Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 41

A Comparative Study of Item Response Theory Item Calibration Methods for the Two Parameter Logistic Model

Kyung Yong Kim[†] *and Won–Chan Lee*[††]

April 2015

[†]Kyung Yong Kim, 210B Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: kyungyong-kim@uiowa.edu)

[††]Won-Chan Lee is Associate Professor and Co–director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

In this article, five item calibration methods were compared in terms of the recovery of item parameters for the two–parameter logistic model. All five methods employed the marginal maximum likelihood estimation method via EM algorithm and consisted of three factors including specification of the ability distribution (Normal and Empirical distributions), numerical integration (Midpoint and Gauss–Hermite quadrature methods), and frame of reference for item parameters (Prior and Posterior ability distributions). Specifically, the five methods were Normal–Midpoint–Prior, Normal–Hermite–Prior, Normal–Midpoint–Posterior, Normal–Hermite–Posterior, and Empirical–Midpoint–Prior. A simulation study was conducted to evaluate the five methods under 12 simulation conditions, which were the combinations of two underlying ability distributions (standard normal and skew normal distributions), two test lengths (30 and 49), and three sample sizes (500, 1,000, and 3,000). In addition, four different numbers of quadrature points (11, 21, 31, and 41) were used to calibrate item parameters under each simulation condition. The accuracy of the item parameter estimates for the five methods were compared based on the bias and root mean squared error statistics. The most important factor that affected the accuracy of the item parameter estimates was the specification of the ability distribution. Under the standard normal ability distribution, the Empirical–Midpoint–Prior method did not perform as accurately as the other methods in recovering the item parameters when the sample size was small, but the accuracy of the item parameter estimates improved as the sample size increased. When the underlying distribution was the skew normal distribution, the Empirical–Midpoint–Prior method produced item parameter estimates with less bias but larger standard errors than the other four methods.

1 Introduction

Most item response theory (IRT) computer packages used in practice employ the marginal maximum likelihood estimation (MMLE) procedure implemented via the EM algorithm (Bock & Aitkin, 1981) and the marginalized Bayesian estimation procedure (Mislevy, 1986) to calibrate item parameters. The marginalized Bayesian estimation procedure often is considered as an extension of MMLE–EM because of their similarity in score functions (i.e., the first derivatives of the log–likelihood with respect to the item parameters). However, within these two frameworks, different IRT programs use different item calibration methods, which results in differences in item parameter estimates. Differences in item parameter estimates might have an impact on subsequent stages such as estimating latent traits or equating test forms.

A couple of studies exist that investigated the recovery of item parameters under different factors. Seong (1990a) examined the sensitivity of item and ability parameters to the specification of the ability distribution. In this study, three underlying ability distributions (normal, positively–, and negatively–skewed distributions) were used to generate response data under two sample–size conditions (100 and 1,000), and the accuracy of item and ability parameter estimates were compared for different specifications of the ability distribution. Using a simulation study (five replications per condition), it turned out that item parameters were more accurately estimated when the ability distribution was correctly specified and the sample size was large. Abilities were also more precisely estimated under the correct specification of the ability distribution even when the sample size was small. Seong (1990b) compared the accuracy of item and ability parameter estimates for two numerical integration methods: the midpoint and Gauss–Hermite quadrature methods. Seong (1990b) found that the midpoint method estimated item and ability estimates more accurately when a small number of quadrature points were used for estimation; whereas no significant differences were observed between the two methods when a large number of quadrature points were chosen. However, the above studies focused on only one factor (specification of the ability distribution or numerical integration method) that affected the item parameter estimates without considering the interaction among multiple factors. Moreover, the results were based on a single data set or a small number of simulated data sets due to the limitation of the processing power for computers at the time the studies were conducted.

The main purpose of this study is to extend the scope of the previous studies and compare the recovery of item parameters for five item calibration methods. The five methods are combinations of the following three factors:

1. Specification of the ability distribution
   (a) fixed at the standard normal distribution (N)
   (b) estimated concurrently with item parameters (E)
2. Numerical integration

   (a) midpoint method (M)
   (b) Gauss–Hermite quadrature method (H)

3. Frame of reference for item parameters
   (a) prior ability distribution (Pr)
   (b) posterior ability distribution (Po)

Among all possible combinations of the three factors, NMPr, NHPr, NMPo, NHPo, and EMPr are considered in the present study. The reason for comparing these five methods among eight possible combinations is that they are the methods that are implemented by at least one of the following four commonly used IRT computer packages: BILOG–MG (Zimowski, Muraki, Mislevy, & Bock, 2003), PARSCALE (Muraki & Bock, 2003), flexMIRT (Cai, 2013), and ICL (Hanson, 2002). The specific objectives are:

1. to provide an in–depth description of the three factors considered in this study;
2. to compare the four IRT computer programs in terms of their default settings and options; and
3. to compare the recovery of item parameters for the two–parameter logistic model (2PL) under five item calibration methods.

2 Description of the Three Factors

This section provides an in–depth description of the three factors that are considered in this study. As mentioned previously, the three factors are specification of the ability distribution, numerical integration, and frame of reference for item parameters.

