Research Report

Comparing Different Approaches of Bias Correction for Ability Estimation in IRT Models

Yi-Hsuan Lee

Jinming Zhang

March 2008

ETS RR-08-13

Comparing Different Approaches of Bias Correction for Ability Estimation in IRT Models

Yi-Hsuan Lee and Jinming Zhang
ETS, Princeton, NJ

March 2008

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

The method of maximum-likelihood is typically applied to item response theory (IRT) models when the ability parameter is estimated while conditioning on the true item parameters. In practice, the item parameters are unknown and need to be estimated first from a calibration sample. Lewis (1985) and Zhang and Lu (2007) proposed the expected response functions (ERFs) and the corrected weighted-likelihood estimator (CWLE), respectively, to take into account the uncertainty regarding item parameters for purposes of ability estimation. In this paper, we investigate the performance of ERFs and of the CWLE in different situations, such as various test lengths and levels of measurement error in item parameter estimation. Our empirical results indicate that ERFs can cause the bias in ability estimation to fall within $[-0.2, 0.2]$ for all conditions, whereas the CWLE can effectively reduce the bias in ability estimation provided that it has a good foundation to start from.

Key words: Item response theory, maximum-likelihood estimator, expected response functions, measurement-error modeling, weighted-likelihood estimator, corrected weighted-likelihood estimator

1 Introduction

Traditionally, the maximum-likelihood method is applied to item response theory (IRT) models when the examinee ability parameter θ is estimated, which results in the maximum-likelihood estimator (MLE) $\hat{\theta}_m$. In that case, item parameters are usually estimated first from a calibration sample, and then the MLEs of θ's in the target sample are calculated with the estimated item parameters treated as fixed and known. When the item parameter estimation is both accurate and precise, replacing the real item parameters with the estimated ones is not unreasonable, and it is acceptable to estimate the ability parameters in the regular way. Under the assumption that the real item parameters are known, both Lord (1983) and Warm (1989) proposed corrections of $\hat{\theta}_m$ for bias based on the asymptotic expansion. It seems reasonable to assume that, if the uncertainty of item parameters does not introduce extra biases in $\hat{\theta}_m$, then these corrections should remain applicable, too.

However, it has been found that, when the estimated item parameters are used as substitutes for the real ones, neither the MLE with Lord bias-correction (MLE-LBC) nor Warm's weighted-likelihood estimator (WLE) will be as effective as they are supposed to be, especially when employing the 3PL model (Zhang, 2005). As a result, the measurement error in item parameter estimation must be considered a potential contaminator to the ability estimation process as well. Lewis (1985) came up with the idea of expected response functions (ERFs), which incorporate the uncertainty of item parameters by averaging out the noises induced by estimation in the item response functions (IRFs). Along this line, Mislevy, Wingersky, and Sheehan (1994) provided the operational procedures for applied work with ERFs. On the other hand, Song (2003) and Zhang, Xie, Song, and Lu (2007) derived the bias-correction formulas for the MLE and WLE of θ by asymptotic expansion with imperfect estimated item parameters under certain regularity conditions. This method is called measurement-error modeling. Zhang and Lu (2007) embraced the thinking behind measurement-error modeling and proposed the corrected weighted-likelihood estimator (CWLE) method.

To obtain an accurate ability estimate when employing the 3PL model, measurement error in ability estimation and in item parameter estimation clearly has to be taken into account, but

how these methods compare to each other is of concern. Accordingly, the main purpose of this study is to investigate the performance of ERFs and of the CWLE in different situations, such as various test lengths and levels of measurement error in item parameter estimation. It is also of practical interest to know when these bias-correction procedures should be applied.

2 Methods

In this section, the two bias-correction procedures are briefly introduced. Suppose there is a test with N dichotomously scored 3PL items, where X_n is the score of a randomly selected examinee on item n in the calibration sample and Y_n is the score of an examinee on item n in the target sample. (Examinees in the calibration sample are not necessarily the same group of people as those in the target sample.) Let $\beta = (a, b, c)$ denote the vector of item parameters. The 3PL IRF is defined as the probability of answering an item correctly by a randomly selected examinee with ability θ, that is,

$$F(\theta;\beta) \equiv c + (1-c)\frac{1}{1+\exp\{-1.7a(\theta-b)\}}. \tag{1}$$

2.1 Expected Response Functions

A corresponding ERF is defined as

$$F^*(\theta) \equiv \mathrm{E}_\beta[F(\theta;\beta)] = \int F(\theta;\beta)\cdot p(\beta)d\beta, \tag{2}$$

where $p(\beta) \equiv p(\beta|x)$ is the posterior distribution of β with prior knowledge from the calibration sample. This ERF will be used to replace the original IRF in the likelihood function.

Because of the integration, it is computationally intensive to take expectations whenever the ERFs are calculated. Thus, Mislevy et al. (1994) described an operational procedure as an alternative, which is the following:

1. Obtain an estimate of the posterior distribution $p(\beta_n),\ n = 1, \ldots, N$.
2. Specify a grid of J θ values across the ability range of interest. Let θ_j denote the jth grid point, $j = 1, \ldots, J$.
3. Draw K item parameter vectors from $p(\beta_n)$. Let $\beta_n^{(k)}$ be the kth such draw.

