Research Report

Refinement of a Bias-Correction Procedure for the Weighted Likelihood Estimator of Ability

Jinming Zhang

Ting Lu

Research & Development

May 2007
RR-07-23

Refinement of a Bias-Correction Procedure for the Weighted Likelihood Estimator of Ability

Jinming Zhang and Ting Lu

ETS, Princeton, NJ

May 2007

Abstract

In practical applications of item response theory (IRT), item parameters are usually estimated first from a calibration sample. After treating these estimates as fixed and known, ability parameters are then estimated. However, the statistical inferences based on the estimated abilities can be misleading if the uncertainty of the item parameter estimates is ignored. Instead, estimated item parameters can be regarded as covariates measured with error. Along the line of this measurement-error-model approach, asymptotic expansions of the maximum likelihood estimator (MLE) and weighted likelihood estimator (WLE) of ability were derived by Zhang, Xie, Song, and Lu (2007). In this paper, we propose an estimator of an ability parameter based on the asymptotic formula of the WLE. A simulation study shows that the new estimator effectively reduces the bias of the MLE or WLE of ability caused by the uncertainty of the item parameter estimates not taken into account.

Key words: Bias reduction, item response theory (IRT), maximum likelihood estimator (MLE), measurement error, weighted likelihood estimator (WLE).

1 Introduction

In practical applications of item response theory (IRT), item parameters are usually estimated first from a calibration sample. After treating these estimates as fixed and known, ability parameters are then estimated and further statistical inferences are made. When item parameter estimation is sufficiently accurate, it may not be problematic to substitute the estimated item parameters for the true ones in the IRT models when estimating ability parameters. However, when the measurement errors in item parameter estimates are no longer ignorable, the statistical inferences based on such a substitution could be misleading. For instance, Tsutakawa and Johnson (1990) demonstrated that both the maximum likelihood and empirical Bayes approaches underestimate the variance of ability when the uncertainty of item parameter estimates is ignored.

Given item parameters, Lord (1983, 1986) and Samejima (1993a, 1993b) used Taylor's expansion of the likelihood equation to obtain an approximation for the bias and its standard error formulae for the maximum likelihood estimator (MLE) of ability in the context of different IRT models. Based on Lord's bias function, Warm (1989) used the weighted likelihood estimation method to estimate ability parameters and showed that the weighted likelihood estimator (WLE) is less biased than the MLE with the same asymptotic variance and normal distribution. Assuming item parameters are known, the WLE method is effective in reducing bias.

However, when item parameters are unknown and estimated item parameters are used as substitutes for the true ones in likelihood functions, as would be the case in all applications, the WLE method for ability estimation is not as effective for the 3PL case (Zhang, 2005). As a result, the measurement error in item parameter estimation must be considered as a potential contaminator of the ability estimation as well. The bias of the MLE of ability based on fixed estimated item parameters comes from two sources: (a) the bias of the MLE of ability given true item parameters, and (b) the measurement error from the uncertainty of the item parameters. Lord (1983, 1986), Warm (1989), and Samejima (1993a, 1993b) only investigated the first of these sources. Various approaches have also been proposed to address the measurement error resulting from the uncertainty of item parameters (Lewis, 1985, 2001; Mislevy, Wingersky, & Sheehan, 1994; Song, 2003; Tsutakawa & Johnson, 1990; Zhang, Xie, Song, & Lu, 2007). One of these approaches, the measurement-error-model approach, treats estimated item parameters as covariates measured with errors, instead of treating them as being fixed in nature (Song, 2003; Zhang, et al., 2007). Thus, a bias-correction formula can be developed along the line of what has been done in research on measurement error models (Stefanski & Carroll, 1985). In this paper, we propose a bias-corrected estimator of an ability parameter based on the asymptotic expansion formula of the WLE of ability. A simulation study is conducted to compare the new method with the MLE and WLE methods in terms of the bias and the root mean squared errors (RMSE) of estimated abilities. The result shows that the new estimator effectively reduces the bias in the cases considered in the simulation study.

2 The Effect of Uncertainty About Item Parameters on Ability Estimation

Suppose a test consists of n dichotomous items. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be the response vector of an examinee with $y_i = 1$ (correct) or $y_i = 0$ (incorrect) for $i = 1, \ldots, n$. The item response function (IRF) of a 3PL model is

$$P_i(\theta) = P(\theta; a_i, b_i, c_i) = P(y_i = 1|\theta) = c_i + (1 - c_i)\frac{1}{1 + \exp\{-1.7a_i(\theta - b_i)\}}, \tag{1}$$

where a_i, b_i, and c_i are the item discrimination, difficulty, and guessing parameters, respectively. Let

$$P_i^*(\theta) = \frac{1}{1 + \exp\{-1.7a_i(\theta - b_i)\}} \tag{2}$$

denote a 2PL model. Thus, $P_i(\theta) = c_i + (1 - c_i)P_i^*(\theta)$. The 3PL model is often rewritten as

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + \exp\{-1.7(a_i\theta + d_i)\}}, \tag{3}$$

where $d_i = -a_i b_i$ is the intercept parameter.

The MLE of examinee's ability is commonly used in practice (Birnbaum, 1968; Wang & Vispoel, 1998; Yi, Wang, & Ban, 2001). Under the assumptions of *local* or *conditional* independence (Lord, 1980), the likelihood function for the response vector $\mathbf{y}$ is

$$L(\mathbf{y} \mid \theta) = \prod_{i=1}^{n} P_i^{y_i}(\theta)\, Q_i^{1-y_i}(\theta), \tag{4}$$

where $Q_i(\theta) = 1 - P_i(\theta)$. If item parameters (a_i, b_i, c_i) in these models are known, the MLE $\hat{\theta}_m$ of ability is defined as the value of θ that maximizes (4). In practice, $\hat{\theta}_m$ is often found by setting the derivative of the likelihood function to zero; that is, $\hat{\theta}_m$ satisfies

$$\frac{\partial \ln L(\mathbf{y} \mid \theta)}{\partial \theta} = \sum_{i=1}^{n} \left(\frac{y_i - P_i(\theta)}{P_i(\theta) Q_i(\theta)} \right) P_i^{'}(\theta) = 0, \tag{5}$$

where $P_i^{'}(\theta)$ is the first derivative of $P_i(\theta)$ with respect to θ (see Lord, 1980). Since $P_i^{'}(\theta) = 1.7 a_i P_i^*(\theta) Q_i(\theta)$, the likelihood equation (5) becomes

$$\sum_{i=1}^{n} a_i K_i(\theta)(y_i - P_i(\theta)) = 0, \tag{6}$$

where

$$K_i(\theta) = K(\theta; a_i, b_i, c_i) = \frac{P_i^*(\theta)}{P_i(\theta)} = \frac{1}{1 + c_i \, \exp\{-1.7 a_i(\theta - b_i)\}}.$$

Let

$$I(\theta) = \sum_{i=1}^{n} \frac{(P_i^{'}(\theta))^2}{P_i(\theta) Q_i(\theta)}$$

be the Fisher test information function. The variance of the MLE of θ is $\mathrm{Var}(\hat{\theta}) = 1/I(\hat{\theta})$. After some calculations,

$$I(\theta) = 1.7^2 \sum_{i=1}^{n} a_i^2 (1 - c_i) P_i^*(\theta) Q_i^*(\theta) K_i(\theta), \tag{7}$$

where $Q_i^*(\theta) = 1 - P_i^*(\theta)$.