2.1 Specification of the Ability Distribution

For the MMLE–EM procedure, the marginal likelihood of the item parameters $L(\Delta)$ is computed as

$$\begin{aligned} L(\Delta) &= P(U \mid \Delta) = \prod_{i=1}^{N} P(u_i \mid \Delta) \\ &= \prod_{i=1}^{N} \int_{-\infty}^{\infty} P(u_i \mid \Delta, \theta_i) h(\theta_i)\, d\theta_i \\ &= \prod_{i=1}^{N} \int_{-\infty}^{\infty} \prod_{j=1}^{n} P(u_{ij} \mid \delta_j, \theta_i) h(\theta_i)\, d\theta_i, \end{aligned} \tag{1}$$

and under the marginalized Bayesian estimation procedure, the posterior distribution of the item parameters $\pi(\Delta|U)$ is obtained by

$$\begin{aligned}\pi(\Delta|U) &\propto P(U \mid \Delta)\pi(\Delta) = \prod_{i=1}^{N} P(u_i \mid \Delta)\pi(\Delta) \\ &= \prod_{i=1}^{N} \int_{-\infty}^{\infty} P(u_i \mid \Delta, \theta_i) h(\theta_i)\pi(\Delta)\, d\theta_i \\ &= \prod_{i=1}^{N} \int_{-\infty}^{\infty} \prod_{j=1}^{n} P(u_{ij} \mid \delta_j, \theta_i) h(\theta_i)\pi(\delta_j)\, d\theta_i, \end{aligned} \tag{2}$$

where N is the number of examinees; n is the number of items; u_i is a vector of length n containing the response data for examinee i; $U = (u_1, \cdots, u_N)^T$ is a $N \times n$ response data matrix; δ_j is a vector of size ν (the number of item parameters in the model) containing the item parameters for item j; $\Delta = (\delta_1, \delta_2, \ldots, \delta_n)^T$ is a $n \times \nu$ item parameter matrix; $\pi(\Delta)$ is the prior distribution of the item parameters; and $h(\theta_i)$ is the ability distribution for examinee i. For the 2PL model, if examinee i (whose ability is θ_i) correctly responds to item j, then $P(u_{ij} \mid \delta_j, \theta_i)$ is

$$P(u_{ij} = 1 \mid \delta_j, \theta_i) = \frac{e^{1.7a_j(\theta_i - b_j)}}{1 + e^{1.7a_j(\theta_i - b_j)}}, \tag{3}$$

and

$$P(u_{ij} = 0 \mid \delta_j, \theta_i) = 1 - P(u_{ij} = 1 \mid \delta_j, \theta_i) \tag{4}$$

otherwise, where a_j and b_j are the item discrimination and difficulty parameters of item j, respectively. As can be seen from Equations 1 and 2, the ability distribution $h(\theta_i)$ needs to be specified to evaluate the marginal likelihood $L(\Delta)$ or the posterior distribution $f(\Delta \mid U)$. In many cases, $h(\theta_i)$ is fixed at the standard normal distribution for every examinee i, which implies that examinees' abilities are a random sample from the standard normal distribution. Thus, the index i in $g(\theta_i)$ can be dropped. Alternatively, the ability distribution $h(\theta)$ can be estimated concurrently with the item parameters at the M–step of each EM cycle by estimating the quadrature weight at each quadrature point. For iteration s of the EM algorithm, the new weights $w_q^{(s+1)}\,(q = 1, \ldots, Q)$ are estimated by the following equation:

$$w_q^{(s+1)} = \frac{1}{N} \sum_{i=1}^{N} \frac{P(u_i \mid \Delta^{(s)}, x_q) w_q^{(s)}}{\sum_{q'=1}^{Q} P(u_i \mid \Delta^{(s)}, x_{q'}) w_{q'}^{(s)}}, \tag{5}$$

where $\Delta^{(s)}$ and the Q values of $w_q^{(s)}$ are the item parameter estimates and quadrature weights obtained from the previous iteration, respectively. Equation 5 is the same as the one presented in Bock and Aitkin (1981).

2.2 Numerical Integration

Most IRT calibration programs employ the midpoint method, which is often called the Mislevy's histogram solution in the IRT literature, for evaluating the marginal likelihood or joint posterior distribution of the item parameters. The midpoint method approximates a definite integral by a collection of rectangles whose heights are the values of the integrand at the quadrature points that are equally spaced. If Q quadrature points $(x_1, \ldots, x_Q)$ are used to discretize the probability distribution of a continuous ability variable θ, the mathematical form of the midpoint method is

$$\int_a^b P(u \mid \Delta, \theta) h(\theta)\, d\theta \approx \sum_{q=1}^{Q} P(u \mid \Delta, x_q) h(x_q)\, \Delta x, \tag{6}$$

where a and b can be any real numbers (including $-\infty$ and ∞) and Δx is the length of every subinterval (note that no indices are used in Δx since all the subintervals have an equal length).

Another numerical integration method that is not widely used but provided by some IRT calibration programs is the Gauss–Hermite quadrature method. This method is used to integrate a polynomial function f that has the form $f(x) = g(x)e^{-x^2}$. There are three major differences between the Gauss–Hermite quadrature and midpoint methods. First, instead of using equally spaced quadrature points, the Gauss–Hermite quadrature method uses unequally spaced points, which are the roots of one of the polynomials belonging to the family of Hermite polynomials. For example, in order to use Q quadrature points for this method, the roots of the Qth degree Hermite polynomial need be found. Second, the definite integral is always evaluated between $-\infty$ and ∞. In order words, unlike the midpoint method, the lower and upper limits of the definite integral cannot be specified. Finally, when Q quadrature points are used, the Gauss–Hermite quadrature method yields an exact result for any definite integral of a polynomial function of degree $2Q - 1$ or less (if a function can be approximated by a polynomial, then the method produces an approximate result).