4. For each of the K sets of item parameters, determine $\mathrm{P}_{nj}^{(k)}$, the probability of a correct response to item n at θ_j, where

$$\mathrm{P}_{nj}^{(k)} = p(y_n = 1 | \theta = \theta_j, \beta_n = \beta_n^{(k)}). \tag{3}$$

5. Compute the expectation at each point θ_j by averaging the probabilities obtained in Step 4:

$$F_n^*(\theta_j) = \frac{1}{K} \sum_{k=1}^{K} \mathrm{P}_{nj}^{(k)}, \ j = 1, \ldots, J. \tag{4}$$

For the nth item, the collection of points $\{(\theta_j, F_n^*(\theta_j)) : j = 1, \ldots, J\}$ is referred to as a *nonparametric* ERF because it does not assume any parametric form. The nonparametric ERF is further approximated by a close-fitting 3PL curve F_n^{**}. The MLEs for the 3PL item parameters $\beta_n^{**} = (a_n^{**}, b_n^{**}, c_n^{**})$ that best approximate F_n^* are found by maximizing

$$\prod_{j=1}^{J} \left\{ F_n^{**}(\theta_j; \beta_n^{**})^{F_n^*(\theta_j)} \left[1 - F_n^{**}(\theta_j; \beta_n^{**}) \right]^{1 - F_n^*(\theta_j)} \right\}^{W_j} \tag{5}$$

over the J-point θ grid, where W_j is a weight that specifies the relative importance of fitting F^{**} at θ_j. The resulting 3PL approximation is referred to as a *fitted* ERF of the nth item. The likelihood function is thus determined with the fitted ERFs of all items serving as substitutes for the IRFs, and standard approaches to estimating ability parameters such as MLE and EAP can be applied immediately.

2.2 Corrected Weighted-Likelihood Estimator

The other method considered here is measurement-error modeling (Song, 2003; Zhang et al., 2007). In this method, the bias of $\hat{\theta}_m$ can be decomposed into two sources of measurement error: One is the bias of $\hat{\theta}_m$ given item parameters, and the other is the bias resulting from the uncertainty of item parameters. The former has been addressed by Lord's MLE-LBC and by Warm's WLE. If the latter can be quantified, subtracting it from either MLE-LBC or WLE should result in an unbiased (or less biased) estimate. In fact, Zhang and Lu (2007) made good use of the idea and proposed the CWLE method, which is described in the theorem that follows.

Let $F_n(\theta) \equiv F(\theta; \beta_n)$ be the probability of the nth item being answered correctly by a randomly selected examinee with ability θ. Under the assumption of local independence (Lord,

1980), the likelihood function for θ given the responses $\mathbf{y}=\{y_1,\ldots,y_N\}$ is

$$L(\theta|\mathbf{y}) = \prod_{n=1}^{N}[F_n(\theta)]^{y_n}\cdot[1-F_n(\theta)]^{1-y_n}. \tag{6}$$

If the item parameters β_n are known, the MLE $\hat{\theta}_m$ is the maximizer of Equation 6, which also satisfies the likelihood equation:

$$\frac{\partial \ln L(\theta|\mathbf{y})}{\partial\theta} = 1.7\sum_{n=1}^{N} a_n K_n(\theta)(y_n - F_n(\theta)) = 0, \tag{7}$$

where

$$K_n(\theta) = K(\theta; a_n, b_n, c_n) \equiv \frac{P_n(\theta)}{F_n(\theta)} = \frac{1}{1+c_n\exp\{-1.7a_n(\theta-b_n)\}} \tag{8}$$

and

$$P_n(\theta) \equiv \frac{1}{1+\exp\{-1.7a_n(\theta-b_n)\}} \tag{9}$$

is the 2PL model. Let $Q_n(\theta) = 1 - P_n(\theta)$.

Let $I_n(\theta)$ be the item-information function for item n,

$$I_n(\theta) = 1.7^2 a_n^2(1-c_n)P_n(\theta)Q_n(\theta)K_n(\theta), \tag{10}$$

and let

$$I(\theta) = \sum_{n=1}^{N} I_n(\theta) \tag{11}$$

be the test-information function. Given item parameters, Lord (1983) applied the asymptotic expansion to the likelihood equation and obtained the the following bias function for $\hat{\theta}_m$:

$$B(\theta) = \frac{1.7}{I^2(\theta)}\sum_{n=1}^{N} a_n I_n(\theta)\bigg(P_n(\theta) - 0.5\bigg). \tag{12}$$

The MLE-LBC of θ is defined as $\hat{\theta}_c = \hat{\theta}_m - B(\hat{\theta}_m)$.

Warm (1989) proposed the WLE based on Lord's work. The WLE, $\hat{\theta}_w$, is the maximizer of the function $f(\theta)L(\theta|\mathbf{y})$, where $f(\theta)$ is a suitable chosen function satisfying

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

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

$$\frac{\partial \ln[f(\theta)L(\theta|\mathbf{y})]}{\partial\theta} = 1.7\sum_{n=1}^{N} a_n K_n(\theta)(y_n - F_n(\theta)) - B(\theta)I(\theta) = 0. \tag{13}$$

The WLE can be shown to be less biased than the MLE with the same asymptotic variance and normal distribution.