The likelihood function is strictly increasing or decreasing for an all-correct-response pattern (i.e., a perfect score) or an all-incorrect-response pattern (i.e., a zero score). Thus, the MLE of ability corresponding to a perfect score or a zero score is $+\infty$ or $-\infty$. Bayes estimators of ability corresponding to perfect scores and zero scores can be finite if an informative prior distribution of ability is appropriately used. This is a major reason why a Bayesian method is sometimes preferred. In practice, examinees with perfect scores or zero scores are usually assigned the highest or lowest scores, such as an 800 or a 200 in SAT® subject tests. A Bayesian method basically also gives a fixed value for each of the two extreme cases given a fixed prior distribution. In effect, any value can be a reasonable estimate of abilities of examinees with perfect scores as long as the value is at least as large as the ability estimates of all other examinees, regardless of estimation methods. Similarly, a reasonable estimate of ability with a zero score should be no larger than the ability estimates of all other examinees. Therefore, the shortcoming of the MLE of ability corresponding to perfect scores and zero scores can be easily overcome by constraining the range of ability on a closed, but large enough, interval, say $[-4, 4]$, so that the MLE of ability for a perfect score or a zero score is the upper or lower endpoint of the interval (see Lord, 1983; Zhang, 2005). Note that this paper will not further consider the bias of the ability estimates in these extreme cases.

Given item parameters, Lord (1983) obtained the following bias function for the MLE of θ:

$$B(\theta) = \frac{1.7}{I^2(\theta)} \sum_{i=1}^{n} a_i I_i(\theta) \left(P_i^*(\theta) - \frac{1}{2} \right), \tag{8}$$

where $I_i(\theta)$ is the item information function of item i, that is,

$$I_i(\theta) = \frac{(P_i^{'}(\theta))^2}{P_i(\theta) Q_i(\theta)} = 1.7^2 a_i^2 (1 - c_i) P_i^*(\theta) Q_i^*(\theta) K_i(\theta).$$

The MLE with Lord bias-correction (MLE-LBC) of θ is defined as

$$\hat{\theta}_c = \hat{\theta}_m - B(\hat{\theta}_m).$$

The bias of $\hat{\theta}_c$, BIAS($\hat{\theta}_c$), is $o(n^{-1})$ (i.e., $\lim_{n \to \infty} n\text{BIAS}(\hat{\theta}_c) = 0$) while BIAS($\hat{\theta}_m$) is $O(n^{-1})$ (i.e., nBIAS($\hat{\theta}_m$) are bounded for all n) under the assumption that the true values of the item parameters are known.

Based on Lord's work, Warm (1989) proposed the weighted likelihood estimation method and showed that the WLE of ability is less biased than the MLE with the same asymptotic variance and normal distribution. The WLE $\hat{\theta}_w$ is defined as the value of θ that maximizes

$$f(\theta) L(\mathbf{y} \mid \theta) = f(\theta) \prod_{i=1}^{n} P_i^{y_i}(\theta)\, Q_i^{1-y_i}(\theta),$$

where $f(\theta)$ is a suitably chosen function satisfying

$$\frac{\partial \ln f(\theta)}{\partial \theta} = -B(\theta) I(\theta).$$

Therefore, $\hat{\theta}_w$ satisfies the following weighted likelihood equation,

$$\frac{\partial \ln[f(\theta) L(\mathbf{y} \mid \theta)]}{\partial \theta} = \sum_{i=1}^{n} \left(\frac{y_i - P_i(\theta)}{P_i(\theta) Q_i(\theta)} \right) P_i^{'}(\theta) - B(\theta) I(\theta) = 0.$$

That is,

$$1.7 \sum_{i=1}^{n} a_i K_i(\theta) [y_i - P_i(\theta)] - B(\theta) I(\theta) = 0. \tag{9}$$

In reality, both item and ability parameters are unknown. As mentioned in the previous section, it is a common practice to estimate item parameters first and then treat the estimates as if they were the true quantities in estimating ability parameters. That is, the MLE of an ability parameter is obtained by assuming estimated item parameters $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ are fixed as substitutes for true parameters. Thus, $\hat{\theta}_m$ satisfies

$$\sum_{i=1}^{n} \hat{a}_i \hat{K}_i(\theta)(y_i - \hat{P}_i(\theta)) = 0 \tag{10}$$

instead of (6), where $\hat{P}_i(\theta) = P(\theta; \hat{a}_i, \hat{b}_i, \hat{c}_i)$ and $\hat{K}_i(\theta) = K(\theta; \hat{a}_i, \hat{b}_i, \hat{c}_i)$, while $\hat{\theta}_w$ satisfies

$$1.7 \sum_{i=1}^{n} \hat{a}_i \hat{K}_i(\theta)[y_i - \hat{P}_i(\theta)] - \hat{B}(\theta)\hat{I}(\theta) = 0 \tag{11}$$

instead of (9). The MLE, or the WLE, based on these fixed estimated item parameters will converge to some value, say θ^*, according to large sample theory under proper regularity conditions, when the number of items becomes larger and larger. However, θ^* will not necessarily be the true ability parameter θ. Thus, the WLE and MLE-LBC methods actually try to reduce the "bias" against θ^*, not the bias against the true θ, since these methods just aim to reduce the bias of MLE given item parameters.

In order to correct the bias properly, uncertainty about item parameters or errors of estimated item parameters should be also considered. Specifically, item parameter estimators can be regarded as covariates measured with error. Suppose that item parameters are estimated using a calibration sample with J examinees. Let $\hat{a}_i$, $\hat{b}_i$, $\hat{c}_i$, and $\hat{d}_i$ be the item parameter estimators. Note that these estimators are related to J. The label J is usually suppressed in these and other related quantities for convenience, unless necessary. Let

$$E(\hat{a}_i) = a_i + \delta_{ai}, \quad E(\hat{b}_i) = b_i + \delta_{bi}, \quad E(\hat{c}_i) = c_i + \delta_{ci},$$

$$\mathrm{Var}(\hat{a}_i) = \sigma_{ai}^2, \quad \mathrm{Var}(\hat{b}_i) = \sigma_{bi}^2, \quad \mathrm{Var}(\hat{c}_i) = \sigma_{ci}^2, \tag{12}$$

$$\mathrm{Cov}(\hat{a}_i, \hat{b}_i) = \sigma_{abi}, \quad \mathrm{Cov}(\hat{b}_i, \hat{c}_i) = \sigma_{bci}, \quad \mathrm{Cov}(\hat{a}_i, \hat{c}_i) = \sigma_{aci}, \tag{13}$$

where δ_{ai}, δ_{bi}, and δ_{ci} are the biases of corresponding item parameter estimators; σ_{ai}, σ_{bi}, and

σ_{ci} are the corresponding standard errors; and σ_{abi}, σ_{bci}, and σ_{aci} are the covariances of item parameter estimators. In other words, item parameter estimators are measured with error,

$$\begin{aligned} \hat{a}_i &= a_i + \delta_{ai} + \varepsilon_{ai}, \\ \hat{b}_i &= b_i + \delta_{bi} + \varepsilon_{bi}, \\ \hat{c}_i &= c_i + \delta_{ci} + \varepsilon_{ci}, \end{aligned}$$

where $\{(\varepsilon_{ai}, \varepsilon_{bi}, \varepsilon_{bi})\}$ is an independent sequence of random vectors[1] with mean zero and covariance matrix

$$\begin{pmatrix} \sigma_{ai}^2 & \sigma_{abi} & \sigma_{aci} \\ \sigma_{abi} & \sigma_{bi}^2 & \sigma_{bci} \\ \sigma_{aci} & \sigma_{bci} & \sigma_{ci}^2 \end{pmatrix}.$$

The theorem below requires the *regularity* conditions. These conditions and their explanations or justifications will be presented first.

Regularity Conditions

(C0) Item parameters a_i and b_i are uniformly bounded and c_i is bounded away from 1. θ is a bounded variable.

(C1) There exists n_0 such that for any $n > n_0$,

$$\lim_{J\to\infty} \sigma_n^2 = 0,$$

where

$$\sigma_n^2 = \max_{1\le i\le n} \{\sigma_{ai}^2, \sigma_{bi}^2, \sigma_{ci}^2, \delta_{ai}^2, \delta_{bi}^2, \delta_{ci}^2\}.$$

(C2)

$$\lim_{J\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}[(\hat{a}_i - a_i)^2] = 0, \quad \lim_{J\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}[(\hat{b}_i - b_i)^2] = 0,$$

$$\lim_{J\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}[(\hat{a}_i - a_i)(\hat{b}_i - b_i)] = 0, \quad \lim_{J\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}[(\hat{c}_i - c_i)^2] = 0,$$

$$\lim_{J\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}[(\hat{a}_i - a_i)(\hat{c}_i - c_i)] = 0, \quad \lim_{J\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}[(\hat{b}_i - b_i)(\hat{c}_i - c_i)] = 0.$$

(C3) $(\hat{a}_i - a_i)/\sigma_{ai}$, $(\hat{b}_i - b_i)/\sigma_{bi}$, and $(\hat{c}_i - c_i)/\sigma_{ci}$ have uniformly bounded 4 moments.