The quadrature points and weights of the Gauss–Hermite quadrature method can be obtained using the `gauss.quad` function in R (R Core Team, 2014), which is included in the `statmod` package (Smyth, Hu, Dunn, Phipson, & Chen, 2014). However, when the standard normal distribution is used as the ability distribution, the quadrature points and weights obtained from the `gauss.quad` function need to be modified by multipling $\sqrt{2}$ to the quadrature points $(x_1, \ldots, x_Q)$ and dividing $\sqrt{\pi}$ from the weights $(w_1, \ldots, w_Q)$. This is because the Gauss–Hermite quadrature method cannot be directly applied to the function $P(u \mid \Delta, \theta)h(\theta)$, where $h(\theta)$ is the standard normal density (note that the kernel of the standard normal density is $e^{-\theta^2/2}$ not $e^{-\theta^2}$, which is required to apply

the Gauss–Hermite quadrature method). That is,

$$\begin{aligned}
\int_{-\infty}^{\infty} P(u \mid \Delta, \theta) h(\theta)\, d\theta &= \int_{-\infty}^{\infty} P(u \mid \Delta, \theta) \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2}\, d\theta \\
&= \int_{-\infty}^{\infty} P(u \mid \Delta, \theta) \frac{1}{\sqrt{2\pi}} e^{-(\theta/\sqrt{2})^2}\, d\theta \\
&= \int_{-\infty}^{\infty} P(u \mid \Delta, \sqrt{2}t) \frac{1}{\sqrt{2\pi}} e^{-t^2} \sqrt{2}\, dt \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} P(u \mid \Delta, \sqrt{2}t) e^{-t^2}\, dt \\
&\approx \frac{1}{\sqrt{\pi}} \sum_{q=1}^{Q} P(u \mid \Delta, \sqrt{2}x_q) w_q \\
&= \sum_{q=1}^{Q} P(u \mid \Delta, \sqrt{2}x_q) \frac{w_q}{\sqrt{\pi}}, \qquad (7)
\end{aligned}$$

where $t = \theta/\sqrt{2}$.

2.3 Frame of Reference for Item Parameters

Within the IRT framework, the origin and unit of measurement of the ability scale are not fixed, which is referred to as scale indeterminacy. This issue occurs because two linearly related ability scales produce the same probability of correctly responding to an item as long as the item parameters are also linearly related the same way. To be more specific, the probability of correctly responding to an item given ability θ and item parameters a and b is equivalent to that given ability $\theta^* = A\theta + B$ and item parameters $a^* = a/A$ and $b^* = Ab + B$, where A and B are constants. Equation 8 shows this relationship mathematically for the 2PL model:

$$\begin{aligned}
P(u = 1 \mid a^*, b^*, \theta^*) &= \frac{e^{1.7a^*(\theta - b^*)}}{1 + e^{1.7a^*(\theta - b^*)}} \\
&= \frac{e^{1.7(a/A)[(A\theta + B) - (Ab + B)]}}{1 + e^{1.7(a/A)[(A\theta + B) - (Ab + B)]}} \\
&= \frac{e^{1.7a(\theta - b)}}{1 + e^{1.7a(\theta - b)}} \\
&= P(u = 1 \mid a, b, \theta). \qquad (8)
\end{aligned}$$

In general, the issue of scale indeterminacy is solved by fixing the mean and standard deviation of the ability distribution at some specific values (e.g., a mean of 0 and a standard deviation of 1). However, because some IRT calibration programs produce more than one ability distribution, finding the frame of reference for the item parameters is not always straightforward. For instance, both BILOG–MG and PARSCALE produce two different ability distributions

by default, which are called the prior and posterior distributions and the item parameter estimates can be expressed on either scale (this issue will be discussed in more detail in the following section). It is important to identify the ability distribution that is used as the frame of reference for the item parameters because the correct distribution needs to be used with the item parameter estimates in subsequent stages such as estimating abilities or conducting IRT equating.

3 IRT Calibration Program Comparison

The midpoint method is the default numerical integration method for all four computer packages mentioned previously, but with different numbers of quadrature points and ranges. For a single group calibration, BILOG–MG, PARSCALE, and ICL use 10, 30, and 40 quadrature points, respectively, over the range of -4 and 4, and flexMIRT uses 49 quadrature points over the range of -6 and 6. However, each program has an option to change these default values. In contrast, the Gauss–Hermite quadrature method is only implemented in PARSCALE as an option. In PARSCALE, the $DIST = n$ keyword in the *CALIB* command determines the numerical integration method and if $n = 3$, the quadrature points and weights are obtained from the Gauss–Hermite quadrature method (the default value is $n = 2$, which is the midpoint method). BILOG–MG does not provide this method internally, but can be implemented by entering the Gauss–Hermite quadrature points and weights manually. In order to do so, the value n of the $IDIST = n$ keyword in the *CALIB* command needs to be set to 2 and the quadrature points and weights have to be specified in the *QUAD* command.

For all four programs, the standard normal distribution is used by default as the initial (prior) ability distribution, which is unchanged during the EM algorithm, for calibrating item parameters. However, the frame of reference for the item parameter estimates are not the same for the four programs. PARSCALE, flexMIRT, and ICL use the prior ability distribution to deal with the issue of scale indeterminacy. Therefore, the frame of reference for the item parameter estimates is the standard normal distribution for these programs. BILOG–MG employs a slightly different approach compared to the other three programs. After the final EM cycle, the prior ability distribution is updated using the item parameter estimates obtained from the last iteration of the EM algorithm and then standardized to have a mean of 0 and a standard deviation of 1 (this updated distribution is called the posterior ability distribution). Then the item parameter estimates obtained from the last EM cycle are rescaled based on the posterior distribution, and, thus, the posterior distribution is the frame of reference for the item parameters. In order to deal with the issue of scale indeterminacy using the prior distribution in BILOG–MG, the *NOADJUST* keyword can be used in the *CALIB* command. This keyword prevents the posterior distribution from being standardized, and as a result, the item parameter estimates remain unchanged. Thus, the prior distribution is the frame of reference for the item parameter estimates. Standardizing this posterior distribution and applying the

same transformation to the item parameter estimates results in the same posterior distribution and item parameter estimates obtained with BILOG–MG's default setting. As mentioned previously, PARSCALE also provides the posterior distribution. However, based on empirical evidences, it seems that the item parameter estimates obtained from the last EM cycle are not rescaled using this distribution. Finally, when the ability distribution is estimated along with the item parameters at each iteration of the EM algorithm, the issue of scale indeterminacy is solved by standardizing the estimated distribution obtained at each M–step (this distribution is often called the empirical ability distribution). In this case, the final item parameter estimates are on the scale of the empirical distribution that is used as the prior distribution at the last EM cycle to estimate the item parameters.