In reality, β's are estimated in the calibration sample, and then their estimates, $\hat{\beta}$'s, are fixed as substitutes for the true β's. When the maximum-likelihood method is applied to estimate the examinees' θ's in the target sample, instead of Equation 7, $\hat{\theta}_m$ must satisfy

$$\sum_{n=1}^{N} \hat{a}_n \hat{K}_n(\theta)(y_n - \hat{F}_n(\theta)) = 0 \tag{14}$$

with $\hat{K}_n(\theta) = K(\theta; \hat{a}_n, \hat{b}_n, \hat{c}_n)$ and $\hat{F}_n(\theta) = F(\theta; \hat{a}_n, \hat{b}_n, \hat{c}_n)$. Similarly, $\hat{\theta}_w$ must satisfy

$$1.7 \sum_{n=1}^{N} \hat{a}_n \hat{K}_n(\theta)(y_n - \hat{F}_n(\theta)) - \hat{B}(\theta)\hat{I}(\theta) = 0. \tag{15}$$

To account for the uncertainty of item parameters, they are assumed to be measured with errors. Let

$$\mathrm{E}(\hat{a}_n) = a_n + \delta_{an}, \quad \mathrm{E}(\hat{b}_n) = b_n + \delta_{bn}, \quad \mathrm{E}(\hat{c}_n) = c_n + \delta_{cn},$$

$$\mathrm{Var}(\hat{a}_n) = \sigma_{an}^2, \quad \mathrm{Var}(\hat{b}_n) = \sigma_{bn}^2, \quad \mathrm{Var}(\hat{c}_n) = \sigma_{cn}^2,$$

$$\mathrm{Cov}(\hat{a}_n, \hat{b}_n) = \sigma_{abn}, \quad \mathrm{Cov}(\hat{b}_n, \hat{c}_n) = \sigma_{bcn}, \quad \text{and } \mathrm{Cov}(\hat{a}_n, \hat{c}_n) = \sigma_{acn},$$

where δ_{an}, δ_{bn}, and δ_{cn} are the biases of $\hat{a}_n$, $\hat{b}_n$, and $\hat{c}_n$, respectively. σ_{an}^2, σ_{bn}^2, σ_{cn}^2, σ_{abn}, σ_{bcn}, and σ_{acn} are elements of the variance/covariance matrix of the triplet ($\hat{a}_n$, $\hat{b}_n$, $\hat{c}_n$). No distributional assumptions are needed, but the following *regularity conditions* are required to establish the whole theory of CWLE:

(A0) Item parameters a_n and b_n are uniformly bounded, and c_n is bounded away from 1. θ is a bounded variable.

(A1) There exists n_0 such that for any $n > n_0$, $\lim_{M\to\infty} \sigma_N^2 = 0$, where M is the calibration sample size and $\sigma_N^2 = \max_{1\le n\le N}\{\sigma_{an}^2, \sigma_{bn}^2, \sigma_{cn}^2, \delta_{an}^2, \delta_{bn}^2, \delta_{cn}^2\}$.

(A2)

$$\lim_{M\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathrm{Var}[(\hat{a}_n-a_n)^2]=0, \qquad \lim_{M\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathrm{Var}[(\hat{b}_n-b_n)^2]=0,$$

$$\lim_{M\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathrm{Var}[(\hat{c}_n-c_n)^2]=0, \qquad \lim_{M\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathrm{Var}[(\hat{a}_n-a_n)(\hat{b}_n-b_n)]=0,$$

$$\lim_{M\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathrm{Var}[(\hat{a}_n-a_n)(\hat{c}_n-c_n)]=0, \quad \text{and} \quad \lim_{M\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathrm{Var}[(\hat{b}_n-b_n)(\hat{c}_n-c_n)]=0.$$

(A3) $(\hat{a}_n-a_n)/\sigma_{an}$, $(\hat{b}_n-b_n)/\sigma_{bn}$, and $(\hat{c}_n-c_n)/\sigma_{cn}$ have uniformly bounded four moments.

(A4) For any fixed θ, there exists $c_0(\theta)>0$ such that $\liminf_{N\to\infty} I(\theta)/N \geq c_0(\theta)>0$.

The details of these regularity conditions can be found in Zhang and Lu (2007).

In the following theorem, notation $o_p(\cdot)$ is needed. If $F_N = G_N + o_p(H_N)$, it means that $(F_N - G_N)/H_N$ converges to zero in probability (Serfling, 1980).

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

Suppose that $\hat{\theta}_w$ is the regular WLE of θ and satisfies Equation 15, where the estimated item parameters, $\hat{\beta}$'s, are regarded as fixed and known. Assume that the regularity conditions (A0)-(A4) hold. Then

$$\hat{\theta}_w = \theta + [J(\theta)+Q(\theta)+Z(\theta)-B(\theta)I(\theta)]/I(\theta) + o_p\left(\max\left(\sigma_N^2, \frac{1}{\sqrt{N}}\right)\right), \tag{16}$$

where $I(\theta)$ is the test-information function given by Equation 11, $B(\theta)$ is given by Equation 12, and

$$\begin{aligned}
J_1(\theta) &= -1.7^2\sum_{n=1}^{N}(\theta-b_n)(1-c_n)P_n(\theta)Q_n(\theta)K_n(\theta) \\
&\quad \{1.7a_n(\theta-b_n)[0.5-P_n(\theta)+c_nL_n(\theta)]+1\}(\sigma_{an}^2+\delta_{an}^2), \\
J_2(\theta) &= -1.7^3\sum_{n=1}^{N}a_n^3(1-c_n)P_n(\theta)Q_n(\theta)K_n(\theta)[0.5-P_n(\theta)+c_nL_n(\theta)](\sigma_{bn}^2+\delta_{bn}^2), \\
J_3(\theta) &= 1.7^2\sum_{n=1}^{N}2a_n(1-c_n)P_n(\theta)Q_n(\theta)K_n(\theta) \\
&\quad \{1.7a_n(\theta-b_n)[0.5-P_n(\theta)+c_nL_n(\theta)]+1\}(\sigma_{abn}+\delta_{an}\delta_{bn}),
\end{aligned}$$

$$J_4(\theta) = 1.7 \sum_{n=1,\ c_n>0}^{N} a_n Q_n(\theta) K_n(\theta) L_n(\theta)(\sigma_{cn}^2 + \delta_{cn}^2),$$