(C4) For any fixed θ, there exists $c_0(\theta) > 0$ such that

$$\liminf_{n\to\infty} I(\theta)/n \geq c_0(\theta) > 0.$$

In effect, (C0), which is also required by Lord (1983), holds in all applications. Regularity Condition (C1) states that the biases and standard errors of item parameter estimators converge to zero when the calibration sample size tends to infinity, which means that item parameter estimation results from the calibration sample are reasonable. So is (C2). Regularity Condition (C3) is a very weak assumption under (C0). Regularity Condition (C4) should hold for all well-designed tests with reasonable IRT models when θ is bounded. In fact, it is commonly assumed. For example, Chang and Stout (1993) also required this condition when proving the asymptotic posterior normality of the latent ability.

Under the regularity conditions, Zhang et al. (2007) obtained the following asymptotic expansion results for the MLE and WLE of ability. In the following theorem, notations $o_p(\cdot)$ and $O_p(\cdot)$ are needed, so that $F_m = G_m + o_p(H_m)$ means that $(F_m - G_m)/H_m$ converges to zero in probability, and $F_m = O_p(1)$ means that $\{F_m\}$ are bounded in probability, whereas $o(\cdot)$ and $O(\cdot)$ are in regular sense (see Serfling, 1980). Let

$$L_i(\theta) = \frac{Q_i^*(\theta)}{P_i(\theta)} = \frac{1}{c_i + \exp\{1.7a_i(\theta - b_i)\}}.$$

Theorem (Zhang, Xie, Song, & Lu, 2007)

Suppose that $\hat{\theta}_m$ is the regular MLE of θ and satisfies (10) and $\hat{\theta}_w$ is the regular WLE of θ and satisfies (11), where estimated item parameters are regarded as fixed and known. Assume that Regularity Condition (C0)–(C4) hold. Then

$$\hat{\theta}_m = \theta + [J_n(\theta) + Q_n(\theta) + Z_n(\theta)]/I(\theta) + o_p\left(\max\left(\sigma_n^2, \frac{1}{\sqrt{n}}\right)\right), \tag{14}$$

and

$$\hat{\theta}_w = \theta + [J_n(\theta) + Q_n(\theta) + Z_n(\theta) - B(\theta)I(\theta)]/I(\theta) + o_p\left(\max\left(\sigma_n^2, \frac{1}{\sqrt{n}}\right)\right), \tag{15}$$

where $I(\theta)$ is the Fisher test information function given by (7), $B(\theta)$ is given by (8) and

$$\begin{aligned}
J_{n,1}(\theta) &= -1.7^2 \sum_{i=1}^{n} (\theta - b_i)(1 - c_i) P_i^*(\theta) Q_i^*(\theta) K_i(\theta) \\
&\qquad \left\{ 1.7 a_i (\theta - b_i) \left[\frac{1}{2} - P_i^*(\theta) + c_i L_i(\theta) \right] + 1 \right\} (\sigma_{ai}^2 + \delta_{ai}^2), \\
J_{n,2}(\theta) &= -1.7^3 \sum_{i=1}^{n} a_i^3 (1 - c_i) P_i^*(\theta) Q_i^*(\theta) K_i(\theta) \left[\frac{1}{2} - P_i^*(\theta) + c_i L_i(\theta) \right] (\sigma_{bi}^2 + \delta_{bi}^2), \\
J_{n,3}(\theta) &= 1.7^2 \sum_{i=1}^{n} 2 a_i (1 - c_i) P_i^*(\theta) Q_i^*(\theta) K_i(\theta) \\
&\qquad \left\{ 1.7 a_i (\theta - b_i) \left[\frac{1}{2} - P_i^*(\theta) + c_i L_i(\theta) \right] + 1 \right\} (\sigma_{abi} + \delta_{ai} \delta_{bi}), \\
J_{n,4}(\theta) &= 1.7 \sum_{i=1,\ c_i>0}^{n} a_i Q_i^*(\theta) K_i(\theta) L_i(\theta) (\sigma_{ci}^2 + \delta_{ci}^2), \\
J_{n,5}(\theta) &= 1.7 \sum_{i=1\ c_i>0}^{n} Q_i^*(\theta) K_i(\theta) \left\{ 1.7 a_i (\theta - b_i) [1 - 2 c_i L_i(\theta)] - 1 \right\} (\sigma_{aci} + \delta_{ai} \delta_{ci}), \\
J_{n,6}(\theta) &= -1.7^2 \sum_{i=1\ c_i>0}^{n} a_i^2 Q_i^*(\theta) K_i(\theta) [1 - 2 c_i L_i(\theta)] (\sigma_{bci} + \delta_{bi} \delta_{ci}), \\
J_n(\theta) &= J_{n,1}(\theta) + J_{n,2}(\theta) + J_{n,3}(\theta) + J_{n,4}(\theta) + J_{n,5}(\theta) + J_{n,6}(\theta), \\
Q_{n,1}(\theta) &= -1.7^2 \sum_{i=1}^{n} a_i (\theta - b_i)(1 - c_i) P_i^*(\theta) Q_i^*(\theta) K_i(\theta) \delta_{ai}, \\
Q_{n,2}(\theta) &= 1.7^2 \sum_{i=1}^{n} a_i^2 (1 - c_i) P_i^*(\theta) Q_i^*(\theta) K_i(\theta) \delta_{bi}, \\
Q_{n,3}(\theta) &= -1.7 \sum_{i=1, c_i>0}^{n} a_i Q_i^*(\theta) K_i(\theta) \delta_{ci}, \\
Q_n(\theta) &= Q_{n,1}(\theta) + Q_{n,2}(\theta) + Q_{n,3}(\theta), \\
Z_n(\theta) &= 1.7 \sum_{i=1}^{n} a_i K_i(\theta) (y_i - P_i(\theta)).
\end{aligned}$$

The theorem provides the error terms or biases of the naive MLE and WLE of ability obtained by treating estimated item parameters as though they were the true values while they are actually associated with measurement error. The bias is asymptotically a function of the biases $\{\delta_{ai}, \delta_{bi}, \delta_{ci}\}$ and covariance matrixes $\{\Sigma_i\}$ of item parameter estimators. Therefore, given $\{\delta_{ai}, \delta_{bi}, \delta_{ci}\}$

and $\{\Sigma_i\}$, one may calculate the values of biases of the MLE and WLE of ability using (14) and (15), respectively. One can also determine the range of the bias of the MLE or WLE of ability if the range of the biases, the variances, and the covariances of item parameter estimators are known. Thus, one can evaluate the impact of measurement errors of item parameter estimators on ability estimation and decide whether the naive MLE or WLE is accurate enough in the situation considered.

Note that the term $Q_n(\theta)/I(\theta)$ represents the component of the bias of the ability estimator that is caused by $\{\delta_{ai},\ \delta_{bi},\ \delta_{ci}\}$ only. The term $J_n(\theta)/I(\theta)$ relates to the components of the bias caused by both $\{\delta_{ai},\ \delta_{bi},\ \delta_{ci}\}$ and $\{\Sigma_i\}$, while $B(\theta)$, $I(\theta)$, and $Z_n(\theta)$ are independent from any of those quantities. Notice that $Q_{n,2}(\theta)$ can be rewritten as $Q_{n,2}(\theta) = \sum_{i=1}^{n} I_i(\theta)\delta_{bi}$, where $I_i(\theta)$ is the item information function of item i, and $\sum_{i=1}^{n} I_i(\theta) = I(\theta)$. Thus, $Q_{n,2}(\theta)/I(\theta)$ is the weighted average bias of item difficulty parameters with item information as the weights.