Table 1 summarizes the default item calibration methods among the five methods considered in this study for the four IRT programs and the options each program provides. The NMPo method is the default method for BILOG–MG and the NMPr method is the default method for PARSCALE, flexMIRT, and ICL. In addition, the EMPr method can be implemented with all four programs by using the keywords that are inside the parentheses in the row labelled Ability Distribution.

Table 1: IRT program comparison based on the three factors considered in the present study

	BILOG–MG	PARSCALE	flexMIRT	ICL
Default Setting	NMPo	NMPr	NMPr	NMPr
Integration Method				
Midpoint (M)	Default	Default	Default	Default
Hermite (H)		Option (`DIST = 3`)		
Ability Distribution				
N(0, 1) (N)	Default	Default	Default	Default
Empirical (E)	Option (`EMPIRICAL`)	Option (`FREE`)	Option (`EmpHist=YES`)	Option (`estim_dist`)
Frame of Reference				
Prior (Pr)	Option (`NOADJUST`)	Default	Default	Default
Posterior (Po)	Default			

4 Simulation Method

A simulation study was conducted to compare the five calibration methods in terms of the recovery of item parameters. To generate the data sets used in this study, large–scale high school computation and social studies assessments that consisted of 30 and 49 multiple–choice items, respectively, were first calibrated using BILOG–MG with the default settings except that no prior distributions were employed for the item discrimination parameters (BILOG–MG uses a log–normal prior by default) and that 41 quadrature points were used instead of 10.

These item parameter estimates were assumed to be the true item parameters. The means and standard deviations of the item discrimination and difficulty parameters for both tests are provided in Table 2. Then ability values of size 500,

Table 2: First two moments of the item parameter estimates

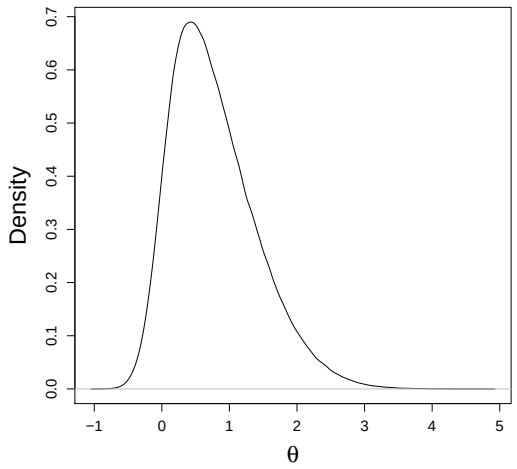
	30–item test		49–item test	
Moment	a	b	a	b
Mean	0.7535	0.0607	0.6410	0.2594
SD	0.2205	0.8331	0.2026	0.6342

1,000, and 3,000 were sampled from each of two underlying ability distributions: the standard normal distribution and a skew normal distribution with a location parameter $\xi = 0$, a scale parameter $\omega = 1$, and a shape parameter $\alpha = 4$. The probability density function of this skew normal distribution is

$$f(x) = \frac{1}{\pi} e^{x^2/2} \int_{-\infty}^{4x} e^{-t^2/2}\, dt, \tag{9}$$

which has a mean of 0.77, a standard deviation of 0.63, and a skewness of 0.78. For the skew normal distribution, the sampled ability values were standardized so that the origin and unit of measurement of the ability scale were 0 and 1, respectively (hereafter, this distribution is referred to as SN(0, 1, 4)). Figure 1 depicts the probability density function of SN(0, 1, 4). Using both the values

Figure 1: Probability density function of SN(0, 1, 4)



of the item and ability parameters, the probability $P(u_{ij} = 1 \,|\, a_j, b_j, \theta_i)$ was

computed for each simulee i and item j. Then, a random number was sampled from a uniform distribution between 0 and 1. If $P(u_{ij} = 1 \,|\, a_j, b_j, \theta_i)$ was larger than the random number, the response to the item was coded 1 and 0 otherwise. This process was repeated 100 times under all 12 conditions (combinations of two test lengths, three sample sizes, and two underlying ability distributions).

All the data sets were calibrated with the five item calibration methods using a series of quadrature points (11, 21, 31, and 41) over the range of -4 and 4 for the midpoint method and over the range of $-\infty$ and ∞ for the Gauss–Hermite quadrature method using a program written in R for this study. The purpose for writing an R program instead of using one of the IRT computer packages mentioned previously was not only to take full control of the program but also to control all the other factors that were not considered in this study but might affect the item parameter estimates. The maximum number of EM cycles was set to 1,000 with a convergence criterion of 0.001. In addition, no prior distributions were employed for the item parameters. There were no convergence issues during the estimation process.