$$J_5(\theta) = 1.7 \sum_{n=1,\ c_n>0}^{N} Q_n(\theta) K_n(\theta)\{1.7 a_n(\theta - b_n)[1 - 2c_n L_n(\theta)] - 1\}(\sigma_{acn} + \delta_{an}\delta_{cn}),$$

$$J_6(\theta) = -1.7^2 \sum_{n=1,\ c_n>0}^{N} a_n^2 Q_n(\theta) K_n(\theta)[1 - 2c_n L_n(\theta)](\sigma_{bcn} + \delta_{bn}\delta_{cn}),$$

$$J(\theta) = J_1(\theta) + J_2(\theta) + J_3(\theta) + J_4(\theta) + J_5(\theta) + J_6(\theta),$$

$$Q_1(\theta) = -1.7^2 \sum_{n=1}^{N} a_n(\theta - b_n)(1 - c_n) P_n(\theta) Q_n(\theta) K_n(\theta)\delta_{an},$$

$$Q_2(\theta) = 1.7^2 \sum_{n=1}^{N} a_n^2(1 - c_n) P_n(\theta) Q_n(\theta) K_n(\theta)\delta_{bn},$$

$$Q_3(\theta) = -1.7 \sum_{n=1,\ c_n>0}^{N} a_n Q_n(\theta) K_n(\theta)\delta_{cn},$$

$$Q(\theta) = Q_1(\theta) + Q_2(\theta) + Q_3(\theta), \text{ and}$$

$$Z(\theta) = 1.7 \sum_{n=1}^{N} a_n K_n(\theta)(y_n - F_n(\theta)).$$

This theorem gives the error terms of bias in $\hat{\theta}_w$, which are obtained by treating $\hat{\beta}$'s estimated from a calibration sample as though they are the true values when they actually have associated statistical errors. According to Equation 16, the total bias of $\hat{\theta}_w$ is given by

$$[J(\theta) + Q(\theta) + Z(\theta) - B(\theta)I(\theta)]/I(\theta),$$

which is the sum of (a) the bias of $\hat{\theta}_w$ given $\hat{\beta}$'s as the true values, which equals $[Z(\theta) - B(\theta)I(\theta)]/I(\theta)$; and (b) the bias of substituting $\hat{\beta}$'s for β's, which is $[J(\theta) + Q(\theta)]/I(\theta)$. When the theorem is applied to a practical situation and θ is estimated by $\hat{\theta}_w$, $I(\theta)$, $J(\theta)$, $Q(\theta)$, $Z(\theta)$, and $B(\theta)$ have to be replaced by their estimates, $\hat{I}(\hat{\theta}_w)$, $\hat{J}(\hat{\theta}_w)$, $\hat{Q}(\hat{\theta}_w)$, $\hat{Z}(\hat{\theta}_w)$, and $\hat{B}(\hat{\theta}_w)$, respectively. It is clear that $\hat{Z}(\hat{\theta}_w) - \hat{B}(\hat{\theta}_w)\hat{I}(\hat{\theta}_w) = 0$ from Equation 15. Accordingly, the new bias-corrected ability estimator is defined as

$$\hat{\theta}_{wc} = \hat{\theta}_w - [\hat{J}(\hat{\theta}_w) + \hat{Q}(\hat{\theta}_w)]/\hat{I}(\hat{\theta}_w), \tag{17}$$

where $\hat{\theta}_{wc}$ denotes the CWLE and $\hat{J}$, $\hat{Q}$, and $\hat{I}$ are all evaluated at $\hat{\theta}_w$. Apparently, the CWLE works best when there is a reasonable $\hat{\theta}_w$ to start from. (This will be demonstrated in Section 3.)

3 Simulation Studies

3.1 Design

Some simulation studies were conducted to allow for a comparison of the ERFs and CWLE. For all studies, two test lengths (N=35 and N=70) are considered, and all items are modeled by 3PL models. The real item parameters are generated marginally in MATLAB (The MathWorks, Inc., 2007): Log(a) follows a normal distribution with mean -0.1 and standard deviation 0.4, b follows a normal distribution with mean 0 and standard deviation 1.5, and c follows a beta distribution with mean 0.22 and standard deviation 0.05. The true β's are shown in Table 1.

Our goal is to examine the performance of ERFs and CWLE with different levels of measurement error in β estimation from the calibration sample. The use of different software or calibration sample sizes would yield different levels of measurement error, and it is hard to enumerate all possible combinations of these conditions. For example, large bias in β estimation may result from bad software, a calibration sample size that is too small, or both. To have better control on the magnitude of these measurement errors, δ_a's, δ_b's, and δ_c's are generated in MATLAB. Three levels of measurement error in β estimation are considered: large bias, median bias, and negligible bias, which can be compared to the operational scenarios where β's are estimated from a calibration sample of size < 500, 1,000, and $> 5,000$, respectively. (We assume that the larger the calibration sample size, the better the item parameter estimation, regardless of the software used.) Without loss of generality, the triplet (δ_a,δ_b,δ_c) for each item is assumed to follow a multivariate normal distribution, the mean and variances/covariances of which vary from item to item. Besides, it is natural to assume that the triplets for different items are independent. Each time we generate a set of $\{(\delta_{an},\delta_{bn},\delta_{cn}) : n = 1,\ldots,N\}$ for a total of N items, which are treated as the item bias estimates, and the *estimated* item parameters $(\hat{a}_n, \hat{b}_n, \hat{c}_n)$ for the nth item are calculated by

$$\hat{a}_n = a_n + \delta_{an}, \quad \hat{b}_n = b_n + \delta_{bn}, \quad \text{and } \hat{c}_n = c_n + \delta_{cn}.$$