The theorem does not restrict the method of item parameter estimation for a calibration sample. Hence, any regular joint MLE, marginal MLE, or Bayesian estimation methods can be used to estimate item parameters before applying the theorem. However, the effectiveness of the theorem obviously depends on the accuracy of the estimation of the item parameters, the biases, the variances, and the covariances of item parameter estimators.

When applying the theorem to a practical situation, one needs the estimates of $\{\delta_{ai},\ \delta_{bi},\ \delta_{ci}\}$ and $\{\Sigma_i\}$. Usually, a calibration program provides a set of estimation results either for parameters (a_i, b_i, c_i) of (1) or for parameters (a_i, d_i, c_i) of (3). For example, PARSCALE presents an estimate of the covariance matrix of $(\hat{a}_i, \hat{d}_i, \hat{c}_i)$. Using the delta method, given the estimation results based on (3), one can obtain the results based on (1) and vice-versa. If a calibration program could not provide accurate enough estimates of $\{\Sigma_i\}$, one can always calculate the appropriate information matrix or Hessian matrix to obtain estimates of these covariance matrixes. However, estimates of δ_{ai}, δ_{bi}, and δ_{ci} are typically not directly available from a calibration program. A Monte-Carlo simulation with some replications or the bootstrap method (Efron, 1982) is needed to obtain estimates of biases of item parameter estimators in practice. For example, one may use estimated item and ability parameters to generate 100 sets of simulated response data using a setting as similar as possible to the original data and then calibrate each set of these data. The average of the discrepancies of newly estimated item parameters from the original (estimated) item parameters across 100 replications can be used as estimates for the bias of item parameter estimators. The sample covariance matrix of $(\hat{a}_i, \hat{b}_i, \hat{c}_i)$ based on the 100 replications can also be calculated and used as a substitute for the estimate of the covariance matrix of $(\hat{a}_i, \hat{b}_i, \hat{c}_i)$. Thus, $B(\theta)$, $I(\theta)$, $J_n(\theta)$, $Q_n(\theta)$, and $Z_n(\theta)$ can be replaced by their estimates, $\hat{B}(\hat{\theta})$, $\hat{I}(\hat{\theta})$, $\hat{J}_n(\hat{\theta})$, $\hat{Q}_n(\hat{\theta})$, and $\hat{Z}_n(\hat{\theta})$, respectively. Here $\hat{\theta}$ is either $\hat{\theta}_m$ or $\hat{\theta}_w$.

In this paper, we focused on $\hat{\theta}_w$ because the WLE produces slightly better results than the MLE and MLE-LBC (see Hoijtink & Boomsma, 1995; Zhang, 2005). By (11), we know that $\hat{Z}_n(\hat{\theta}_w) - \hat{B}(\hat{\theta}_w)\hat{I}(\hat{\theta}_w) = 0$. Thus, we may only need to correct the bias in $[J_n(\theta) + Q_n(\theta)]/I(\theta)$ from the corresponding WLE to further reduce the bias of the WLE of θ. That is, the bias-corrected ability parameter estimator is

$$\hat{\theta}_{wc} = \hat{\theta}_w - [\hat{J}_n(\hat{\theta}_w) + \hat{Q}_n(\hat{\theta}_w)]/\hat{I}(\hat{\theta}_w). \tag{16}$$

This estimator is called the corrected weighted likelihood estimator (CWLE), indicating that the final estimator corrects error terms based on the WLE.

3 A Simulation Study

A simulation study was conducted to compare MLE, WLE, and CWLE. Specifically, the study attempted to determine which method produces the best ability-estimation result. The estimated item parameters from the 1998 National Assessment of Educational Progress (NAEP) grade 4 reading assessment were used to generate simulated response data (Allen, Donoghue, & Schoeps, 2001). Among 60 items used in the simulation study, there are 26 2PL items and 34 3PL items. These item parameters are presented in Table 1.

The simulation study has two stages. In the first stage, item parameters are estimated from a simulated calibration sample. These estimated item parameters are used as fixed to estimate individual ability parameters of a simulated target sample in the second stage.

Table 1

Item Parameters Used in the Simulation Study

Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
1	0.623	-0.872	0.000	31	1.342	-0.457	0.175
2	0.920	1.008	0.000	32	1.110	0.148	0.244
3	1.052	1.009	0.000	33	1.228	0.259	0.247
4	0.754	0.015	0.000	34	0.951	-0.864	0.319
5	0.763	-0.284	0.000	35	1.472	1.204	0.167
6	1.025	0.107	0.000	36	1.859	0.213	0.265
7	0.647	-1.008	0.000	37	1.133	0.916	0.297
8	0.520	-1.425	0.000	38	1.374	0.307	0.269
9	0.757	-0.630	0.000	39	0.504	-0.932	0.247
10	0.832	1.118	0.000	40	1.415	0.891	0.271
11	1.123	1.057	0.000	41	2.303	0.609	0.418
12	0.814	0.306	0.000	42	0.966	-1.318	0.244
13	0.506	-1.272	0.000	43	1.029	0.327	0.300
14	0.269	-0.904	0.000	44	0.721	-1.193	0.247
15	1.172	0.645	0.000	45	0.941	0.401	0.264
16	0.877	-0.523	0.000	46	0.793	0.642	0.247
17	0.761	-1.242	0.000	47	1.032	0.507	0.248
18	0.619	-1.113	0.000	48	0.533	-0.835	0.218
19	1.154	0.645	0.000	49	1.203	0.257	0.165
20	1.536	1.192	0.000	50	1.104	-0.155	0.247
21	0.597	1.341	0.000	51	1.464	0.774	0.138
22	0.970	0.906	0.000	52	2.300	0.416	0.264
23	1.086	-0.060	0.000	53	0.562	-0.073	0.237
24	0.795	-0.238	0.000	54	0.883	-1.015	0.310
25	0.838	-0.076	0.000	55	1.261	1.084	0.206
26	1.031	-0.310	0.000	56	0.597	-0.206	0.156
27	1.506	-0.495	0.215	57	0.938	-1.691	0.294
28	0.607	0.712	0.251	58	1.414	-0.608	0.275
29	1.288	0.554	0.190	59	1.185	-0.590	0.312
30	1.798	-0.899	0.248	60	0.579	-0.688	0.276

Note. Data from the 1998 NAEP Grade 4 Reading Assessment.

The numbers of examinees in simulated calibration samples are 250, 500, and 1,000. Examinees' ability parameters were independently generated from a standard normal distribution. Based on these ability parameters and the item parameters shown in Table 1, 100 sets (for 100 replications) of calibration response data were generated using IRT method for each of the three sample sizes. Each simulated data set was used to estimate item parameters separately. In this study, a NAEP version of PARSCALE (Allen et al., 1999; Muraki & Bock, 1991) was used to estimate item parameters. Tables 2–4 present the bias of estimated item parameters based on 100 replications for sample sizes 250, 500, and 1,000, respectively. The covariance matrixes are not reported here because their sizes are too large. Each of these 300 sets of estimated item parameters will be used as fixed and known when estimating ability parameters in the next stage.

In the second stage, the MLE, WLE, and CWLE methods are used to estimate ability parameters. Since the performance of MLE, WLE, or CWLE might be different at the different ability levels, we evaluated the ability-estimation accuracy at several ability levels. That is, we compared the results from the three ability-estimation methods to determine which method gives the best ability estimation result at these ability levels. Specifically, we chose 13 ability levels in this simulation. They are -3.0, -2.5, $\ldots$, 2.5, and 3.0. Using these ability values and the item parameters in Table 1, simulated response data were generated again using IRT method. Regarding the estimated item parameters from a calibration sample in the first stage as fixed and known, $\hat{\theta}_m$ and $\hat{\theta}_w$ were obtained for each examinee in the newly simulated response data. Then, the biases, variances, and covariances of estimated item parameters obtained in the first stage were used to calculate the bias-correction term in (16), $[\hat{J}_n(\hat{\theta}_w) + \hat{Q}_n(\hat{\theta}_w)]/\hat{I}(\hat{\theta}_w)$, so as to obtain $\hat{\theta}_{wc}$. For each set of estimated item parameters, the process was repeated 100 times.