The recovery of the item parameters for the five item calibration methods was compared using two measures of accuracy: the absolute bias (BIAS) and root mean squared error (RMSE) statistics. BIAS for item j is defined as

$$BIAS_j = \left| \frac{1}{100} \sum_{r=1}^{100} \hat{\delta}_{jr} - \delta_j \right|, \tag{10}$$

where $\hat{\delta}_j$ and δ_j denote the estimated and true item parameters (either discrimination or difficulty), respectively. A small value of BIAS indicates that the mean value of an estimator is not systematically different from the true value. However, BIAS does not provide any indication of variability. RMSE takes into account both bias and standard error of an estimator and is computed as follows (for item j):

$$RMSE_j = \sqrt{\frac{1}{100} \sum_{r=1}^{100} (\hat{\delta}_{jr} - \delta_j)^2}. \tag{11}$$

Under each condition, values of BIAS and RMSE of the item discrimination and difficulty parameter estimates were first computed for each item and the average of all the values of BIAS (ABIAS) and the average of all the values of RMSE (ARMSE) were computed for ease of comparison.

5 Results

Tables 3 and 4 display results of the accuracy of item parameters under the N(0, 1) ability distribution. The values of ABIAS and ARMSE for the item discrimination parameters are given below the columns a–ABIAS and a–ARMSE, respectively, and those for the item difficulty parameters are provided respectively below the columns b–ABIAS and b–ARMSE. Similarly, results of the

accuracy of item parameter estimates under the SN(0, 1, 4) ability distribution are summarized in Tables 5 and 6.

Regardless of the test length and the underlying ability distribution, the values of ARMSE decreased as the sample size increased. This result was expected since more stable item parameter estimates would be obtained with larger sample sizes. However, the values of ABIAS showed different patterns for the two underlying ability distributions. Under the N(0, 1) ability distribution, the values of ABIAS decreased as the sample size increased for all five calibration methods, while under the SN(0, 1, 4) ability distribution, only the values of ABIAS for the EMPr method decreased as the sample size increased.

5.1 N(0, 1) Underlying Ability Distribution

For the 30–item test, the NMPr, NHPr, NMPo, and NHPo methods that assumed a fixed standard normal ability distribution (hereafter, these four methods are referred to as methods in Group 1) during the EM algorithm produced similar values of ABIAS and ARMSE when more than 11 quadrature points were used for any sample–size condition. This indicates that the two numerical integration methods, two frames of reference for the item parameters, and number of quadrature points barely affect the recovery of item parameters when the ability distribution is fixed at N(0, 1) during the estimation process and a moderate to large number of quadrature points is used. In addition, these values were smaller than those of the EMPr method for the 500 and 1,000 sample–size conditions, especially for the item discrimination parameters. However, the differences in the values of the four criteria for the two groups of methods decreased as the sample size increased, and they became almost identical at the 3,000 sample–size condition.

There were three noteworthy observations for the 49–item test that were different from the results observed for the 30–item test. First, when the ability distribution was fixed at N(0, 1) during the estimation process, the methods using the posterior distribution (NMPo and NHPo) as the frame of reference for the item parameters produced smaller values of b–ABIAS than those using the prior distribution (NMPr and NHPr). Despite of these differences in the values of b–ABIAS, the values of b–ARMSE were almost identical. The second finding was that the EMPr method estimated item parameters as accurately as the other four methods even for the 1,000 sample–size condition. Also, for the 500 sample–size condition, the values of the four criteria for the methods in Group 1 were still smaller than those of the EMPr method, but the differences were much smaller than the ones observed for the 30–item test. Finally, the values of all four criteria except a–ARMSE for the NHPr method started to increase at the 21 quadrature–point condition, while the values of the same criteria for the other four methods increased at the 11 quadrature–point condition.

Table 3: Values of the ABIAS and ARMSE for the 30–item test when the underlying distribution is N(0, 1)

	a–ABIAS				b–ABIAS				a–ARMSE				b–ARMSE			
	Quadrature Points				Quadrature Points				Quadrature Points				Quadrature Points			
Method	41	31	21	11	41	31	21	11	41	31	21	11	41	31	21	11
N = 500																
NMPr	0.009	0.009	0.009	0.010	0.015	0.015	0.015	0.023	0.098	0.098	0.098	0.096	0.132	0.132	0.132	0.137
NHPr	0.009	0.009	0.008	0.030	0.015	0.015	0.016	0.041	0.098	0.098	0.098	0.098	0.132	0.132	0.133	0.149
NMPo	0.009	0.009	0.009	0.007	0.015	0.015	0.015	0.016	0.098	0.098	0.098	0.097	0.134	0.134	0.134	0.135
NHPo	0.009	0.009	0.008	0.011	0.015	0.015	0.016	0.021	0.098	0.098	0.098	0.096	0.134	0.134	0.134	0.136
EMPr	0.035	0.034	0.024	0.012	0.019	0.020	0.015	0.015	0.115	0.118	0.109	0.101	0.136	0.137	0.134	0.134
N = 1,000																
NMPr	0.007	0.007	0.007	0.011	0.007	0.007	0.007	0.015	0.069	0.069	0.069	0.067	0.091	0.091	0.091	0.093
NHPr	0.007	0.007	0.005	0.032	0.007	0.007	0.008	0.033	0.069	0.069	0.068	0.072	0.091	0.091	0.091	0.103
NMPo	0.006	0.006	0.006	0.005	0.007	0.007	0.007	0.008	0.069	0.069	0.069	0.068	0.092	0.092	0.092	0.092
NHPo	0.006	0.006	0.006	0.010	0.007	0.007	0.007	0.012	0.069	0.069	0.068	0.068	0.092	0.092	0.092	0.093
EMPr	0.014	0.013	0.009	0.006	0.008	0.008	0.007	0.007	0.073	0.073	0.070	0.069	0.092	0.093	0.092	0.092
N = 3,000																
NMPr	0.004	0.004	0.004	0.011	0.004	0.004	0.004	0.012	0.039	0.039	0.039	0.040	0.050	0.050	0.050	0.053
NHPr	0.004	0.004	0.004	0.033	0.004	0.004	0.005	0.030	0.039	0.039	0.039	0.050	0.050	0.050	0.051	0.064
NMPo	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.005	0.039	0.039	0.039	0.039	0.051	0.051	0.051	0.051
NHPo	0.004	0.004	0.004	0.011	0.004	0.004	0.005	0.009	0.039	0.039	0.039	0.040	0.051	0.051	0.051	0.052
EMPr	0.005	0.004	0.004	0.004	0.005	0.004	0.004	0.006	0.041	0.040	0.040	0.039	0.052	0.052	0.052	0.052

