Published in final edited form as:
Psychometrika. 2010 June ; 75(2): 280–291. doi:10.1007/s11336-009-9144-x.

THE IMPACT OF FALLIBLE ITEM PARAMETER ESTIMATES ON LATENT TRAIT RECOVERY

Ying Cheng and **Ke-Hai Yuan**
University of Notre Dame

Abstract

In this paper we propose an upward correction to the standard error (SE) estimation of $\hat{\theta}_{ML}$, the maximum likelihood (ML) estimate of the latent trait in item response theory (IRT). More specifically, the upward correction is provided for the SE of $\hat{\theta}_{ML}$ when item parameter estimates obtained from an independent pretest sample are used in IRT scoring. When item parameter estimates are employed, the resulting latent trait estimate is called pseudo maximum likelihood (PML) estimate. Traditionally the SE of $\hat{\theta}_{ML}$ is obtained on the basis of test information only, as if the item parameters are known. The upward correction takes into account the error that is carried over from the estimation of item parameters, in addition to the error in latent trait recovery itself. Our simulation study shows that both types of SE estimates are very good when θ is in the middle range of the latent trait distribution, but the upward-corrected SEs are more accurate than the traditional ones when θ takes more extreme values.

Keywords

IRT scoring; pseudo maximum likelihood (PML); standard error; upward correction; latent trait estimation

Introduction

Item response theory (IRT) scoring offers various advantages. For example, the examinees can be placed on the same scale as the items. IRT also offers *conditional* standard errors (SE) of latent trait estimates. In IRT scoring, usually the item parameter values are considered known (see Baker & Kim, 2004; Embretson & Reise, 2000), unless the joint maximum likelihood (JML) method (Birnbaum, 1968) is used, which can produce estimates of the item and latent trait parameters simultaneously. JML estimation used to be a standard practice for item parameter estimation as implemented in LOGIST (Wingersky, Barton, & Lord, 1982). However, using JML estimation for tests of finite length may result in estimates that are not statistically consistent (Harwell & Baker, 1991).

Nowadays, it is a common practice in psychological and educational testing to first estimate item parameters from a pretest sample, and then treat the item parameter estimates as true values when estimating the latent trait in the scoring sample. However, the item parameter estimates are "fallible" in the sense that they do not equal the true parameter values, and they have their own standard errors. When item parameter estimates are treated as true values in scoring, the errors will affect the precision of resulting latent trait estimates.

Please address correspondence to Ying Cheng, 118 Haggar Hall, Notre Dame, IN, 46556, ycheng4@nd.edu, Phone: (574) 631-7649, Fax: (574) 631-8883.

It is important to obtain good estimates of the standard errors (SE) for the latent trait estimates. For instance, the SE can serve as the termination criterion in variable-length computerized adaptive testing. The test stops for an examinee when his or her ability estimate is sufficiently precise, i.e., when the SE of the final latent trait estimate is small enough. If the SE is underestimated, the test may end prematurely.

Researchers have long been aware of the issue that, when item parameter estimates are used instead of the true values, the standard error of final ability estimates can be underestimated. In the context of Bayes estimation with the three-parameter logistic (3PL) model (Hambleton & Swaminathan, 1985), Tsutakawa & Johnson (1990) discussed two sources of error in latent trait estimation, one being the scoring process itself, and the other being the error carried over from item calibration. They proposed approximations of the posterior mean and variance of latent trait and found that the uncertainty in latent trait estimation may be seriously underestimated when the pretest sample is small or moderately large. A remedy is to use posterior standard deviations from the Markov chain Monte Carlo (MCMC) output as standard errors. Specifically, one can obtain the latent trait estimates along with the standard errors by marginalizing over other dimensions of the joint posterior of item and person parameters (Patz & Junker, 1999).

Instead of adopting the Bayes framework, we will study the effect of fallible item parameter estimates on the precision of latent trait estimates in the context of maximum likehood (ML) estimation, which is widely used in IRT applications (see chapter 3 of Baker & Kim, 2004; chapter 7 of Embretson & Reise, 2000). In the ML approach with known item parameters, the SE of the estimated latent trait is obtained from the test information. When the item parameters are estimated rather than known, the true SE of the estimated latent trait will be inflated. Therefore the theory of ML has to take into account the extra errors. The procedure is called pseudo ML (PML) (Gong & Samaniego, 1981; Parke, 1986) estimation, which leads to an asymptotic quantification of how the errors in item parameter estimates affect the estimation of latent trait. The development allows us to obtain a correction to the information-based SE. It can also be shown that, as long as the item parameter estimates are consistent, the ML estimates (MLE) of latent traits are also consistent. We will use Monte Carlo methods to compare the corrected SE and the original SE against the empirical SE, with calibration sample of different sizes. The simulation study shows that the commonly used formula for SE underestimates the amount of variability in the final latent trait estimates for examinees with high or low latent trait levels.

Methodology

Dichotomous IRT models relate the probability of correct responses to the person parameter, θ. A widely used IRT model is the two-parameter logistic model (2PL, Birnbaum, 1968), which includes two item parameters, item discrimination α and item difficulty β:

$$P(U_{ij}=1|\theta_i,\alpha_j,\beta_j)=\{1+\exp[-\alpha_j(\theta_i-\beta_j)]\}^{-1}, \tag{1}$$

where $U_{ij} = 1$ represents a correct answer of subject i to item j, and $U_{ij} = 0$ otherwise.

Without loss of generality, the 2PL model can be re-parameterized as follows:

$$P(U_{ij}=1|\theta_i,\gamma_{j0},\gamma_{j1})=\{1+\exp[-(\gamma_{j0}+\gamma_{j1}\theta_i)]\}^{-1}. \tag{2}$$

Naturally, $\gamma_{j0} = -\alpha_j\beta_j$ and $\gamma_{j1} = \alpha_j$.