The measurement precision of $\hat{\theta}_m$, $\hat{\theta}_w$, and $\hat{\theta}_{wc}$ was evaluated by comparing the conditional bias and the RMSE at each ability level. RMSE is the square root of the average of the squared deviations of estimated parameters from the true one. Tables 5–7 and Figures 1–3 show the biases and RMSEs of $\hat{\theta}_m$, $\hat{\theta}_w$, and $\hat{\theta}_{wc}$ at each of the 13 ability levels when item parameters are estimated from calibration samples of 250, 500, and 1,000 examinees, respectively.

Table 2

Bias of Estimated Item Parameters With Calibration Sample Size 250, Based on 100 Replications

Item	a	b	c	Item	a	b	c
1	0.0225	-0.0468	0.0000	31	0.0458	-0.0051	0.0330
2	-0.0051	-0.0341	0.0000	32	-0.0429	-0.0958	-0.0162
3	-0.0269	-0.0122	0.0000	33	-0.0729	-0.0894	-0.0238
4	0.0309	-0.0633	0.0000	34	-0.0097	-0.2190	-0.0920
5	0.0159	-0.0375	0.0000	35	-0.1097	0.0286	0.0132
6	0.0029	-0.0437	0.0000	36	-0.2607	-0.1298	-0.0367
7	0.0440	-0.0343	0.0000	37	-0.1691	-0.1687	-0.0528
8	0.0531	0.0427	0.0000	38	-0.1416	-0.1409	-0.0372
9	0.0272	-0.0594	0.0000	39	0.0865	0.0689	-0.0149
10	0.0173	-0.0293	0.0000	40	-0.2276	-0.0913	-0.0334
11	-0.0060	-0.0287	0.0000	41	-0.9640	-0.3142	-0.1201
12	0.0299	-0.0488	0.0000	42	0.0271	-0.0589	-0.0189
13	0.0532	0.0286	0.0000	43	-0.0962	-0.2028	-0.0646
14	0.0857	0.1224	0.0000	44	0.0444	-0.0651	-0.0201
15	-0.0321	-0.0431	0.0000	45	-0.0303	-0.1490	-0.0337
16	0.0148	-0.0428	0.0000	46	0.0021	-0.0784	-0.0159
17	0.0633	-0.0024	0.0000	47	-0.0092	-0.0763	-0.0172
18	0.0407	-0.0471	0.0000	48	0.0661	0.0435	0.0144
19	0.0059	-0.0556	0.0000	49	0.0641	-0.0092	0.0331
20	-0.0643	-0.0098	0.0000	50	-0.0173	-0.1002	-0.0192
21	0.0333	-0.0598	0.0000	51	-0.0161	-0.0055	0.0301
22	0.0075	-0.0250	0.0000	52	-0.4895	-0.1168	-0.0377
23	-0.0034	-0.0534	0.0000	53	0.0671	-0.0647	-0.0027
24	0.0105	-0.0681	0.0000	54	-0.0231	-0.2349	-0.0822
25	0.0179	-0.0542	0.0000	55	-0.0725	-0.0435	-0.0052
26	0.0175	-0.0554	0.0000	56	0.0997	0.1442	0.0724
27	-0.0040	-0.0713	-0.0003	57	-0.0140	-0.2004	-0.0681
28	0.0866	-0.0958	-0.0127	58	-0.0977	-0.1343	-0.0505
29	-0.0025	-0.0218	0.0104	59	-0.0890	-0.2341	-0.0837
30	-0.1218	-0.1043	-0.0291	60	0.0521	-0.1053	-0.0440

Table 3

Bias of Estimated Item Parameters With Calibration Sample Size 500, Based on 100 Replications

Item	a	b	c	Item	a	b	c
1	0.0198	-0.0388	0.0000	31	0.0851	0.0180	0.0394
2	-0.0074	-0.0307	0.0000	32	0.0084	-0.0452	-0.0007
3	-0.0065	-0.0281	0.0000	33	-0.0328	-0.0624	-0.0096
4	0.0224	-0.0562	0.0000	34	-0.0346	-0.1923	-0.0752
5	0.0100	-0.0322	0.0000	35	-0.0627	0.0180	0.0115
6	-0.0051	-0.0356	0.0000	36	-0.1362	-0.0887	-0.0237
7	0.0265	-0.0335	0.0000	37	-0.0959	-0.1177	-0.0288
8	0.0232	0.0178	0.0000	38	-0.0955	-0.0986	-0.0239
9	0.0194	-0.0507	0.0000	39	0.0568	0.0543	0.0009
10	0.0079	-0.0297	0.0000	40	-0.1474	-0.0616	-0.0186
11	-0.0004	-0.0285	0.0000	41	-0.6196	-0.1671	-0.0617
12	0.0214	-0.0359	0.0000	42	0.0125	-0.0379	-0.0056
13	0.0367	0.0105	0.0000	43	-0.0631	-0.1243	-0.0441
14	0.0499	0.0597	0.0000	44	0.0206	-0.0255	-0.0056
15	-0.0217	-0.0368	0.0000	45	-0.0167	-0.1042	-0.0211
16	0.0081	-0.0361	0.0000	46	0.0100	-0.0484	-0.0051
17	0.0228	-0.0344	0.0000	47	-0.0030	-0.0491	-0.0078
18	0.0238	-0.0269	0.0000	48	0.0525	0.0701	0.0285
19	0.0177	-0.0385	0.0000	49	0.0816	0.0053	0.0298
20	-0.0234	-0.0191	0.0000	50	-0.0322	-0.0672	-0.0092
21	0.0191	-0.0592	0.0000	51	0.0255	-0.0025	0.0257
22	0.0079	-0.0231	0.0000	52	-0.3427	-0.0794	-0.0231
23	0.0013	-0.0431	0.0000	53	0.0427	-0.0052	0.0115
24	0.0114	-0.0467	0.0000	54	-0.0358	-0.1956	-0.0681
25	0.0078	-0.0440	0.0000	55	-0.0121	-0.0165	0.0016
26	0.0130	-0.0412	0.0000	56	0.0878	0.1686	0.0809
27	0.0334	-0.0426	0.0077	57	-0.0111	-0.1504	-0.0548
28	0.0553	-0.0319	0.0019	58	-0.0568	-0.0839	-0.0306
29	0.0378	-0.0182	0.0131	59	-0.0811	-0.1833	-0.0612
30	-0.0566	-0.0742	-0.0149	60	0.0334	-0.0872	-0.0283

Table 4

Bias of Estimated Item Parameters With Calibration Sample Size 1,000, Based on 100 Replications