Table 4: Values of the ABIAS and ARMSE for the 49–item test when the underlying distribution is N(0, 1)

	a–ABIAS				b–ABIAS				a–ARMSE				b–ARMSE			
	Quadrature Points				Quadrature Points				Quadrature Points				Quadrature Points			
Method	41	31	21	11	41	31	21	11	41	31	21	11	41	31	21	11
N = 500																
NMPr	0.008	0.008	0.008	0.019	0.018	0.018	0.018	0.032	0.085	0.085	0.085	0.084	0.149	0.149	0.149	0.165
NHPr	0.008	0.008	0.006	0.044	0.019	0.019	0.023	0.049	0.085	0.085	0.084	0.090	0.149	0.150	0.155	0.183
NMPo	0.008	0.008	0.008	0.007	0.015	0.015	0.015	0.016	0.085	0.085	0.085	0.084	0.149	0.149	0.149	0.149
NHPo	0.008	0.008	0.007	0.010	0.015	0.015	0.016	0.020	0.085	0.085	0.085	0.083	0.149	0.149	0.149	0.151
EMPr	0.014	0.012	0.010	0.007	0.013	0.014	0.015	0.016	0.090	0.088	0.087	0.085	0.149	0.149	0.149	0.150
N = 1,000																
NMPr	0.004	0.004	0.004	0.025	0.014	0.014	0.014	0.026	0.058	0.058	0.058	0.060	0.100	0.100	0.100	0.113
NHPr	0.004	0.004	0.008	0.051	0.014	0.014	0.016	0.046	0.058	0.058	0.057	0.074	0.100	0.101	0.104	0.127
NMPo	0.004	0.004	0.004	0.006	0.010	0.010	0.010	0.011	0.058	0.058	0.058	0.057	0.100	0.100	0.100	0.101
NHPo	0.004	0.004	0.004	0.014	0.010	0.010	0.010	0.015	0.058	0.058	0.058	0.058	0.100	0.100	0.100	0.102
EMPr	0.006	0.005	0.004	0.004	0.008	0.008	0.009	0.011	0.059	0.059	0.058	0.058	0.100	0.100	0.100	0.101
N = 3,000																
NMPr	0.003	0.003	0.003	0.027	0.010	0.010	0.010	0.023	0.034	0.034	0.034	0.042	0.058	0.058	0.058	0.067
NHPr	0.003	0.003	0.008	0.053	0.010	0.010	0.012	0.044	0.034	0.034	0.034	0.062	0.058	0.058	0.060	0.082
NMPo	0.003	0.003	0.003	0.005	0.005	0.005	0.005	0.006	0.034	0.034	0.034	0.034	0.057	0.057	0.057	0.058
NHPo	0.003	0.003	0.003	0.013	0.005	0.005	0.006	0.011	0.034	0.034	0.034	0.036	0.057	0.057	0.057	0.059
EMPr	0.003	0.003	0.003	0.004	0.006	0.006	0.007	0.007	0.035	0.034	0.034	0.034	0.057	0.057	0.058	0.059

Table 5: Values of the ABIAS and ARMSE for the 30–item test when the underlying distribution is SN(0, 1, 4)

	a–ABIAS				b–ABIAS				a–ARMSE				b–ARMSE			
	Quadrature Points				Quadrature Points				Quadrature Points				Quadrature Points			
Method	41	31	21	11	41	31	21	11	41	31	21	11	41	31	21	11
N = 500																
NMPr	0.030	0.030	0.030	0.035	0.031	0.031	0.031	0.036	0.102	0.102	0.102	0.103	0.134	0.134	0.134	0.139
NHPr	0.030	0.030	0.031	0.047	0.031	0.031	0.032	0.048	0.102	0.102	0.102	0.107	0.134	0.134	0.135	0.148
NMPo	0.030	0.030	0.030	0.032	0.032	0.032	0.032	0.033	0.102	0.102	0.102	0.103	0.135	0.135	0.135	0.136
NHPo	0.030	0.030	0.031	0.036	0.032	0.032	0.032	0.036	0.102	0.102	0.102	0.103	0.135	0.135	0.135	0.137
EMPr	0.042	0.041	0.031	0.013	0.023	0.023	0.020	0.018	0.122	0.123	0.115	0.104	0.133	0.133	0.131	0.130
N = 1,000																
NMPr	0.030	0.030	0.030	0.034	0.021	0.021	0.021	0.025	0.075	0.075	0.075	0.077	0.091	0.091	0.091	0.094
NHPr	0.030	0.030	0.030	0.046	0.021	0.021	0.022	0.039	0.075	0.075	0.075	0.083	0.091	0.091	0.091	0.103
NMPo	0.030	0.030	0.030	0.031	0.021	0.021	0.021	0.022	0.075	0.075	0.075	0.076	0.092	0.092	0.092	0.092
NHPo	0.030	0.030	0.030	0.036	0.021	0.021	0.021	0.024	0.075	0.075	0.075	0.078	0.092	0.092	0.092	0.093
EMPr	0.023	0.023	0.017	0.008	0.015	0.016	0.012	0.010	0.081	0.082	0.078	0.072	0.091	0.092	0.090	0.088
N = 3,000																
NMPr	0.033	0.033	0.033	0.038	0.026	0.026	0.026	0.031	0.054	0.054	0.054	0.057	0.060	0.060	0.061	0.064
NHPr	0.033	0.033	0.034	0.052	0.026	0.026	0.027	0.046	0.054	0.054	0.054	0.068	0.061	0.061	0.061	0.074
NMPo	0.034	0.034	0.034	0.036	0.026	0.026	0.026	0.027	0.054	0.054	0.054	0.055	0.061	0.061	0.061	0.061
NHPo	0.034	0.034	0.034	0.040	0.026	0.026	0.026	0.030	0.054	0.054	0.054	0.059	0.061	0.061	0.061	0.063
EMPr	0.006	0.004	0.003	0.008	0.004	0.005	0.005	0.011	0.042	0.042	0.041	0.041	0.051	0.052	0.052	0.053