The Fisher information measures the amount of information that an observable random variable X carries about an unknown parameter θ (Cover & Thomas, 1991), given a model. In the context of 2PL, the Fisher information of item j can be expressed as:

$$I_j(\theta;\gamma_j)=\gamma_{j1}^2 P(\theta,\gamma_{j0},\gamma_{j1})\,(1-P(\theta,\gamma_{j0},\gamma_{j1})), \tag{3}$$

where $P(\theta, \gamma_{j0}, \gamma_{j1})$ is a shorthand expression of $P(U_j = 1|\theta, \gamma_{j0}, \gamma_{j1})$, and $\gamma_j = (\gamma_{j0}, \gamma_{j1})'$. Assuming local independence, the test information is the sum of item information:

$$I_T(\theta;\gamma)=\sum_{j=1}^{m} I_j(\theta;\gamma_j), \tag{4}$$

where m is test length, and γ is a $2m \times 1$ vector containing all the γ_{j0} and γ_{j1}, $j = 1, 2,\ldots, m$.

The square root of the inverse of test information evaluated at θ',

$$\sqrt{I_T^{-1}\left(\widehat{\theta};\gamma\right)}, \tag{5}$$

provides a consistent estimate of the SE of $\hat{\theta}$. In reality, however, γ is never known with certainty. They are often replaced with estimates from a pretest sample. In other words, the standard error of $\hat{\theta}$ is usually approximated by (see Embretson & Reise, 2000):

$$\sqrt{I_T^{-1}\left(\widehat{\theta};\widehat{\gamma}\right)}. \tag{6}$$

However, Eq. 6 may underestimate the standard error of $\hat{\theta}$. Let's denote the realization of U_{ij} as u_{ij}. The log likelihood function, under the assumption of local independence, is given by:

$$l(\theta,\gamma)=\sum_{i=1}^{n}\sum_{j=1}^{m}\{u_{ij}\log[P(\theta_i,\gamma_{j0},\gamma_{j1})]+(1-u_{ij})\log[Q(\theta_i,\gamma_{j0},\gamma_{j1})]\},$$

where n is the size of the scoring sample, $Q(\theta_i, \gamma_{j0}, \gamma_{j1}) = 1 - P(\theta_i, \gamma_{j0}, \gamma_{j1})$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)'$, $\gamma=(\gamma_1', \gamma_2', \ldots, \gamma_m')'$ with $\gamma_j = (\gamma_{j0}, \gamma_{j1})'$.

Suppose that a consistent estimator $\hat{\gamma}$ is available and we treat it as known when using the ML procedure to estimate $\boldsymbol{\theta}$, our objective function is:

$$l(\theta,\widehat{\gamma})=\sum_{i=1}^{n}\sum_{j=1}^{m}\{u_{ij}\log[P(\theta_i,\widehat{\gamma}_{j0},\widehat{\gamma}_{j1})]+(1-u_{ij})\log[1-P(\theta_i,\widehat{\gamma}_{j0},\widehat{\gamma}_{j1})]\}.$$

When the IRT model defined in Eq. 2 is correct, such an approach to estimation is the aforementioned pseudo maximum likelihood (PML) estimation in the statistics literature (Gong & Samaniego, 1981;Parke, 1986). Under standard regularity conditions the pseudo MLE of θ_i, $\hat{\theta}_i$, satisfies the following estimating equation:

$$g_i(\widehat{\theta}_i, \widehat{\gamma}) = \mathbf{0}, \ i = 1, 2, \ldots, n, \tag{7}$$

where

$$\begin{aligned} g_i(\theta_i, \widehat{\gamma}) \quad &= \frac{1}{m} \sum_{j=1}^{m} \left(\frac{u_{ij}}{P(\theta_i, \widehat{\gamma}_{j0}, \widehat{\gamma}_{j1})} - \frac{1 - u_{ij}}{Q(\theta_i, \widehat{\gamma}_{j0}, \widehat{\gamma}_{j1})} \right) \frac{\partial P(\theta_i, \widehat{\gamma}_{j0}, \widehat{\gamma}_{j1})}{\partial \theta_i} \\ &= \frac{1}{m} \sum_{j=1}^{m} [u_{ij} - P(\theta_i, \widehat{\gamma}_{j0}, \widehat{\gamma}_{j1})] \widehat{\gamma}_{j1}. \end{aligned}$$

When $g_i(\theta_i, \gamma)$ is unbiased, i.e. when $E[g_i(\theta_i, \gamma)] = 0$, $\hat{\theta}_i$ is consistent (e.g., Godambe & Thompson, 1974).

We can use the Taylor expansion to see how the distribution of $\hat{\theta}_i$ is affected by treating $\hat{\gamma}$ as non-stochastic. Let $\dot{g}_{i\theta}(\theta_i, \gamma)$ and $\dot{g}_{i\gamma}(\theta_i, \gamma)$ denote the partial derivatives of $g_i(\theta_i, \hat{\gamma})$ with respect to θ_i and γ, respectively. Then

$$\dot{g}_{i\theta}(\theta_i, \gamma) = -\frac{1}{m} \sum_{j=1}^{m} P(\theta_i, \gamma_{j0}, \gamma_{j1}) Q(\theta_i, \gamma_{j0}, \gamma_{j1}) \gamma_{j1}^{2},$$

$$\dot{g}_{i\gamma} = (a_{10}, a_{11}; a_{20}, a_{21}; \ldots; a_{m0}, a_{m1})/m$$

with

$$a_{j0} = -P(\theta_i, \gamma_{j0}, \gamma_{j1}) Q(\theta_i, \gamma_{j0}, \gamma_{j1}) \gamma_{j1},$$

and

$$a_{j1} = [u_{ij} - P(\theta_i, \gamma_{j0}, \gamma_{j1})] - P(\theta_i, \gamma_{j0}, \gamma_{j1}) Q(\theta_i, \gamma_{j0}, \gamma_{j1}) \gamma_{j1} \theta_i.$$