Item	a	b	c	Item	a	b	c
1	0.0132	-0.0342	0.0000	31	0.0750	0.0020	0.0357
2	-0.0007	-0.0413	0.0000	32	0.0078	-0.0477	-0.0005
3	-0.0055	-0.0388	0.0000	33	-0.0213	-0.0454	-0.0027
4	0.0072	-0.0498	0.0000	34	-0.0302	-0.1765	-0.0667
5	0.0050	-0.0317	0.0000	35	-0.0065	-0.0145	0.0060
6	-0.0059	-0.0437	0.0000	36	-0.0674	-0.0719	-0.0125
7	0.0172	-0.0290	0.0000	37	-0.0587	-0.0871	-0.0156
8	0.0151	0.0006	0.0000	38	-0.0454	-0.0801	-0.0182
9	0.0176	-0.0424	0.0000	39	0.0367	0.0403	0.0067
10	0.0053	-0.0347	0.0000	40	-0.0768	-0.0557	-0.0110
11	-0.0068	-0.0350	0.0000	41	-0.3395	-0.0972	-0.0293
12	0.0074	-0.0367	0.0000	42	0.0096	-0.0320	-0.0004
13	0.0162	-0.0181	0.0000	43	-0.0546	-0.1030	-0.0337
14	0.0317	0.0147	0.0000	44	0.0117	-0.0314	-0.0002
15	-0.0081	-0.0421	0.0000	45	-0.0055	-0.0833	-0.0127
16	0.0042	-0.0373	0.0000	46	0.0113	-0.0500	-0.0011
17	0.0165	-0.0322	0.0000	47	-0.0071	-0.0436	-0.0045
18	0.0116	-0.0351	0.0000	48	0.0423	0.0622	0.0337
19	0.0101	-0.0451	0.0000	49	0.0589	-0.0073	0.0206
20	-0.0029	-0.0363	0.0000	50	-0.0044	-0.0575	-0.0064
21	0.0122	-0.0452	0.0000	51	0.0277	-0.0318	0.0153
22	0.0052	-0.0338	0.0000	52	-0.1896	-0.0680	-0.0138
23	0.0112	-0.0381	0.0000	53	0.0321	-0.0089	0.0147
24	0.0062	-0.0397	0.0000	54	-0.0434	-0.1845	-0.0631
25	0.0027	-0.0415	0.0000	55	0.0022	-0.0332	0.0007
26	0.0147	-0.0412	0.0000	56	0.0835	0.1716	0.0789
27	0.0634	-0.0285	0.0087	57	-0.0101	-0.1329	-0.0498
28	0.0389	-0.0122	0.0084	58	-0.0317	-0.0802	-0.0202
29	0.0602	-0.0269	0.0098	59	-0.0632	-0.1485	-0.0477
30	-0.0646	-0.0673	-0.0114	60	0.0229	-0.0792	-0.0196

Table 5

Bias and RMSE of Ability Estimates When Item Parameters Are Estimated With Calibration Sample Size 250

	Bias			RMSE		
Ability	MLE	WLE	CWLE	MLE	WLE	CWLE
-3.0	-0.0428	0.2309	0.0756	0.7396	0.6978	0.7483
-2.5	-0.1206	0.0919	-0.0160	0.7432	0.6254	0.6990
-2.0	-0.0749	0.0490	0.0000	0.6056	0.5057	0.5662
-1.5	-0.0666	-0.0064	-0.0037	0.4396	0.3871	0.4281
-1.0	-0.0555	-0.0306	0.0117	0.3239	0.3027	0.3211
-0.5	-0.0595	-0.0514	0.0128	0.2627	0.2563	0.2595
0.0	-0.0735	-0.0716	-0.0012	0.2347	0.2323	0.2196
0.5	-0.0695	-0.0743	-0.0184	0.2301	0.2283	0.2107
1.0	-0.0341	-0.0514	-0.0105	0.2477	0.2431	0.2332
1.5	0.0110	-0.0366	-0.0112	0.3369	0.3092	0.2967
2.0	0.1080	-0.0191	-0.0169	0.5420	0.4360	0.4148
2.5	0.2134	-0.0383	-0.0607	0.7094	0.5394	0.5145
3.0	0.2117	-0.1541	-0.1960	0.7176	0.5818	0.5691

Table 6

Bias and RMSE of Ability Estimates When Item Parameters Are Estimated With Calibration Sample Size 500

	Bias			RMSE		
Ability	MLE	WLE	CWLE	MLE	WLE	CWLE
-3.0	-0.0931	0.1805	0.0870	0.7282	0.6793	0.7204
-2.5	-0.1682	0.0498	-0.0105	0.7475	0.6208	0.6741
-2.0	-0.1105	0.0216	0.0010	0.6204	0.5062	0.5467
-1.5	-0.0846	-0.0203	-0.0084	0.4471	0.3887	0.4131
-1.0	-0.0573	-0.0305	0.0037	0.3231	0.3010	0.3098
-0.5	-0.0491	-0.0404	0.0053	0.2567	0.2500	0.2516
0.0	-0.0553	-0.0527	-0.0002	0.2234	0.2209	0.2150
0.5	-0.0510	-0.0552	-0.0099	0.2157	0.2132	0.2017
1.0	-0.0222	-0.0392	-0.0038	0.2319	0.2268	0.2208
1.5	0.0134	-0.0338	-0.0043	0.3202	0.2908	0.2856
2.0	0.1010	-0.0277	-0.0044	0.5254	0.4129	0.4075
2.5	0.2029	-0.0581	-0.0397	0.6985	0.5137	0.5083
3.0	0.2021	-0.1838	-0.1677	0.7110	0.5597	0.5527

Table 7

Bias and RMSE of Ability Estimates When Item Parameters Are Estimated With Calibration Sample Size 1,000

	Bias			RMSE		
Ability	MLE	WLE	CWLE	MLE	WLE	CWLE
-3.0	-0.1193	0.1554	0.0986	0.7229	0.6698	0.7055
-2.5	-0.1913	0.0281	-0.0025	0.7497	0.6196	0.6620
-2.0	-0.1299	0.0048	0.0051	0.6253	0.5060	0.5372
-1.5	-0.0984	-0.0323	-0.0082	0.4546	0.3922	0.4087
-1.0	-0.0643	-0.0367	0.0016	0.3245	0.3009	0.3044
-0.5	-0.0510	-0.0421	0.0017	0.2568	0.2499	0.2489
0.0	-0.0525	-0.0494	-0.0005	0.2217	0.2194	0.2151
0.5	-0.0491	-0.0530	-0.0061	0.2101	0.2071	0.1982
1.0	-0.0261	-0.0432	-0.0012	0.2243	0.2197	0.2144
1.5	0.0050	-0.0422	-0.0015	0.3109	0.2821	0.2787
2.0	0.0879	-0.0418	-0.0016	0.5150	0.4022	0.4003
2.5	0.1919	-0.0761	-0.0354	0.6935	0.5004	0.4979
3.0	0.1940	-0.2055	-0.1632	0.7103	0.5507	0.5395

Figures 1–3 illustrate that the MLE has negative bias at low ability levels and positive bias for high ability levels (i.e., outward bias), while the bias of the WLE has an opposite pattern. The figures also clearly show that the CWLE successfully reduced the bias in the cases considered here.

Note that the ability-estimation program used in this study searches for the maximum values of (weighted) likelihood functions only on $[-4, 4]$. This restriction may cause the irregular results at the extreme ability levels considered in this paper, especially at Level -3.

Although it reduces the bias, CWLE does not always reduce the RMSE at the ability levels considered here. It even produces slightly larger RMSEs than WLE when the true ability is at the left side of ability scale (see Tables 5–7 or Figures 1–3). The main reason may be that the error terms in (16) were evaluated at the weighted likelihood estimates, instead of being evaluated at the true ability values.