Table 6: Values of the ABIAS and ARMSE for the 49–item test when the underlying distribution is SN(0, 1, 4)

	a–ABIAS				b–ABIAS				a–ARMSE				b–ARMSE			
	Quadrature Points				Quadrature Points				Quadrature Points				Quadrature Points			
Method	41	31	21	11	41	31	21	11	41	31	21	11	41	31	21	11
N = 500																
NMPr	0.012	0.012	0.013	0.028	0.015	0.015	0.015	0.031	0.083	0.083	0.083	0.084	0.149	0.149	0.149	0.164
NHPr	0.013	0.013	0.015	0.052	0.015	0.016	0.019	0.050	0.083	0.083	0.083	0.093	0.149	0.149	0.153	0.181
NMPo	0.013	0.013	0.013	0.015	0.019	0.019	0.019	0.021	0.083	0.083	0.083	0.083	0.150	0.150	0.150	0.151
NHPo	0.013	0.013	0.013	0.020	0.019	0.019	0.020	0.024	0.083	0.083	0.083	0.083	0.150	0.150	0.150	0.154
EMPr	0.015	0.014	0.010	0.007	0.013	0.013	0.012	0.013	0.091	0.090	0.088	0.085	0.150	0.150	0.150	0.150
N = 1,000																
NMPr	0.010	0.010	0.010	0.029	0.009	0.009	0.008	0.021	0.059	0.059	0.058	0.063	0.098	0.098	0.098	0.109
NHPr	0.010	0.010	0.014	0.054	0.008	0.008	0.010	0.041	0.058	0.058	0.059	0.077	0.098	0.099	0.102	0.124
NMPo	0.010	0.010	0.010	0.013	0.011	0.011	0.011	0.013	0.058	0.058	0.058	0.059	0.099	0.099	0.099	0.100
NHPo	0.010	0.010	0.011	0.019	0.012	0.012	0.012	0.016	0.058	0.058	0.059	0.060	0.099	0.099	0.100	0.101
EMPr	0.010	0.009	0.006	0.005	0.008	0.008	0.008	0.009	0.062	0.061	0.060	0.059	0.098	0.099	0.099	0.099
N = 3,000																
NMPr	0.011	0.011	0.011	0.033	0.011	0.011	0.011	0.023	0.036	0.036	0.036	0.048	0.059	0.059	0.059	0.069
NHPr	0.011	0.012	0.016	0.059	0.011	0.011	0.012	0.044	0.036	0.037	0.038	0.067	0.059	0.059	0.062	0.084
NMPo	0.012	0.012	0.012	0.015	0.017	0.017	0.017	0.018	0.036	0.036	0.036	0.038	0.061	0.061	0.061	0.062
NHPo	0.012	0.012	0.013	0.022	0.017	0.017	0.017	0.022	0.037	0.037	0.037	0.042	0.061	0.061	0.061	0.064
EMPr	0.003	0.003	0.003	0.005	0.006	0.006	0.007	0.008	0.035	0.035	0.035	0.035	0.058	0.059	0.058	0.059

5.2 SN(0, 1, 4) Underlying Ability Distribution

Similar to the 30–item test under the N(0, 1) ability distribution, the four methods in Group 1 produced similar values of ABIAS and ARMSE for the 30–item test. However, the values of ABIAS for the methods in Group 1 were now larger than those of the EMPr method across most of the conditions, and the largest differences were observed at the 3,000 sample–size condition. In contrast, the EMPr method did not consistently perform more accurately than the other methods in terms of ARMSE. As can be seen from Table 5, the methods in Group 1 estimated item parameters with less overall estimation error than the EMPr method in spite of the larger values of ABIAS for the 500 and 1,000 sample–size conditions. This result implies that, when the true ability distribution departs from N(0, 1), using the empirical ability distribution introduces less bias but more standard errors than the methods fixing the ability distribution at N(0, 1). Although this was also true for the 3,000 sample–size condition, the EMPr method still resulted in the smallest values of ARMSE because of the relatively large differences in the values of ABIAS. Similar results were obtained for the 49–item test, but the differences of the four criteria between the two groups of methods were much smaller than those observed for the 30–item test. In addition, opposite to the results observed for the 49–item test under the N(0, 1) ability distribution, the values of b–ABIAS for the methods fixing the ability distribution at N(0, 1) and using the posterior distribution as the frame of reference for the item parameters (NMPo and NHPo methods) were larger than those for the methods using the prior distribution as the frame of reference (NMPr and NMPr methods).

6 Summary and Discussion

This paper provided an in–depth description of three factors that affect the accuracy of item parameter estimates as well as the default settings and options of four IRT computer packages (BILOG–MG, PARSCALE, flexMIRT, and ICL) associated with those factors. Furthermore, five item calibration methods that were combinations of the three factors were compared under various simulation conditions.