Suppose $\hat{\gamma}$ is obtained by another sample that is independent from the current sample, with

$$\sqrt{N}(\widehat{\gamma} - \gamma) \xrightarrow{\mathscr{L}} \mathscr{N}(\mathbf{0}, \mathbf{\Omega}), \tag{8}$$

where $\xrightarrow{\mathscr{L}}$ denotes convergence in distribution, and N the size of the calibration sample. Assuming $m \le N$, the first-order Taylor expansion of $g_i(\hat{\theta}_i, \hat{\gamma})$ leads to

$$0 = g_i(\widehat{\theta}_i, \widehat{\gamma}) = g_i(\theta_i, \gamma) + \dot{g}_{i\theta}(\theta_i, \gamma)\,(\widehat{\theta}_i - \theta_i) + \dot{g}_{i\gamma}(\theta_i, \gamma)\,(\widehat{\gamma} - \gamma) + o_p(1/\sqrt{m}), \tag{9}$$

where $o_p(1/\sqrt{m})$ denotes a random term r_m such that $\sqrt{m} r_m$ converges to 0 in probability when $m \to \infty$ and $N \to \infty$. Notice that

$$\begin{aligned}\dot{g}_{i\gamma}(\theta_i,\gamma)\,(\widehat{\gamma}-\gamma) &= \frac{1}{m}\sum_{j=1}^{m}[a_{j0}(\widehat{\gamma}_{j0}-\gamma_{j0})+a_{j1}(\widehat{\gamma}_{j1}-\gamma_{j1})]\\ &= \frac{1}{m}\sum_{j=1}^{m}[a_{j0}(\widehat{\gamma}_{j0}-\gamma_{j0})+E(a_{j1})\,(\widehat{\gamma}_{j1}-\gamma_{j1})]\\ &+\frac{1}{m}\sum_{j=1}^{m}\{[a_{j1}-E(a_{j1})]\,(\widehat{\gamma}_{j1}-\gamma_{j1})\}.\end{aligned} \tag{10}$$

Because

$$\frac{1}{\sqrt{m}}\sum_{j=1}^{m}[a_{j1}-E(a_{j1})]$$

converges to a normal distribution according to the central limit theorem, it follows from Eqs. 8 and 10 that

$$\sqrt{m}\dot{g}_{i\gamma}(\theta_i,\gamma)\,(\widehat{\gamma}-\gamma)=\frac{1}{\sqrt{m}}\sum_{j=1}^{m}[a_{j0}(\widehat{\gamma}_{j0}-\gamma_{j0})+E(a_{j1})\,(\widehat{\gamma}_{j1}-\gamma_{j1})]+o_p(1). \tag{11}$$

Let

$$a_{i\theta}=-\lim_{m\to\infty}\dot{g}_{i\theta}(\theta_i,\gamma).$$

It follows from Eqs. 9 and 11 that

$$\begin{aligned}\sqrt{m}(\widehat{\theta}_i-\theta_i) &= a_{i\theta}^{-1}[\sqrt{m}g_i(\theta_i,\gamma)+\dot{g}_{i\gamma}(\theta_i,\gamma)\sqrt{m}(\widehat{\gamma}-\gamma)]+o_p(1)\\ &= a_{i\theta}^{-1}\{\sqrt{m}g_i(\theta_i,\gamma)+\frac{1}{\sqrt{m}}\sum_{j=1}^{m}[a_{j0}(\widehat{\gamma}_{j0}-\gamma_{j0})+E(a_{j1})\,(\widehat{\gamma}_{j1}-\gamma_{j1})]\}+o_p(1)\\ &= a_{i\theta}^{-1}\sqrt{m}g_i(\theta_i,\gamma)+\frac{\boldsymbol{a}_{i\gamma}}{\sqrt{m}}(\widehat{\gamma}-\gamma)o_p(1),\end{aligned} \tag{12}$$

where

$$\boldsymbol{a}_{i\gamma}=(a_{10},E(a_{11});a_{20},E(a_{21});\ldots;a_{m0},E(a_{m1}))'.$$

Because $\boldsymbol{a}_{i\gamma}\neq\mathbf{0}$, the distribution and standard error of $\hat{\gamma}$ do affect those of $\hat{\theta}_i$. Since a_{j0} and $E(a_{j1})$ are nonstochastic, the two terms in the braces of Eq. 12 are independent. From Eq. 8 we get

$$\frac{\boldsymbol{a}_{i\gamma}}{\sqrt{m}}(\widehat{\gamma}-\gamma)\xrightarrow{\mathscr{L}}N(0,\tau^2),$$

where

$$\tau^2 = \lim_{N\to\infty} \boldsymbol{a}'_{i\gamma}\boldsymbol{\Omega}\boldsymbol{a}_{i\gamma}/(mN).$$

It follows from Eq. 12 that

$$\sqrt{m}(\widehat{\theta}_i - \theta_i) \xrightarrow{\mathscr{L}} N(0, \omega_i^2), \tag{13}$$

where

$$\omega_i^2 = a_{i\theta}^{-2}(\upsilon_{i11} + \tau^2)$$

and

$$\upsilon_{i11} = \lim_{m\to\infty} \frac{1}{m}\sum_{j=1}^{m} \gamma_{j1}^2 E[u_{\mathrm{ij}} - P(\theta_i, \gamma_{j0}, \gamma_{j1})]^2.$$

Consistent estimates of $a_{i\theta}$, υ_{i11} and τ^2 are obtained using

$$\widehat{a}_{i\theta} = -\frac{1}{m}\sum_{j=1}^{m} P(\widehat{\theta}_i, \widehat{\gamma}_{j0}, \widehat{\gamma}_{j1})Q(\widehat{\theta}_i, \widehat{\gamma}_{j0}, \widehat{\gamma}_{j1})\widehat{\gamma}_{j1}^2,$$

$$\widehat{\upsilon}_{i11} = \frac{1}{m}\sum_{j=1}^{m} \widehat{\gamma}_{j1}^2 [u_{ij} - P(\widehat{\theta}_i, \widehat{\gamma}_{j0}, \widehat{\gamma}_{j1})]^2,$$

