A New Online Calibration Method Based on Lord's Bias-Correction

Applied Psychological Measurement
2017, Vol. 41(6) 456–471
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621617697958
journals.sagepub.com/home/apm

Yinhong He¹, Ping Chen¹, Yong Li¹, and Shumei Zhang¹

Abstract

Online calibration technique has been widely employed to calibrate new items due to its advantages. Method A is the simplest online calibration method and has attracted many attentions from researchers recently. However, a key assumption of Method A is that it treats person-parameter estimates $\hat{\theta}_s$ (obtained by maximum likelihood estimation [MLE]) as their true values θ_s , thus the deviation of the estimated $\hat{\theta}_s$ from their true values might yield inaccurate item calibration when the deviation is nonignorable. To improve the performance of Method A, a new method, MLE-LBCI-Method A, is proposed. This new method combines a modified Lord's bias-correction method (named as *maximum likelihood estimation-Lord's bias-correction with iteration* [MLE-LBCI]) with the original Method A in an effort to correct the deviation of $\hat{\theta}_s$ which may adversely affect the item calibration precision. Two simulation studies were carried out to explore the performance of both MLE-LBCI and MLE-LBCI-Method A under several scenarios. Simulation results showed that MLE-LBCI could make a significant improvement over the ML ability estimates, and MLE-LBCI-Method A did outperform Method A in almost all experimental conditions.

Keywords

CAT, IRT, online calibration, Method A, MLE, error correction

Introduction

Computerized adaptive testing (CAT), which is based on item response theory (IRT), is a new kind of testing mode (e.g., H.-H. Chang, 2015; H.-H. Chang & Zhang, 2002). In CAT, a test taker's ability can be estimated or updated immediately after each response is given by the test taker (Cheng & Chang, 2009; Patton, Cheng, Yuan, & Diao, 2013). Compared with the traditional paper-pencil (P&P) test, CAT has compelling advantages in many aspects including shorter test length, more efficient score reporting, and more flexible testing time (Cheng & Chang, 2009; C. Wang, Chang, & Boughton, 2013; Weiss, 1982). Due to these advantages, CAT becomes popular in large-scale assessment programs such as the Graduate Management Admission Test (GMAT) and the Armed Services Vocational Aptitude Battery (ASVAB)

¹Beijing Normal University, China

Corresponding Author:

Ping Chen, Collaborative Innovation Center of Assessment Toward Basic Education Quality, Beijing Normal University, No.19, Xijiekouwai Street, Haidian District, Beijing 100875, China.
Email: pchen@bnu.edu.cn

(H.-H. Chang, 2012; Cheng, 2008), and also plays an important role in large-scale personalized teaching (H.-H. Chang, 2015).

The maintenance and management of item bank is essential for CAT administration, and item replenishment is even more important than P&P administration due to the accelerated exhausting of items in CAT (Y. C. I. Chang & Lu, 2010). After a period of time, some operational items may no longer be suitable for use due to reasons such as overexposure, obsolescence, or flaw (Wainer & Mislevy, 1990), thus it is very necessary to develop new items to substitute the unsuitable ones. Moreover, the new items have to be precisely calibrated before they are first put into use, because poorly calibrated items may result in bad estimation of latent trait levels of examinees (e.g., Y. C. I. Chang & Lu, 2010; van der Linden & Glas, 2000). Although it is possible to use traditional calibration with anchor design for item calibration, a more cost-effective and commonly used method is online calibration. Online calibration refers to the process of assigning the new items to test takers during the course of their adaptive tests and then estimating the item parameters of new items based on the collected responses (Wainer & Mislevy, 1990). Compared with the traditional anchor design, online calibration technique has several advantages, such as putting the item parameters of the new items on the same scale as the operational items without post hoc linking/scaling (P. Chen, Xin, Wang, & Chang, 2012), saving time and money (Makransky, 2009), and reducing the impact of motivation concerns related to administration of new items to volunteers (Parshall, 1998).

As mentioned by P. Chen and Xin (2014), online calibration design and online calibration method are two important aspects of online calibration. Different from online calibration design, which aims to design the way the new items are assigned to test takers during the course of their adaptive tests for obtaining more accurate calibration results, the main role of online calibration method is to estimate the item parameters of the new items and put them on the same scale as the operational items after response data have been collected. In addition, when calibrating new item online, the item parameters of the operational items are typically fixed and only the new items need to be calibrated. Thus, online calibration methods belong to the fixed parameter calibration (FPC) method according to Kim (2006). Several online calibration methods have been proposed for the unidimensional CAT (UCAT), such as Method A and Method B (Stocking, 1988), “one Expectation Maximization (EM) cycle” method (OEM) (Wainer & Mislevy, 1990), and “multiple EM cycles” method (MEM) (Ban, Hanson, Wang, Yi, & Harris, 2001). Note that in Kim’s (2006) taxonomy system, OEM and MEM are referred to as the no prior weights updating and one EM Cycle (NWU-OEM) method and the no prior weights updating and multiple EM cycles (NWU-MEM) method, respectively.