Table 1

Simulated True Item Parameters for the 70-Item Test

Item	a	b	c	Item	a	b	c
1	1.2145	0.0723	0.2835	36	0.8332	0.6681	0.3200
2	0.9932	2.4498	0.2012	37	0.9514	2.6131	0.3561
3	0.5657	-0.6576	0.3064	38	0.6048	1.1258	0.2698
4	1.2740	1.7239	0.2582	39	0.7591	-1.0534	0.2331
5	0.7467	1.3533	0.2636	40	0.7553	-0.0035	0.2699
6	2.2014	-0.0363	0.2350	41	0.6303	-2.1020	0.2121
7	0.6670	1.0233	0.1862	42	0.8788	1.8163	0.3153
8	0.8617	0.7997	0.2024	43	0.6037	-1.1131	0.2983
9	0.9157	-0.0487	0.2832	44	1.0568	-0.3722	0.2236
10	0.9063	1.2036	0.3496	45	0.5610	-1.4502	0.2884
11	0.4602	-1.7750	0.2747	46	1.5003	-1.9449	0.2368
12	0.8763	1.6128	0.4302	47	0.6732	-2.0891	0.2933
13	0.6917	3.3107	0.2828	48	0.6824	0.0053	0.3418
14	1.2359	0.3697	0.2755	49	0.9529	2.6770	0.2407
15	0.9117	-0.9586	0.3024	50	0.8636	-3.4394	0.3951
16	1.5037	-0.8225	0.3110	51	1.0132	-2.3428	0.2354
17	0.8800	-0.4547	0.3925	52	0.7933	0.2393	0.2555
18	0.6777	-1.9274	0.3303	53	0.6666	-0.0767	0.2359
19	1.7620	0.6543	0.2772	54	0.9970	3.0302	0.1813
20	1.0723	0.4940	0.2354	55	1.2376	-3.5479	0.3135
21	0.9151	0.5579	0.2104	56	0.7182	-1.2628	0.2447
22	1.8462	0.2401	0.2663	57	0.9063	-2.4380	0.2705
23	1.6068	0.0391	0.2892	58	0.5784	-2.7701	0.2560
24	0.9381	0.4437	0.3172	59	1.5682	-0.7267	0.2733
25	0.6205	-0.0567	0.3358	60	1.0389	-1.0317	0.2629
26	1.5043	0.2674	0.2842	61	1.0407	-0.9062	0.2596
27	1.6106	0.2059	0.3404	62	0.6053	-1.2803	0.2780
28	1.1303	0.6481	0.2471	63	0.9283	-0.3579	0.3197
29	1.3088	0.0140	0.1752	64	1.9090	-1.5824	0.3210
30	0.7699	0.7510	0.2383	65	0.8335	-0.7725	0.2208
31	0.9393	-2.2205	0.1912	66	0.7925	2.4777	0.2431
32	1.3867	-0.1077	0.2963	67	0.5538	1.1808	0.2504
33	0.8239	1.8788	0.2107	68	1.3536	-0.5254	0.2669
34	1.0747	-1.3066	0.2006	69	1.1407	0.8589	0.2630
35	1.9959	0.9734	0.2024	70	1.2149	-0.2157	0.2677

Note. The true item parameters for the 35-item test are the first 35 items.

The above procedure is repeated 100 times, and the sample variances/covariances are used to approximate the true variance/covariance matrix of the $\hat{\beta}$'s. In other words, there is no restriction on the correlation between any two estimated item parameters of the same item. The correlation can be positive or negative, depending completely on the resulting estimated item parameters. Table 2 presents the mean simulated bias of the $\hat{\beta}$'s based on 100 replications. The whole

procedure, which is the calibration step, is replicated for each test length.

In the target sample, 13 ability levels are set up with 100 examinees at each level for each replication. The estimated item parameters obtained at the previous stage are fixed to obtain the $\hat{\theta}_w$. The CWLEs are then calculated from Equation 17 using the simulated biases and the estimated variance/covariance matrices. As for the ERFs, the posterior distribution has to