$\widehat{\tau}^2 = \widehat{\boldsymbol{a}}'_{i\gamma}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{a}}_{i\gamma}/(mN)$ with $\hat{\boldsymbol{a}}_{i\gamma}$ being obtained by

$$\widehat{\boldsymbol{a}}_{i\gamma} = (\widehat{a}_{10}, \widehat{a}_{11}; \widehat{a}_{20}, \widehat{a}_{21}; \ldots; \widehat{a}_{m0}, \widehat{a}_{m1})$$

and

$$\widehat{a}_{j0} = -P(\widehat{\theta}_i, \widehat{\gamma}_{j0}, \widehat{\gamma}_{j1})Q(\widehat{\theta}_i, \widehat{\gamma}_{j0}, \widehat{\gamma}_{j1})\widehat{\gamma}_{j1},$$

$$\widehat{a}_{j1} = [u_{ij} - P(\widehat{\theta}_i, \widehat{\gamma}_{j0}, \widehat{\gamma}_{j1})] - P(\widehat{\theta}_i, \widehat{\gamma}_{j0}, \widehat{\gamma}_{j1})Q(\widehat{\theta}_i, \widehat{\gamma}_{j0}, \widehat{\gamma}_{j1})\widehat{\gamma}_{j1}\widehat{\theta}_i.$$

In summary, the standard error of $\hat{\theta}_i$ is given by:

$$\sqrt{\frac{\omega_i^2}{m}} = \sqrt{\frac{a_{i\theta}^{-2}\upsilon_{i11}}{m} + \frac{a_{i\theta}^{-2}\tau^2}{m}}. \tag{14}$$

A consistent estimate of $\hat{\boldsymbol{\Omega}}$ depends on the method and sample that are used when estimating γ. When the model is correct (i.e. when Eq. 1 is the true distribution of U_{ij}) and the local independence assumption holds, $a_{i\theta}=\upsilon_{i11}$ and $a_{i\theta}^{-2}\upsilon_{i11}/m$ is just the inverse of test information. Eq. 14 therefore implies that, when $\hat{\gamma}$ instead of γ is used, the standard error of $\hat{\theta}_i$ will be positively affected by that of $\hat{\gamma}$. The effect depends on the magnitude of $\boldsymbol{a}_{i\gamma}\boldsymbol{\Omega}\boldsymbol{a}_{i\gamma}'/m$ and N.

In the following section, we will give a detailed description of a simulation study, which shows how the upward-corrected SE offered by Eq. 14 better reflects the reality.

Simulation and Results

The simulation study consists of three parts: item calibration, subject scoring, and the computation of $\hat{\boldsymbol{\Omega}}$:

1. A calibration sample of size N is first generated following a certain distribution (in this study $\mathcal{N}(0, 1)$), and their responses to m items are simulated following Eq. 1. Calibration samples of two different sizes are simulated: $N = 2,000$ and $N = 200$. The m items are 40 retired items from a national math test, whose parameter estimates are available from previous large-scale calibration. These estimates are used as true parameter values. Descriptive statistics for these item parameters are summarized in Table 3. In the end, an $N \times m$ matrix of 0's and 1's is generated.

   BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003) is used for item calibration. The output from this step is the calibrated item parameters for the $m = 40$ items. Note that the 2PL model is used, so in total we have 80 item parameter estimates.

2. A scoring sample of size $n = 1,800$ is generated by having 200 subjects at each of the nine, equally-spaced θ points (step-size = 0.5) from −2.0 to 2.0. Their responses to the $m = 40$ items are simulated following Eq. 1 as well. The item parameter estimates from Step 1 are adopted to score the 1,800 subjects using maximum likelihood estimation. The ceiling and floor cases (those whose response sequence is either all 1 or all 0's) are assigned scores of +4.0 or −4.0, respectively. Empirical conditional SEs are obtained within each local θ group of 200 subjects. Note that the empirical SEs are based on $\hat{\theta}_i$s that are "contaminated" by the imprecise item parameter estimates from Step 1.

3. It is clear from the derivation that the corrective procedure requires the estimate of $\boldsymbol{\Omega}$, where $\boldsymbol{\Omega}/N$ is the asymptotic covariance matrix associated with the maximum marginal likelihood estimation (MMLE) of item parameters. To obtain $\hat{\boldsymbol{\Omega}}$, we use $\boldsymbol{\Omega} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, where $\mathbf{A}$ is the negative expected second derivative of the MMLE for a single case, and $\mathbf{B}$ is the expected cross-product of the first derivative of the MMLE for a single case. Note that when the model is correct, $\mathbf{A} = \mathbf{B}$, and consequently, $\boldsymbol{\Omega} = \mathbf{B}^{-1}$. In practice, $\mathbf{A}$ is estimated by the negative average second derivative of the MMLE, and $\mathbf{B}$ by the average cross-product of the first derivative of the MMLE.

In our simulation study, the numerical differences between the two estimates, $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$, only appear at the second decimal places. Therefore $\hat{\mathbf{B}}^{-1}$ is used in subsequent computation in place of $\hat{\boldsymbol{\Omega}}$ for easier implementation. Cai (2008) discussed a supplemented expectation-maximization (EM) algorithm to compute the information matrix of the item parameter estimates, which could be used as well. Another possibility to come up with $\hat{\boldsymbol{\Omega}}$ is to use a parametric bootstrap.