As expected, the values of ARMSE decreased as the sample size increased for all five methods regardless of the underlying ability distribution and test length. However, ABIAS for the five methods showed different patterns under the two underlying ability distributions. Specifically, the values of ABIAS for all five methods decreased as the sample size increased under the N(0, 1) ability distribution, while the method using the empirical distribution showed a decreasing pattern when the underlying ability distribution was SN(0, 1, 4).

In general, for both the 30– and 49–item tests, the numerical integration method, frame of reference for item parameters, and number of quadrature points barely had any impact on the recovery of item parameters when fixing the ability distribution at N(0, 1) during the estimation process and using 21 or more

quadrature points. However, there were two exceptions. First, for the 49–item test, the values of a–ARMSE for the NHPr method started to increase at the 21 quadrature–point condition. The other exception was also observed for the 49–item test where the values of b–ABIAS between the methods using the prior and posterior distributions as the frame of reference for the item parameters showed some differences. To be more specific, the methods using the posterior distribution as the frame of reference for the item parameters produced smaller values of b–ABIAS under the N(0, 1) ability distribution, while the opposite result was observed under the SN(0, 1, 4) ability distribution.

As described earlier, for the two underlying ability distributions, methods using different specifications of the ability distribution produced different results. Under the N(0, 1) ability distribution, the methods fixing the ability distribution at N(0, 1) during the estimation process estimated item parameters more accurately than the method using the empirical ability distribution when the sample size was small. However, the EMPr method performed as accurately as the other methods in recovering the item parameters for the 3,000 sample–size condition and even for the 1,000 sample–size condition for the 49–item test. When the underlying distribution was SN(0, 1, 4), the EMPr method introduced less bias but more standard errors than the other four methods, which resulted in larger values of ARMSE for the 500 and 1,000 sample–size conditions. For the 3,000 sample–size condition, the values of ARMSE for the EMPr method were still the smallest because of the relatively large differences in the values of ABIAS.

The above results suggest that the specification of the ability distribution has the greatest impact on the accuracy of item parameter estimates among the three factors considered in the present study. Therefore, it seems important to specify the ability distribution correctly using theoretical or empirical evidence when calibrating item parameters. When no information about the underlying ability distribution is available, the distribution can be selected based on the sample size of the test. With small sample sizes, using the standard normal distribution during the estimation process produces more accurate item parameter estimates than using the empirical distribution; whereas using the empirical distribution is a better choice with large sample sizes. Note that this suggestion only applies when more than 20 quadrature points are used during the estimation process. As mentioned previously, most IRT computer programs use more than 20 quadrature points when conducting a single group item parameter calibration. However, BILOG–MG uses only 10 quadrature points by default. Thus, when using BILOG–MG for item parameter calibration, users are advised to increase the number of quadrature points to at least 20.

There are some limitations on generalizing the findings of the present study to more general cases because of the following three reasons: (1) this study compared the recovery of item parameters for only the 2PL model under five item calibration methods; (2) only one non–normal ability distribution (with a specific skewness value) was considered; and (3) a single group calibration was used only, as opposed to multiple group calibration. These restrictions will be addressed in a future study by examining the performance of the five item

calibration methods under other dichotomous and polytomous IRT models and considering multiple non–normal ability distributions with different skewness values. In addition, another interesting future study would be to examine the impact of the three factors considered in this study on the recovery of item parameters for multiple group calibration.

References

Baker, F. B. (1987). Methodology review: item parameter estimation under the one–, two–, and three–parameter logistic models. *Applied Psychological Measurement*, *11*, 111–141.

Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. *Applied Psychological Measurement*, *14*, 139–150.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). CRC Press.

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

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179–197.

Cai, L. (2013). *flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software].* Chapel Hill, NC: Vector Psychometric Group.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: The Guilford Press.

Hanson, B. A. (2002). *IRT command language (Version 0.020301).* Monterey, CA.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational and Behavioral Statistics*, *13*, 243–271.

Kim, K. Y., & Lee, W. (2014). *Recovery of item parameters under various IRT item calibration methods.* Paper presented at the International Meeting of the Psychometric Society, Madison, WI.

Lord, F. M. (2008). *Applications of item response theory to practical testing problems.* New York, NY: Routledge.

Mean, A. D., Morris, S. B., & Blitz, D. L. (2007). *Open-source IRT: A comparison of BILOG-MG and ICL features and item parameter recovery.*

Mislevy, R. J. (1986). Bayes model estimation in item response models. *Psychometrika*, *51*, 177–195.

Mislevy, R. J., & Bock, R. D. (1982). *Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program.* Paper presented at the Item Response Theory and Computerized Adaptive Testing Conference, Wayzata, MN.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, *13*, 57–75.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales [Computer software].* Skokie, IL: Scientific Software International, Inc..

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Seong, T. J. (1990a). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, *14*, 299–311.

Seong, T. J. (1990b). *Validity of using two numerical analysis techniques to estimate item and ability Parameters via MMLE: Gauss–Hermite quadrature formula and Mislevy's histogram solution.* Paper presented at the Annual Meeting of the National Council on Measurement in Education.

Smyth, G., Hu, Y., Dunn, P., Phipson, B., & Chen, Y. (2014). statmod: Statistical modeling [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=statmod` (R package version 1.4.20)

Woodruff, D., & Hanson, B. A. (1997). *Estimation for item response models using the EM algorithm for finite mixtures.* Paper presented at the Annual Meeting of the Psychometric Society, Gatlingburg, TN.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, *52*, 275-291.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG–MG 3 for Windows: Multiple–group IRT analysis and test maintenance for binary items [Computer software].* Skokie, IL: Scientific Software International, Inc..