Table 2

Mean Simulated Bias of Estimated Item Parameters Based on 100 Replications

	Large bias			Median bias			Negligible bias		
Item	a	b	c	a	b	c	a	b	c
1	-0.0673	-0.1418	-0.0042	0.0963	-0.1088	0.0233	-0.0088	0.0021	-0.0010
2	-0.0725	-0.1037	0.0232	-0.0147	-0.0974	-0.0296	0.0022	0.0092	-0.0047
3	0.0520	-0.0514	-0.0654	0.0655	0.0147	-0.0010	-0.0113	-0.0144	-0.0002
4	0.0020	-0.0123	-0.0147	0.0593	-0.0047	0.0404	0.0014	-0.0160	0.0020
5	0.0687	-0.0504	-0.0403	-0.1160	-0.0620	0.0235	0.0083	-0.0096	-0.0031
6	-0.0521	-0.0860	-0.0286	-0.0552	0.0431	-0.0333	-0.0033	-0.0107	0.0002
7	-0.0172	-0.1482	-0.0015	-0.1807	-0.0656	-0.0238	0.0081	0.0018	-0.0021
8	-0.0178	-0.0028	0.0033	-0.0541	-0.1072	-0.0578	0.0027	-0.0012	0.0084
9	-0.0743	-0.0716	0.0037	0.1408	-0.0628	-0.0069	0.0132	-0.0017	-0.0007
10	-0.0509	-0.0718	-0.0491	-0.0585	-0.1190	-0.0099	-0.0137	-0.0030	-0.0057
11	-0.0827	0.0238	-0.0088	0.0145	0.0219	0.0184	-0.0020	-0.0006	-0.0065
12	-0.1175	-0.1352	-0.0160	-0.0613	0.0381	0.0085	0.0104	-0.0040	0.0010
13	0.0185	-0.0133	0.0127	0.1680	-0.0603	-0.0003	0.0046	-0.0050	-0.0031
14	-0.0936	-0.1151	-0.0103	0.0231	-0.0888	-0.0037	0.0011	-0.0071	0.0006
15	-0.0236	-0.2164	-0.0191	-0.1471	-0.0322	0.0289	-0.0093	-0.0207	-0.0033
16	-0.0368	-0.0676	-0.0000	-0.0146	-0.0708	0.0024	-0.0040	-0.0046	0.0067
17	0.0098	-0.1156	-0.0495	0.0416	-0.1104	-0.0523	-0.0038	0.0023	-0.0066
18	-0.0244	-0.1088	-0.0248	0.0869	-0.0128	-0.0244	0.0080	0.0038	0.0071
19	-0.0000	-0.1935	-0.0343	0.1507	-0.0266	0.0333	-0.0089	-0.0110	0.0028
20	-0.0343	-0.2045	-0.0079	0.0570	-0.1056	-0.0053	0.0002	-0.0073	-0.0077
21	0.0039	-0.0713	0.0184	-0.0854	-0.0125	-0.0026	-0.0021	0.0013	0.0029
22	0.0417	-0.0951	-0.0464	0.0237	-0.0987	-0.0049	-0.0001	0.0012	-0.0011
23	-0.1374	-0.1076	0.0087	-0.0530	-0.0785	0.0077	0.0061	0.0007	0.0049
24	-0.0566	-0.1159	-0.0492	-0.0525	0.0533	-0.0105	0.0007	-0.0011	0.0015
25	-0.0532	-0.0820	-0.0257	-0.0842	-0.1187	0.0292	-0.0054	-0.0102	-0.0034
26	-0.1202	-0.0441	-0.0444	0.0794	-0.0643	0.0072	0.0006	-0.0175	-0.0001
27	0.0071	-0.1400	-0.0193	-0.1645	0.0070	-0.0058	0.0056	-0.0002	0.0024
28	0.0358	-0.0753	-0.0265	-0.0791	-0.0265	-0.0312	0.0041	-0.0080	-0.0075
29	-0.0828	-0.1192	0.0437	0.0750	-0.0090	-0.0054	0.0013	-0.0110	-0.0022
30	-0.0071	-0.0240	0.0028	0.1542	-0.1247	-0.0480	-0.0019	0.0055	-0.0023
31	0.0383	-0.0165	0.0117	-0.0681	-0.0863	0.0214	-0.0101	-0.0129	-0.0065
32	-0.0517	-0.0264	-0.0211	0.0149	0.0250	-0.0204	0.0037	0.0058	0.0025
33	-0.0508	-0.1032	0.0232	0.1537	-0.1540	0.0165	-0.0018	0.0041	0.0002
34	-0.0232	-0.1583	-0.0337	0.1686	-0.0819	-0.0107	0.0023	-0.0016	0.0035
35	-0.0221	-0.1443	0.0036	0.0151	-0.1214	-0.0112	0.0012	-0.0054	-0.0037

be determined. Approximating the transformed item parameters $\beta^* = (\log(a), b, \text{logit}(c))$ by a multivariate normal distribution, the delta method is applied to get the posterior distribution of β. This process was repeated 100 times to match 100 replications in the calibration step for each test length. As a result, there are 10,000 examinees in total at each of 13 ability levels.

In our studies, the ability estimates of examinees who answered all the items right or wrong were set to be 4 or -4, respectively. In addition, we defined $\hat{\theta}_{wc} = \hat{\theta}_w$ when $\hat{\theta}_w$=4 or -4.

3.2 Evaluation

The accuracy and precision are studied by examining the bias and root mean square error (RMSE) of WLE, CWLE, and ERFs at each ability level. Let $\tilde{\theta}_j$ and θ_j denote the estimated and true abilities for the jth examinee at the Lth ability level, $j = 1, \ldots, 10,000$ and $L = 1, \ldots, 13$. Then,

$$\text{Bias}_L = \frac{1}{10000} \sum_{j=1}^{10000} \{\tilde{\theta}_j - \theta_j\}, \tag{18}$$

and

$$\text{RMSE}_L = \sqrt{\frac{1}{10000} \sum_{j=1}^{10000} \{\tilde{\theta}_j - \theta_j\}^2}. \tag{19}$$

The $\tilde{\theta}_j$ is an estimate resulting from the method of WLE, CWLE, or ERFs.