Table 1 contains the conditional SE of $\hat{\theta}_i$s when the calibration sample size is $N = 2,000$. The second column represents the empirical SE obtained in Step 2. The third column gives the average SE computed using test information among the 2,000 subjects in the same local θ group. The test information is calculated by plugging the item parameter estimates from Step 1 and $\hat{\theta}_i$s in Step 2 into Eq. 4. The fourth column provides the average upward corrected SE in each local θ group, following Eq. 14. The fifth column is the ratio of the information-based SE estimates over the empirical ones. The last column is the ratio of the upward corrected SE estimates over the empirical ones. Clearly, the information-based SE estimates are quite precise when θ ∈ [−1.5, 0.5], where their ratios to the empirical SE estimates are around 1. There are some cases showing overestimation of the SE, but the magnitude is very small (in second decimal places). However, there is serious underestimation of SE when θ is not in the middle range of the ability continuum. For example, when θ = −2, $SE_{uncorrected}/SE_{empirical} = 0.79$, suggesting an error of about 21%; when θ = 2, $SE_{uncorrected}/SE_{empirical} = 0.64$, suggesting an error of about 36%. By contrast, the upward-corrected SE estimates fare much better at these extreme locales: the errors in the corresponding cases are 13% and 6%, respectively. When θ ∈ [−1.5, 0:5], the upward correction also tends to overestimate the SE, but also just slightly (in second decimal places).

Figure 1 shows how $|SE_{uncorrected} - SE_{empirical}|$ and $|SE_{corrected} - SE_{empirical}|$ change with θ when θ ∈ [−2, 2]. Again, it is clear that the SE estimates are more accurate in the middle range of latent trait distribution, regardless of whether the SE estimates are corrected or not. Second, the advantage of using the upward-corrected SE estimate, i.e., the one based on ω, diminishes when θ gets in the vicinity of 0, and is larger when θ deviates from the middle range of the ability distribution.

Table 2 is a replica of Table 1, except that the calibration sample size is $N = 200$. We expect to see a larger SE associated with item parameter estimates when the calibration sample size gets smaller, and consequently the SE estimates of $\hat{\theta}_{ML}$ will be further inflated compared to what is shown in Table 1. This is confirmed by our data. Across all θ levels, the empirical SEs in Table 2 are larger than, or at least equal to, those in Table 1. On the other hand, the inflation in SE estimates of θ is rather small, which suggests that θ recovery in a sense is fairly robust to the size of the calibration sample. This may be due to the fact that we are looking at the 2PL model, which is known to have less stringent requirement on calibration sample size than the 3PL model. Another possible reason is that the simulation study adopts a 40-item test, which is fairly long. The resulting SEs are therefore small.

Again, our focus is on comparing the uncorrected SE estimates with the corrected SE estimates. The last two columns of Table 2 demonstrate the same pattern as Table 1: the ω-based SE more closely matches the empirical SE. In the middle range of θ, both the pure information-based SE estimates and the upward-corrected ones are quite precise. In fact, they both tend to overestimate the SEs when θ is close to 0, but only very slightly. The advantage of upward correction is pronounced when θ moves away from the middle range. Therefore, when the goal is to identify remedial or advanced students, it is important to recognize that the traditional, purely information-based SE estimates are too liberal.

Figure 2 also shows a similar pattern to Figure 1: upward-corrected SE estimates are more accurate when the θ is not in the middle range of the ability distribution. In addition, no matter which SE estimate is adopted, it is safer to use it when θ are in the middle range, where most of the items measure the best. This, again, resonates with what has been shown repeatedly in the literature, i.e., the floor and ceiling effect.

In Figure 1 and Figure 2, one may notice a “dip” at θ = 2.0. The “dip” represents an increase in $a_{i\theta}^{-2}\upsilon_{i11}/m$, the value added to the test information (see Eq. 14). As a matter of fact, moving towards extreme θ levels, the $a_{i\theta}^{-2}\upsilon_{i11}/m$ keeps increasing, as reflected by the difference between the 2nd and 3rd columns in Table 1 and Table 2. The “dip” therefore shows actually the pronounced effect of correction at these θ levels.

It may be worth mentioning that both Table 1 and Table 2 indicate that the empirical SE estimates are not symmetric around 0: they are clearly lower with examinees from lower ability groups (i.e., $\theta < 0$), which suggests that the test measures these examinees better than the higher-achieving students. This phenomenon can be explained by the particular test used in our simulation study. Table 3 presents the summary statistics of the test. The average item difficulty is −0.25. The boxplot and histogram (Figure 3 and Figure 4) for the difficulty parameter further reveal that the test is in general on the easy side for the scoring sample, whose average θ is 0. The maximum item difficulty level is merely 0.8, while the minimum reaches down to −2.0. Therefore it is not surprising that the test measures low-ability groups better.

Summary and Discussion

In this study we propose an upward correction to the standard error estimate of $\hat{\theta}$, the latent trait estimate in item response theory. More specifically, the upward correction is provided for $\hat{\theta}_{ML}$, and it takes into account the error that is carried over from the estimation of item parameters. Our simulation study shows that the upward correction leads to better SE estimates towards the two ends of the latent trait distribution.

It is not uncommon to use item parameter estimates from pretest in scoring, and SE estimates of the final $\hat{\theta}_{ML}$ are used for high-stakes decisions such as terminating a variable-length computerized adaptive test. Hence it is important to understand that the SE estimate based on test information alone might be too small. Findings from this study suggest using the upward-corrected SE estimate for $\hat{\theta}_{ML}$, especially when the item parameter estimates are obtained from a small sample.

On the other hand, to use the corrective procedure, not only the item parameter estimates, but also their variances and covariances need to be available. Therefore we would encourage reporting variances and covariances terms along with the item parameter estimates themselves, unless the calibration sample size is very large. Intra-item variance and covariance terms are available from BILOG-MG, but not inter-item ones. We would like to investigate further if the corrective procedure still works with a reduced covariance structure, i.e., using only the intra-item terms.