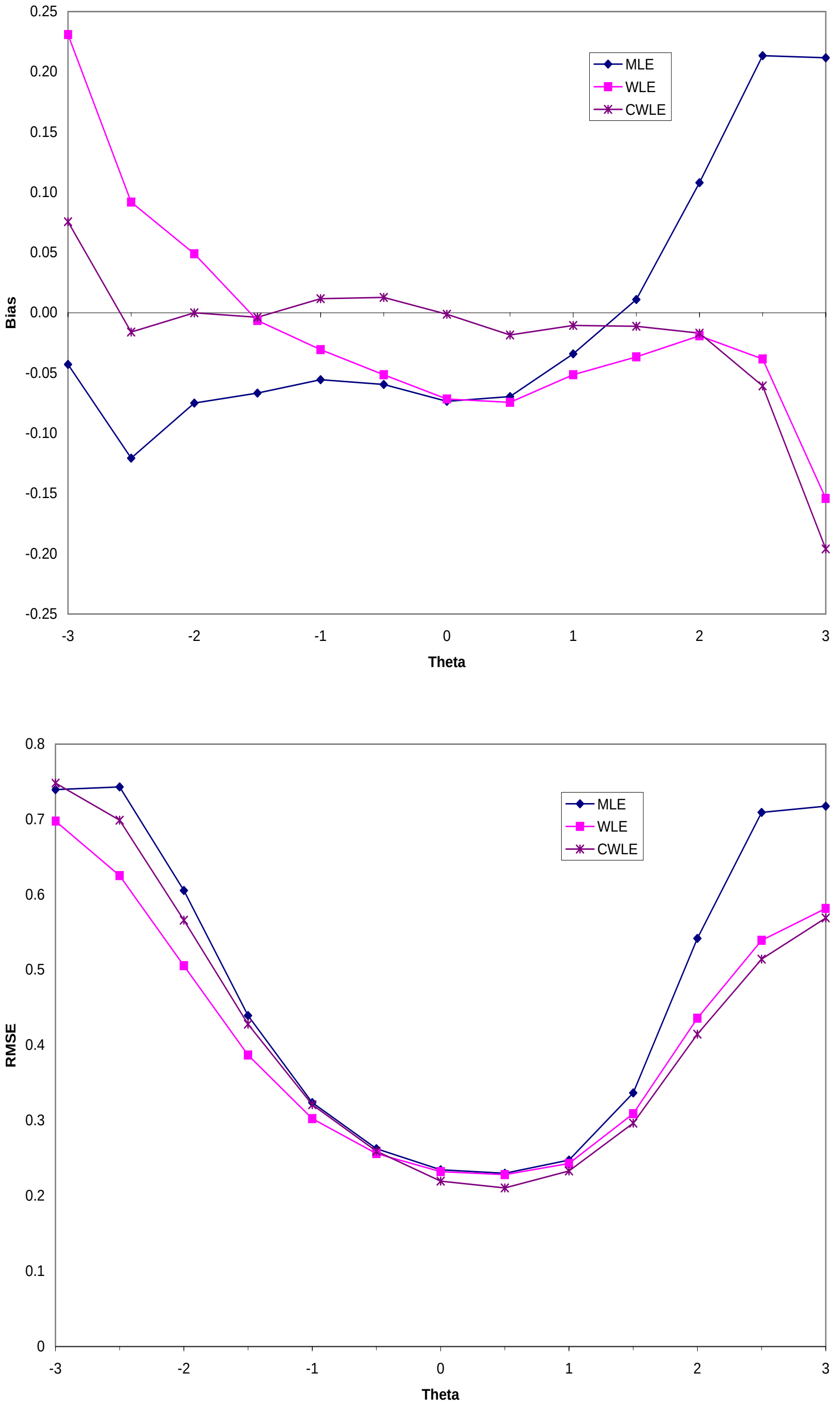

Figure 1. **Bias and RMSE of ability estimates when item parameters are estimated with calibration sample size 250.**

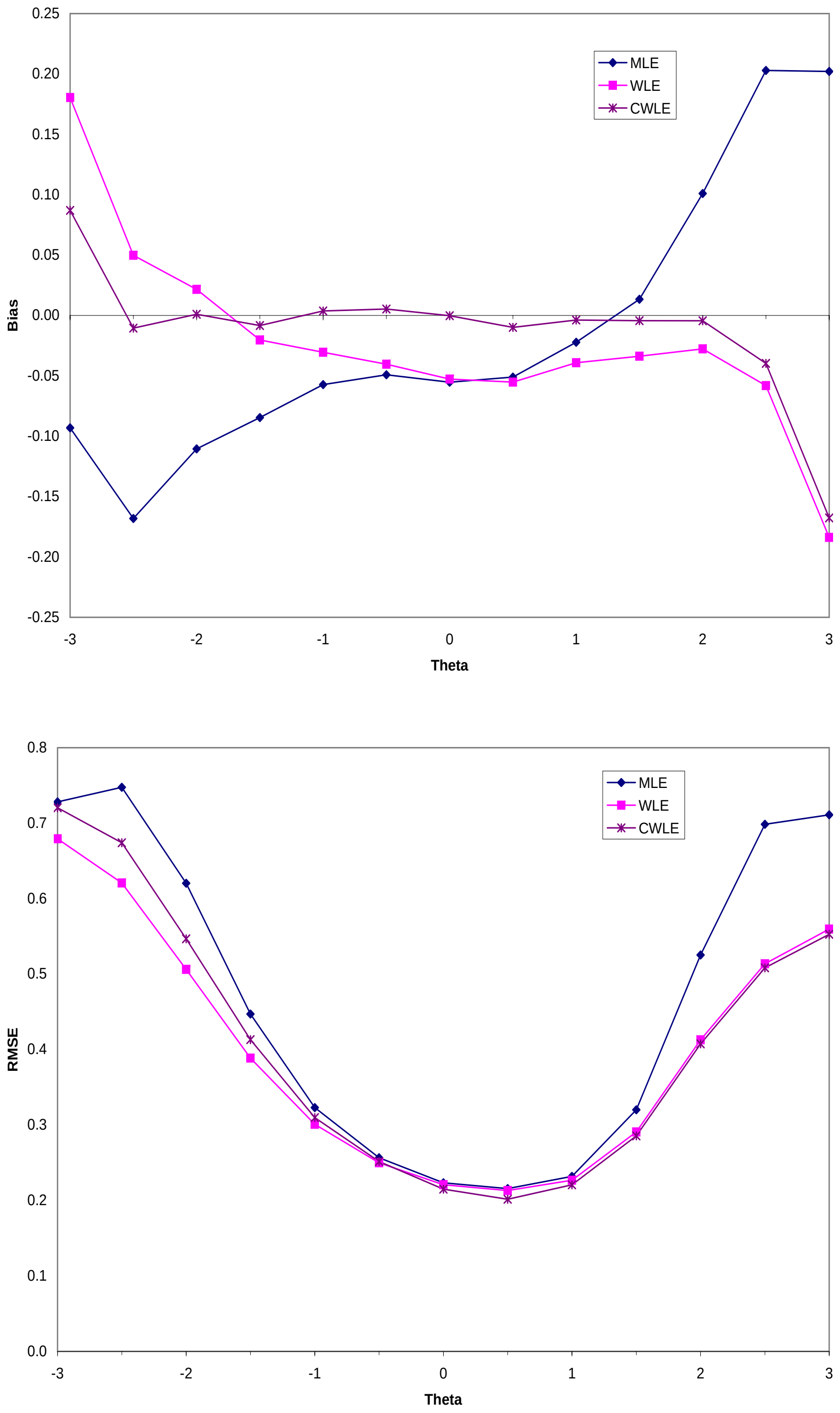

Figure 2. **Bias and RMSE of ability estimates when item parameters are estimated with calibration sample size 500.**