3.3 Results

Figure 1 presents the results of the bias and RMSE of ability estimates for two different test lengths with large bias in $\hat{\beta}$. The top panel shows that, for the 35-item test, the CWLE is better in reducing bias than WLE for θ values higher than -1.5. The bias of the CWLE is fairly close to zero for all θ values higher than -1. On the other hand, ERFs do not appear to work well in reducing bias except for the median θ values. Neither ERFs nor CWLE could reduce the RMSE. The graphs in the bottom panel make obvious that longer tests (i.e., $N = 70$) lead to better ability estimation, in terms of bias and RMSE, when the WLE or CWLE is applied. The maximum bias of the WLE is less than 0.1, while the bias of the CWLE stays around zero for almost all θ values. This means the CWLE performs best when the bias of WLE is within a reasonable range. The bias for the ERFs is not significantly influenced by the change of test lengths because its range is $[-0.2, 0.2]$ in both situations, which agrees with the findings in Lewis

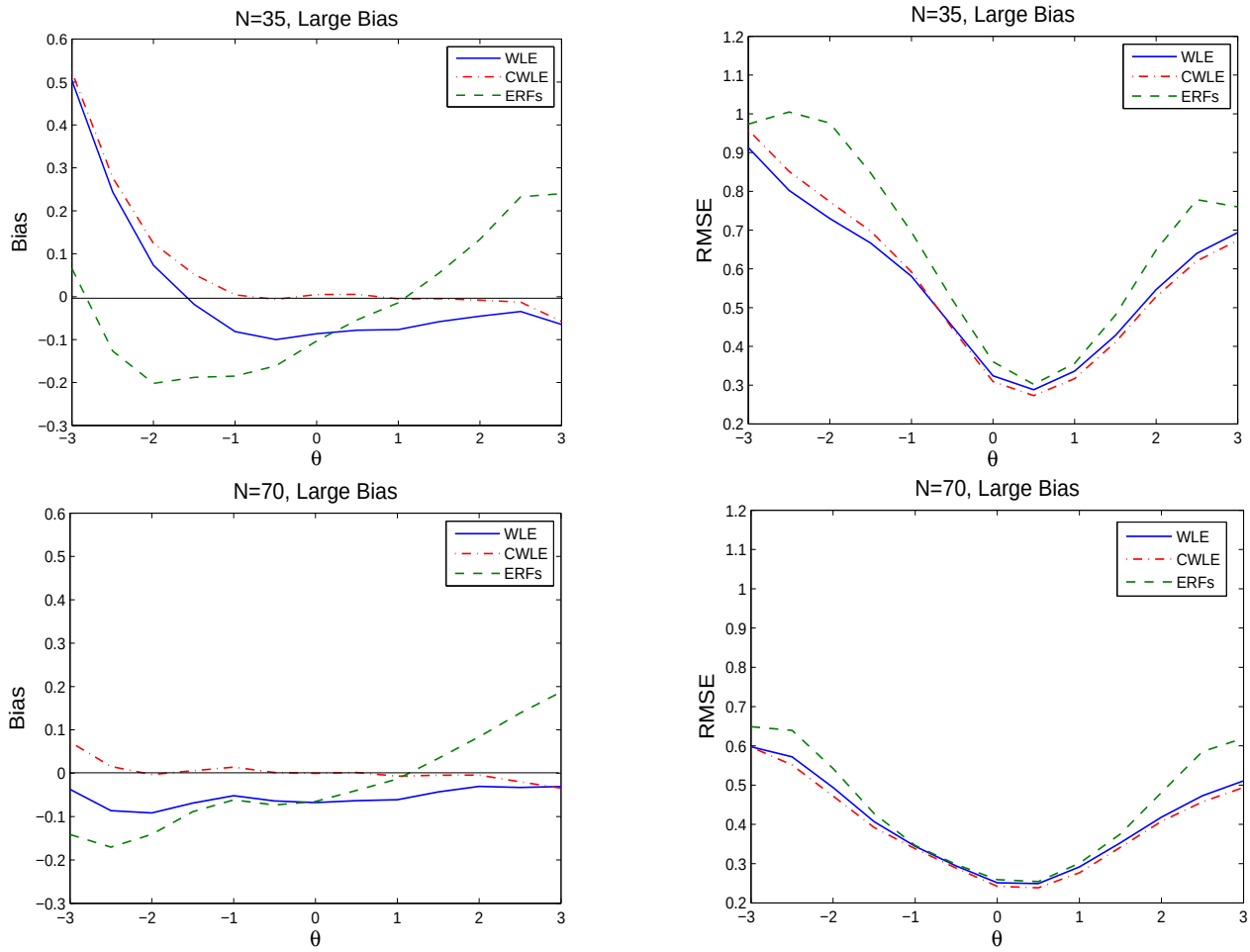

Figure 1. **Bias and root mean square error of ability estimates for two test lengths; large bias in item parameter estimation.**

(2001) that "the greater test lengths, at least in the θ range studied, do not appear to reduce the relative effects of increasing uncertainty about the item parameters." However, the RMSE of the ERFs does decrease for longer tests.

Figure 2 illustrates the relationship between the magnitude of measurement error and the effectiveness of the two bias-correction methods. The left panel indicates that the bias of WLE approaches zero for θ values higher than -2, when the amount of measurement error decreases towards zero. This observation verifies that the WLE performs well as long as the measurement error in $\hat{\beta}$ is negligible. CWLE still performs well in reducing bias with median or greater amounts of measurement error, but the difference between CWLE and WLE decreases as the measurement

error decreases. The two curves overlap when the bias of $\hat{\beta}$ becomes negligible. As expected, the curves of ERFs are almost identical in all cases; the bias is within $[-0.2, 0.2]$ for all θ values. The possible noises are averaged out beforehand, even if the true β can be regarded as known. The right panel shows that the RMSE of the CWLE tends to the RMSE of the WLE when the amount of measurement error approaches zero. The RMSE of the ERFs stays the same for all cases.

It is worth noting that, in both figures, when the θ value is lower than -2, CWLE does not appear to work well for all 35-item tests in the sense that it does not reduce the bias of WLE. One possible explanation is that under our experimental conditions $\hat{\theta}_w$ is not accurate enough for extreme low θ values. The CWLE is a bias-reduction method based on the WLE and is applicable when the WLE can provide a good estimate. Adopting the bias-correction procedure in Equation 17 without paying attention to the results of the WLE could lead to a worse θ estimate. On the other hand, the ERFs method does not rely on any other estimation procedure.

