Sheehan, K. M. (1993). *Dealing with uncertainty about item parameters: Expected response functions* (Research Report No. 94-28). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Yan, D. (1991). *Dealing with uncertainty about item parameters: Multiple imputations and SIR*. Paper presented at the Annual Meeting of the Psychometric Society, Princeton, NJ.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: Chapman & Hall/CRC.

Thissen, D., & Wainer, H. (1990). Confidence envelopes for item response theory. *Journal of Educational Statistics, 15*, 113-128.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*, 371-390.

Tsutakawa, R. K., & Soltys, M. J. (1988). Approximation for Bayesian ability estimation. *Journal of Educational Statistics, 13*, 117-130.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, England: Cambridge University Press.

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