In addition, we feel that the study can be extended in at least the following aspects. In this paper, we have focused on the 2PL model, an extensively used model in psychological and educational measurement. The development can be easily extended to other IRT models, such as the 3PL model or the normal ogive models. Moreover, we have assumed that the probability density function in Eq. 1 or 2 is correctly specified in this study. In case this assumption is violated, consistent SEs for $\hat{\theta}_{ML}$ can still be obtained by replacing the inverse of the information in Eq. 14 by the sandwich-type variance (see Yuan & Jennrich, 2000). In addition, the derivation and simulation study in this paper assume that the calibration sample and the scoring sample are independent. When the same sample is used for calibration first and then the item parameter estimates are employed to score the sample, the rationale of the corrective procedure offered in this paper is still valid. Only one change needs to be made to Eq. 14, which consists of only two components right now: test information, and the error from γ. When the same sample is used for both calibration and scoring, a third term, which represents the covariance of the $g_i(\theta_i, \gamma)$ and $\hat{\gamma} - \gamma$ (see Eq. 12), needs to be incorporated into Eq. 13. Finally, as pointed out by one referee, Bayesian latent trait estimates such as expected a posteriori (EAP) or modal a posterior (MAP) are also used frequently in IRT applications. The technique presented in this paper can be applied in the same fashion to Bayesian posterior probabilities, and can offer similar adjustment to, for example, the SE estimate of $\hat{\theta}_{MAP}$ with a $\mathscr{N}(0, 1)$ prior.

Acknowledgments

Both authors made equal contributions to the paper, and the order of authorship is alphabetical. The authors would like to thank the Associate Editor and two anonymous reviewers for their insightful suggestions and comments. Ke-Hai Yuan's research on this project was supported by grant 4R44CA137841 from the National Cancer Institute.

References

Baker, FB.; Kim, SH. Item response theory: Parameter estimation techniques. 2nd Eds. New York: Marcel Dekker; 2004.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In: Lord, FM.; Novick, MR., editors. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968. p. 397-472.

Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika 1981;46:443–459.

Cai L. SEM of another flavour: Two new applications of the supplemented EM algorithm. British Journal of Mathematical and Statistical Psychology 2008;61:309–329. [PubMed: 17971266]

Cover, TM.; Thomas, JA. Elements of information theory. New York: John Wiley & Sons, Inc.; 1991.

Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.

Godambe VP, Thompson ME. Estimating equations in the presence of a nuisance parameter. The Annals of Statistics 1974;2:568–571.

Gong G, Samaniego F. Pseudo-maximum likelihood estimation: Theory and applications. The Annals of Statistics 1981;9:861–869.

Hambleton, R.; Swaminathan, H. Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff Publishing; 1985.

Harwell MR, Baker FB. The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. Applied Psychological Measurement 1991;15:375–389.

Harwell MR, Baker FB, Zwarts M. Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. Journal of Educational Statistics 1988;13:243–271.

Lord, FM. Applications of item response theory to practical testing problems. Mahwah, NJ: Erlbaum; 1980.

Parke WR. Pseudo maximum likehood estimation: The asymptotic distribution. The Annals of Statistics 1986;14:355–357.

Patz RJ, Junker BW. A straightforward approach to Markov chain Monte Carlo methods for item response models. Journal of Educational and Behavioral Statistics 1999;24:146–178.

Tsutakawa RK, Johnson JC. The effect of uncertainty of item parameter estimation on ability estimates. Psychometrika 1990;55:371–390.

Yuan K-H, Jennrich RI. Estimating equations with nuisance parameters: Theory and applications. Annals of the Institute of Statistical Mathematics 2000;52:343–350.

Wingersky, MS.; Barton, MA.; Lord, FM. LOGIST user's guide. Princeton NJ: Educational Testing Service; 1982.

Zimowski, M.; Muraki, E.; Mislevy, RJ.; Bock, RD. BILOG-MG 3: Item analysis and test scoring with binary logistic models. Chicago, IL: Scientific Software. [Computer software]; 2003.