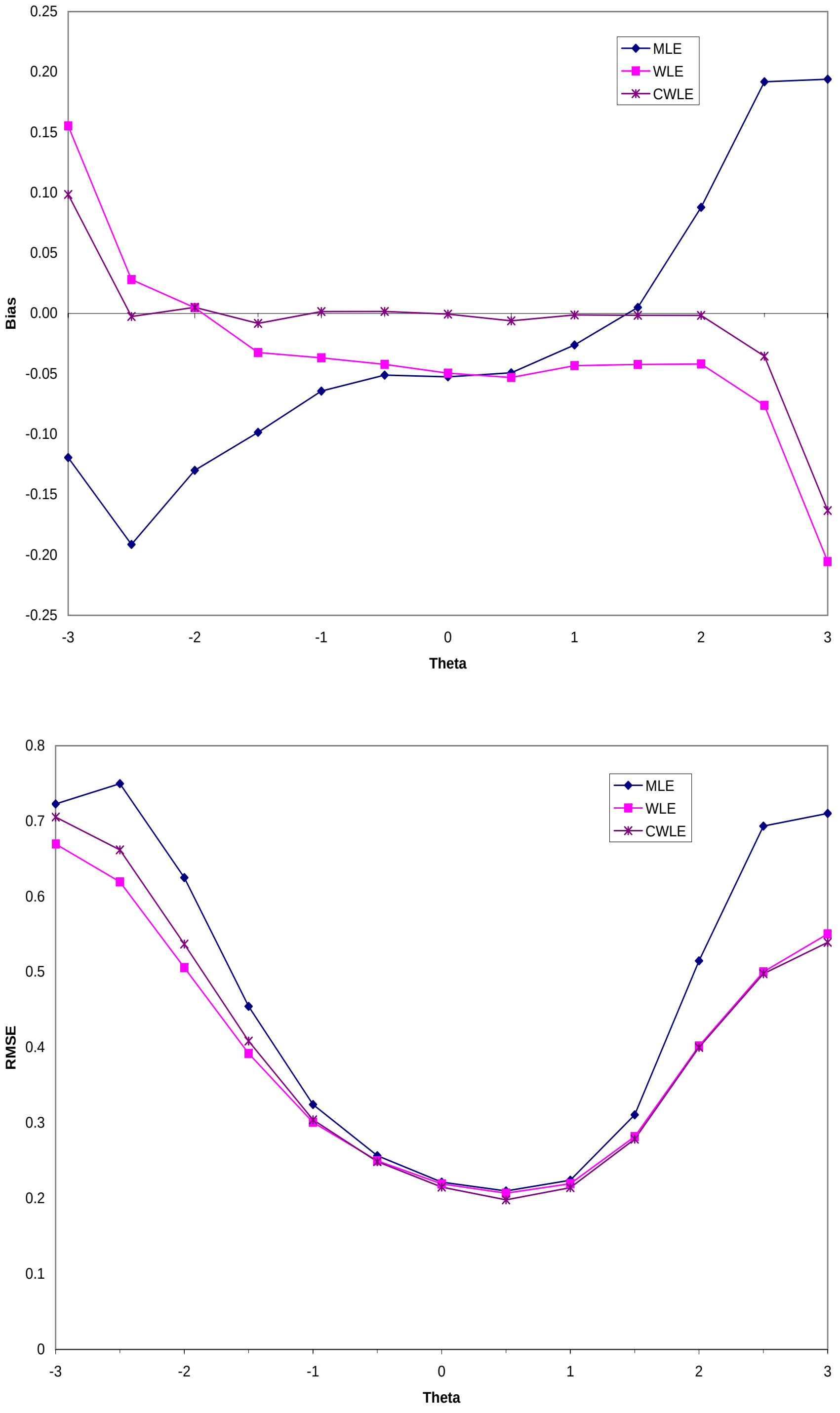

Figure 3. **Bias and RMSE of ability estimates when item parameters are estimated with calibration sample size 1,000.**

Note that the average of the difficulty parameters in Table 1 is around zero (-0.0401). Thus, $\hat{\theta}_w - \theta$ is larger when θ is away from zero and smaller when θ is near zero. Hence, noise has been added when we try to correct the bias by substituting $[\hat{J}_n(\hat{\theta}_w) + \hat{Q}_n(\hat{\theta}_w)]/\hat{I}(\hat{\theta}_w)$ for $[J_n(\theta) + Q_n(\theta)]/I(\theta)$, and the noise may not be small when θ is far away from the average of the difficulty parameters in IRT models. When θ is far away from the average of the difficulty parameters in IRT models, $\hat{\theta}_m$ or $\hat{\theta}_w$ typically has large bias or RMSE, that is, $\hat{\theta}_w$ may be too far away from θ, which violates the assumption $\sqrt{n}(\hat{\theta}_w - \theta) = O_p(1)$. To confirm this, we re-evaluated the error terms in (16) at true ability values rather than estimated ones (but item parameters and their covariance matrixes were still the estimated ones). That is,

$$\hat{\theta}_{wta} = \hat{\theta}_w - [\hat{J}_n(\theta) + \hat{Q}_n(\theta)]/\hat{I}(\theta).$$

Though it is not practical, $\hat{\theta}_{wta}$ did produce slightly, but uniformly smaller RMSE than $\hat{\theta}_w$ produced. In summary, the simulation study suggests that the CWLE is a useful alternative to $\hat{\theta}_m$ or $\hat{\theta}_w$ especially when θ is within the range of difficulty parameters.

4 Discussion

The accuracy of ability estimates is very important because estimated ability scores are the major measurement output of a test that is analyzed using IRT models. This paper tries to reduce the bias of the WLE of ability caused by treating item parameters estimated from a calibration sample as if they were true. Based on the results of the simulation study, the CWLE is effective in reducing the bias of the WLE in the cases considered here. However, CWLE does not reduce the RMSE of the ability estimator when the values of true ability parameters are far below the average of item difficulty parameters in a test. This weakness is not relevant in computerized adaptive testing (CAT), since CAT always tries to match item difficulty level with the examinee's ability level. Therefore, in CAT, CWLE can reduce not only the bias of the ability estimator but also the RMSE. In this study, we only test the CWLE method in limited cases. To determine the capacity and limitations of CWLE, further theoretical and simulation studies are needed.

It is important to note that the effectiveness of CWLE depends on the calibration program (software) used to estimate item parameters and the covariance matrixes of estimated item parameters. Hence, one should check if a calibration program can produce reasonable estimates of covariance matrixes before applying the CWLE method. The CWLE method is effective only when both the bias and the covariance matrixes of estimated item parameters are well estimated. As discussed in Section 3, the bias of estimated item parameters, which is typically not directly available from a calibration program, can be obtained by a Monte-Carlo simulation with some replications or the bootstrap method. As a by-product, one can also obtain sample covariance matrixes of estimated item parameters based on the bootstrap method. These covariance matrixes can be used to check the accuracy of the covariance matrixes provided by the calibration program and/or applied directly to (16).

Another way to deal with the uncertainty about item parameters is to make use of the expected response functions (ERFs; Lewis, 1985, 2001; Mislevy, et al., 1994). An ERF is the expectation of an IRF with respect to the posterior distributions of item parameters. This Bayesian approach takes the uncertainty about item parameters into account by substituting ERFs for IRFs in the likelihood function. Lee and Zhang (2007) to compared this method with CWLE by a simulation study. For details, see Lee and Zhang.

References

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (NCES 2001–509). Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Birnbaum, A. (1968). Some latent ability models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392–479). Reading, MA: Addison-Wesley.

Chang, H., & Stout, W. F. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika, 58*, 37–52.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia, PA: Society for Industrial and Applied mathematics.

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 67–84). New York: Springer-Verlag.

Lee, Y., & Zhang, J. (2007, April). *Comparing different approaches of bias correction for ability estimation in IRT models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

Lewis, C. (2001). Expected response functions. In A. Boomsma, M. van Duijin, & T. Snijders (Eds.), *Essays on item response theory* (pp. 163–171). New York: Springer-Verlag.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48*, 233–245.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157–162.

Mislevy, R. J., Wingersky, M. S., & Sheehan, K. M. (1994). *Dealing with uncertainty about item parameters: Expected response functions* (ETS Research Rep. No. 94-28-ONR). Princeton, NJ: ETS.

Muraki, E., & Bock, R. D. (1991). PARSCALE: Parameter scaling of rating data [Computer software]. Chicago, IL: Scientific Software, Inc.

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

Samejima, F. (1993b). The bias function of the maximum likelihood estimate of ability for dichotomous response level. *Psychometrika, 58*, 195–209.

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

Song, X. (2003). *Item parameter measurement error in item response theory models.* Unpublished doctoral dissertation, Department of Statistics, Rutgers, The State University of New Jersey, New Brunswick, NJ.

Stefanski, L. A., & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Annals of Statistics, 13*, 1335–1351.

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

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109–135.

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

Yi, Q., Wang, T., & Ban, J. (2001). Effect of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement, 38*, 267–292.

Zhang, J. (2005). *Bias correction for the maximum likelihood estimate of ability* (ETS Research Rep. No. RR-05-15). Princeton, NJ: ETS.

Zhang, J., Xie, M., Song, X., & Lu, T. (2007). *Using measurement error models to correct bias in ability estimation.* Unpublished manuscript.

Notes

[1] In Bayesian setting, item parameters are usually assumed to be independent between items, that is, $\{(a_i, b_i, c_i)\}$ is an independent sequence of random vectors (see Lewis, 2001). Lewis argued that this is almost a necessary condition. In practice, only the covariances of item parameter estimators within an item are available and the covariances of item parameter estimators between items are zero. Thus, it is not too unreasonable to assume that $\{(\varepsilon_{ai}, \varepsilon_{bi}, \varepsilon_{bi})\}$ is an independent sequence of random vectors.