4 Conclusions

Several conclusions can be drawn concerning the use of the two bias-correction procedures for θ estimation in IRT models. First, both methods are easy to implement, and their underlying theories are intuitive (i.e., the CWLE corrects bias by removing it from the biased estimate, while the ERFs corrects it by averaging out the noise). Second, the CWLE can effectively reduce the bias in θ estimation even when the bias in $\hat{\beta}$ is large, provided that reasonable results of WLE are found to initiate the bias-correction procedure given by Equation 17. Finally, the ERFs can cause the bias in θ estimation to fall within $[-0.2, 0.2]$, regardless of the test length and the magnitude of bias in item parameter estimation.

Although the last two conclusions are based on simulation studies, there is no reason to expect different results for tests with operational characteristics similar to the experimental conditions. As discussed in Section 3, measurement errors in item parameter estimation are generated in our studies, and there is no restriction on the correlation between any two estimated item parameters of the same item. Further empirical studies are needed to see how different degrees of association between estimated item parameters will affect the efficiency of these two methods in reducing the bias in ability estimation.

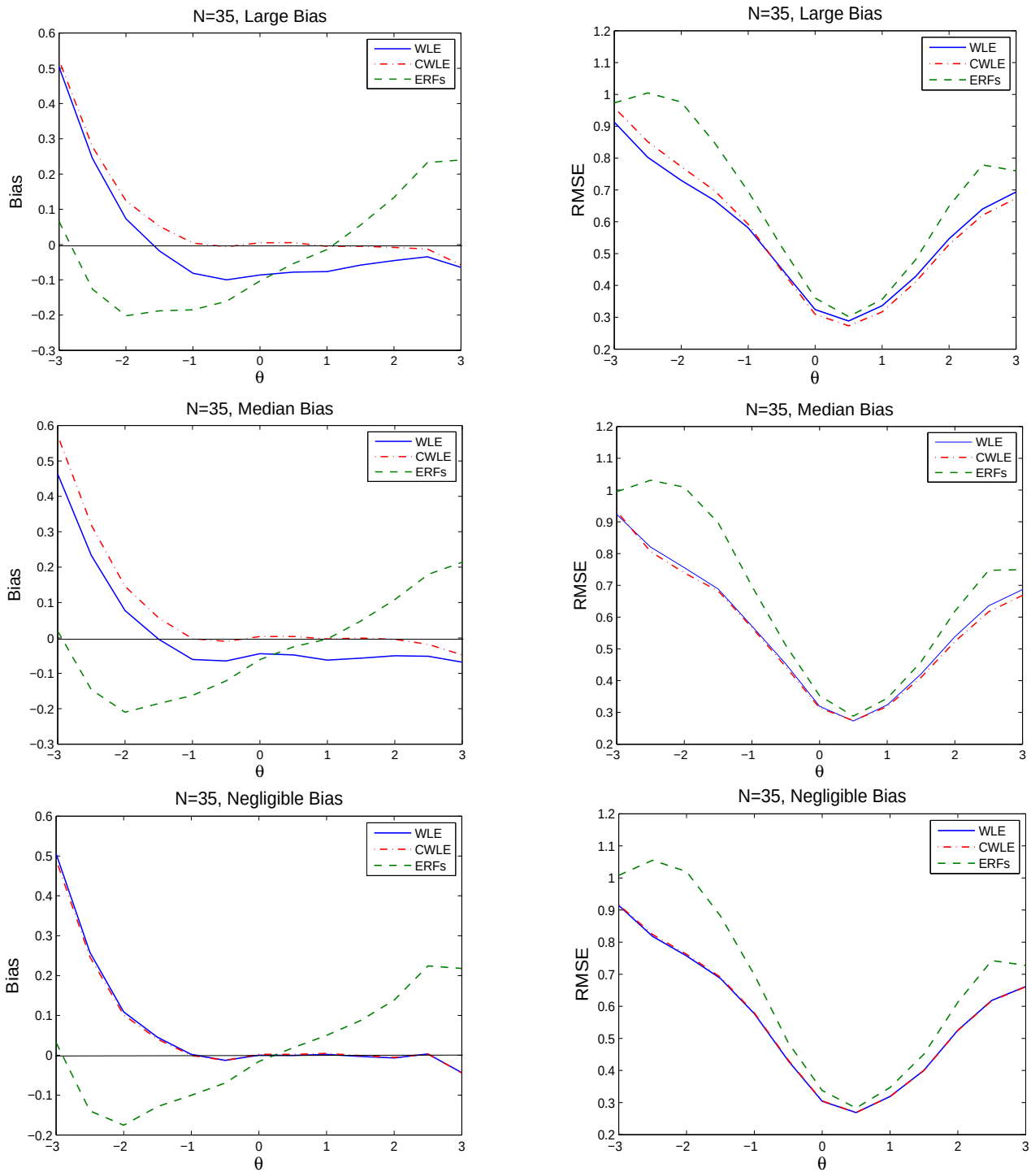

Figure 2. **Bias and root mean square error of ability estimates for three levels of measurement error ($N = 35$).**

References

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

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

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

The MathWorks, Inc. (2007). *Getting started with* MATLAB*, Version 7.5.* Natick, MA: Author.

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

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

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

Zhang, J. & Lu, T. (2007). *Refinements of a bias-correction procedure for the weighted likelihood estimator of ability* (ETS Research Rep. No. RR-07-23). Princeton, NJ: ETS.

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