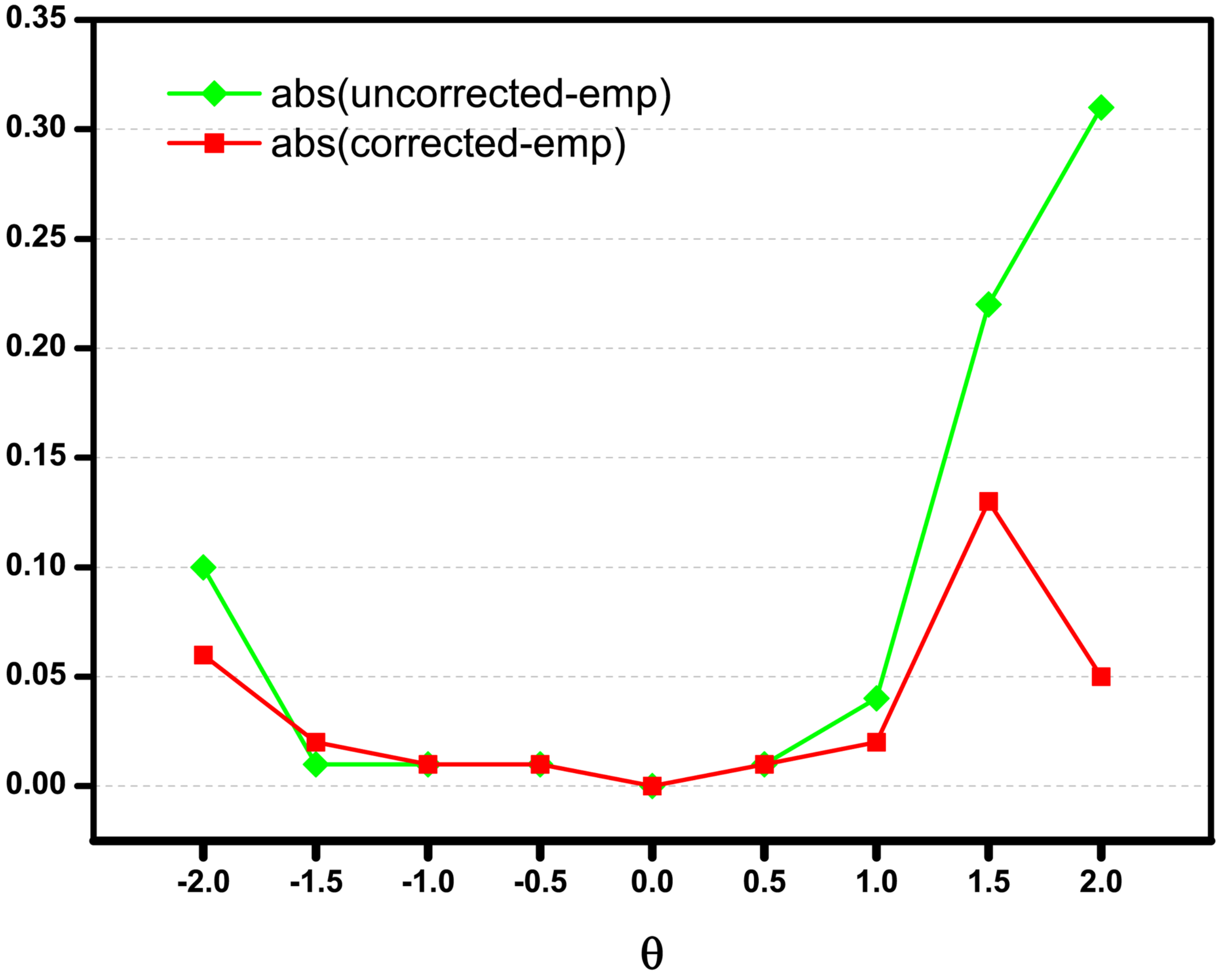

FIGURE 1.

$|SE_{empirical} - SE_{uncorrected}|$ vs. $|SE_{empirical} - SE_{corrected}|$, N = 2000

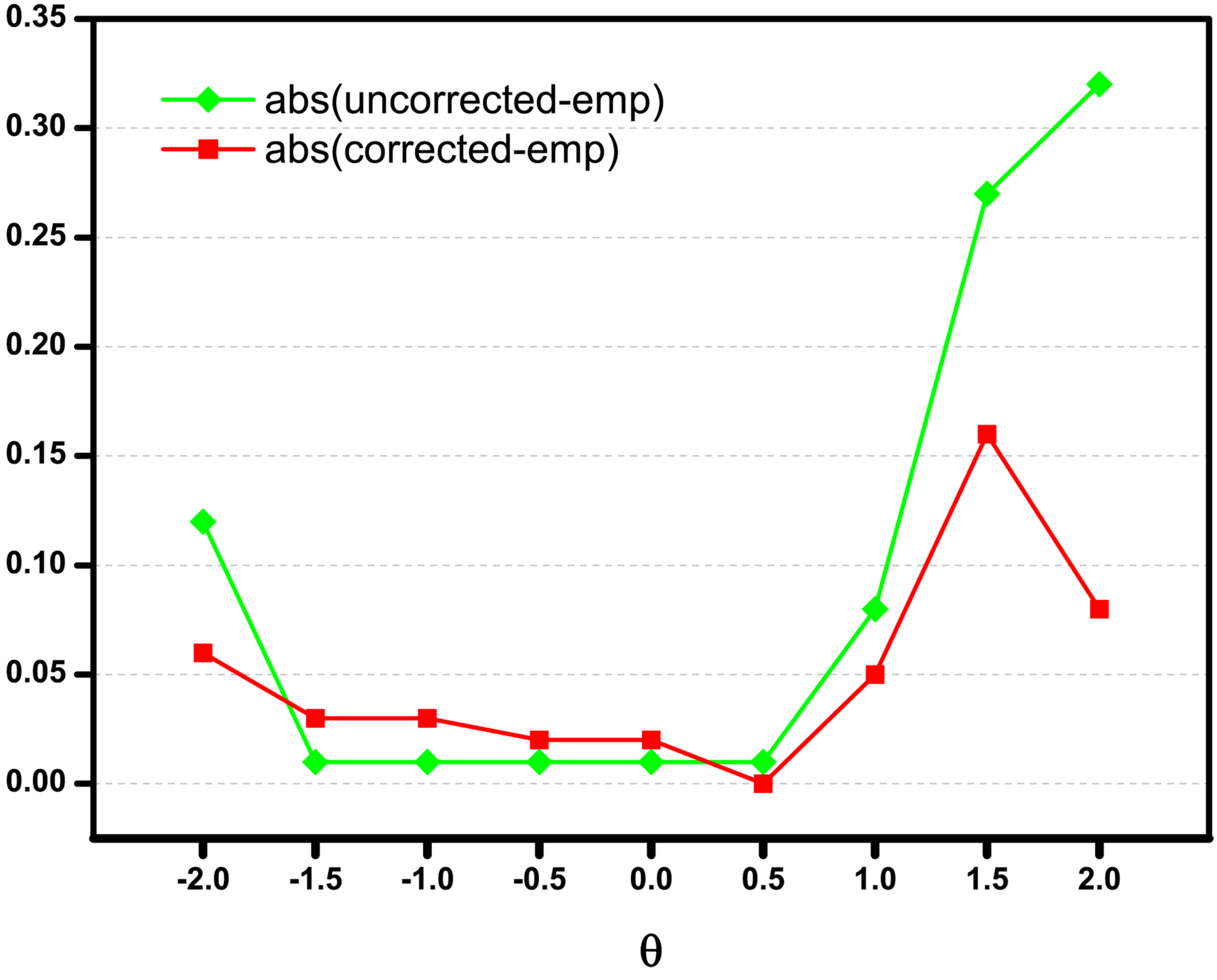

FIGURE 2.
$|SE_{empirical} - SE_{uncorrected}|$ vs. $|SE_{empirical} - SE_{corrected}|$, $N = 200$

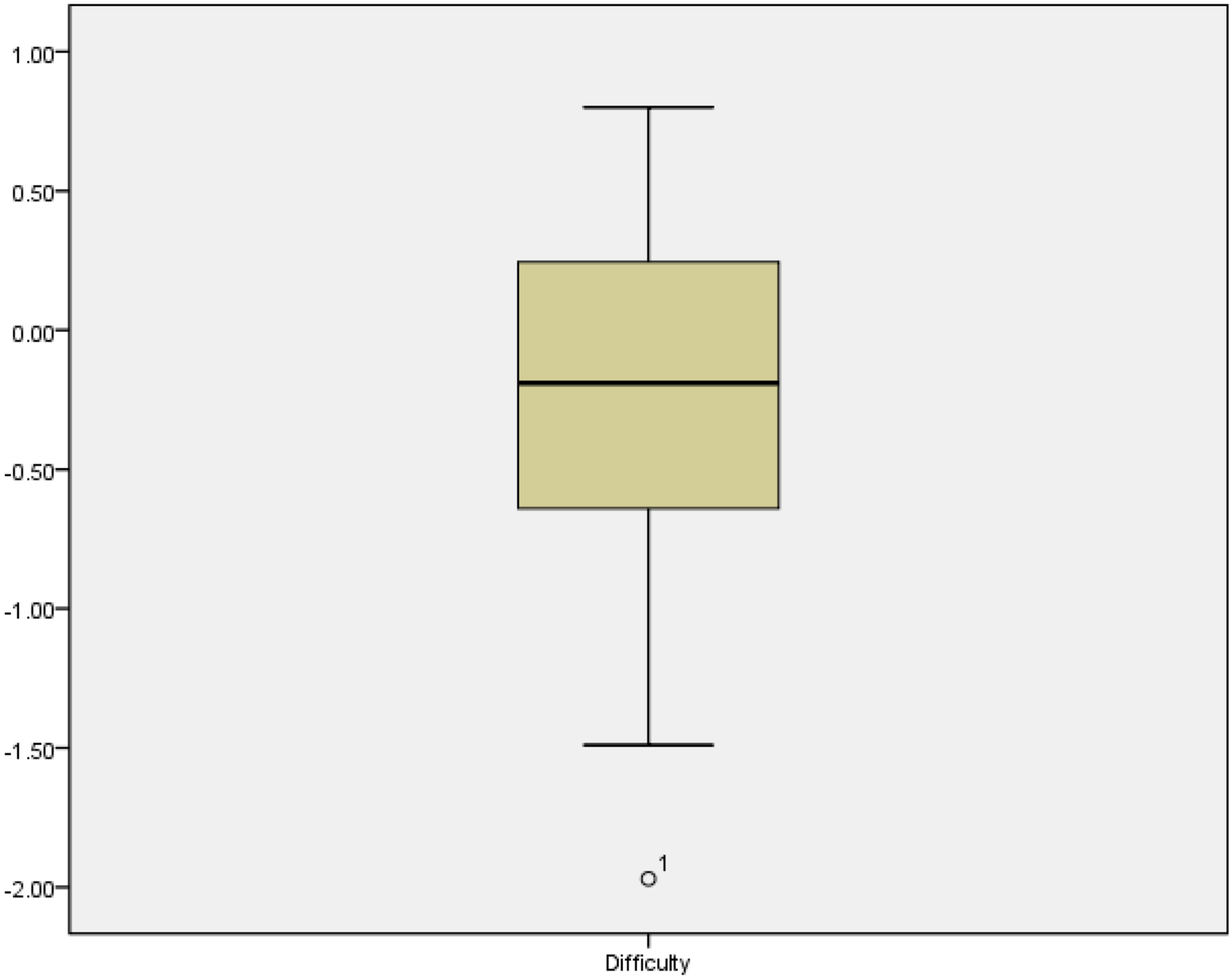

FIGURE 3.
The Boxplot of the Difficulty Parameters of the 40 Items

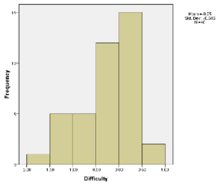

FIGURE 4.

The Histogram of the Difficulty Parameters of the 40 Items

TABLE 1

Comparison of Uncorrected and Upward-corrected SE of $\hat{\theta}$, N = 2,000

θ	**SE**	**Information-based SE**	**ω-based SE**	***Ratio*** $_{uncorrected}$	***Ratio*** $_{corrected}$
	Empirical	**Uncorrected**	**Corrected**	$SE_{uncorrected}/SE_{empirical}$	$SE_{corrected}/SE_{empirical}$
−2.0	0.49	0.39	0.43	0.79	0.87
−1.5	0.31	0.32	0.33	1.04	1.07
−1.0	0.25	0.26	0.26	1.03	1.05
−0.5	0.21	0.22	0.22	1.05	1.06
0	0.21	0.21	0.21	1.02	1.03
0.5	0.24	0.23	0.23	0.95	0.96
1.0	0.33	0.29	0.31	0.88	0.93
1.5	0.61	0.39	0.48	0.64	0.79
2.0	0.86	0.55	0.81	0.64	0.94

TABLE 2

Comparison of Uncorrected and Upward-corrected SE of $\hat{\theta}$, $N = 200$

θ	**SE**	**Information-based SE**	**ω-based SE**	$\boldsymbol{Ratio_{uncorrected}}$	$\boldsymbol{Ratio_{corrected}}$
	Empirical	**Uncorrected**	**Corrected**	$\boldsymbol{SE_{uncorrected}/SE_{empirical}}$	$\boldsymbol{SE_{corrected}/SE_{empirical}}$
−2.0	0.51	0.39	0.45	0.77	0.89
−1.5	0.31	0.32	0.34	1.04	1.12
−1.0	0.25	0.26	0.28	1.02	1.09
−0.5	0.22	0.23	0.24	1.03	1.09
0	0.21	0.22	0.23	1.03	1.08
0.5	0.25	0.24	0.25	0.96	1.01
1.0	0.38	0.30	0.33	0.78	0.86
1.5	0.67	0.40	0.51	0.60	0.76
2.0	0.86	0.57	0.81	0.64	0.91

TABLE 3

Summary Statistics of Item Parameters

	α	β
Mean	0.952	−0.252
SD	0.243	0.646