Among all available calibration methods in UCAT, Method A is the relatively simplest and most straightforward method conceptually and computation wise (P. Chen & Wang, 2016), which can be briefly described as follows: When the response data have been collected, maximum likelihood estimation (MLE) method is used to estimate test takers’ θ s based on their responses to the operational items (the ML ability estimates are denoted as $\hat{\theta}$ s hereafter), and $\hat{\theta}$ s are then treated as the true values and used to estimate the parameters of the new items based on examinees’ responses to the new items. According to Lord (1983), the $\hat{\theta}$ s are characterized by “outward deviation,” meaning that $\hat{\theta}$ s are positively biased for high-ability examinees and negatively biased for low-ability examinees. As a result, the “outward deviation” pattern inherent in $\hat{\theta}$ might yield inaccurate item calibration if Method A is used for calibrating the new items. To overcome the limitation of Method A and improve its performance, $\hat{\theta}$ s should be corrected first before they are put into use.

Generally, correction methods or prevention methods are used to reduce bias in MLE (Firth, 1993), among which the Lord’s bias-correction method (Lord, 1983) is one of the classical corrective approaches (e.g., Firth, 1993; T. Wang, Hanson, & Lau, 1999) and the

weighted likelihood estimation (WLE) method (Firth, 1993; Warm, 1989) is one of the representative preventive approaches. In a preliminary study, Lord's bias-correction method with an iterative process (referred to as MLE-LBCI) was found to perform equally well or even better than WLE for reducing bias in MLE. Thus, the authors propose using the modified Lord's bias-correction method, that is, MLE-LBCI, to correct the deviation of θ s in the original Method A. Then, by combining MLE-LBCI with Method A, a new online calibration method, namely, MLE-LBCI-Method A, is proposed. It should be noted that, during the CAT process, MLE and Bayesian estimation methods, such as expected a posteriori (EAP) and maximum a posteriori (MAP) (Baker & Kim, 2004), are typically used together to update a test taker's ability estimate (Han, 2016), implying that the "outward deviation" also occurs in the provisional ability estimate when MLE is used to update the test taker's ability estimate. Thus, the ML ability estimates can be corrected in two cases: during the adaptive testing process whenever MLE is employed to estimate the test taker's ability, or just at the end of CAT.

This article concentrates on the development of a new robust online calibration method (i.e., MLE-LBCI-Method A) for new items, and conducts two simulation studies to (a) carefully examine whether MLE-LBCI can improve the estimation accuracy compared to the traditional MLE and (b) thoroughly compare the new calibration method with some existing methods under a variety of conditions. The rest of the article is organized as follows: First, the IRT model used in this study is briefly described, followed by the brief introduction of MLE estimation of ability and the modified Lord's bias-correction procedure. Next, the newly proposed online calibration method, MLE-LBCI-Method A, is introduced in detail. Then, the design of the simulation studies is provided, followed by the presentation of the results and conclusions. The final section provides discussion and suggestions for further research.

IRT Model

The three-parameter logistic model (3PLM; Birnbaum, 1968) is one of the mainstream models used for dichotomously scored items, and is widely used in adaptive tests (van der Linden & Ren, 2015). For the j th ($j = 1, 2, \dots, L$) item, the probability of a correct response $y_{ij} = 1$ by a test taker with ability θ_i ($i = 1, 2, \dots, N$) is as follows:

$$P_j(\theta_i) \equiv P(y_{ij} = 1 | \theta_i, a_j, b_j, c_j) = c_j + (1 - c_j) \frac{1}{\{1 + \exp(-Da_j(\theta_i - b_j))\}}, \quad (1)$$

where a_j , b_j , and c_j are the discrimination parameter, difficulty parameter and pseudo-guessing parameter of item j , respectively. The mission of item calibration is to first estimate these item parameters via certain statistical methods and then put them on the existing scale. D is the scale constant, logistic scale is used when $D = 1$, and normal scale is used when $D = 1.702$. For simplicity, D is set to be 1 in this article.

MLE Estimation of Ability and Lord's Bias-Correction Procedure

Suppose a test consists of L dichotomously scored items. Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iL})^T$ denote the response vector of the i th ($i = 1, 2, \dots, N$) test taker. Under the assumption of local independence or conditional independence (Lord, 1980), the likelihood function of observing the response vector $\mathbf{y}_i$ is as follows:

$$L(\mathbf{y}_i|\theta_i) = \prod_{j=1}^L \left[(P_j(\theta_i))^{y_{ij}} (Q_j(\theta_i))^{1-y_{ij}} \right], \quad (2)$$

where $Q_j(\theta_i) = 1 - P_j(\theta_i)$. Assuming that the item parameters $(a_j, b_j, c_j)^T$ are known and fixed, then the ML estimate $\hat{\theta}$ is defined as the value of θ that maximizes Equation 2. Typically, $\hat{\theta}$ is found by setting the derivative of the log-likelihood function with respect to θ to zero; that is, $\hat{\theta}$ satisfies the following estimating equation:

$$\frac{\partial \ln L(\mathbf{y}_i|\theta_i)}{\partial \theta_i} = \sum_{j=1}^L \left[\left(\frac{y_{ij} - P_j(\theta_i)}{P_j(\theta_i)Q_j(\theta_i)} \right) P'_j(\theta_i) \right] = 0, \quad (3)$$

where $P'_j(\theta_i) = a_j P_j^*(\theta_i) Q_j(\theta_i)$, and $P_j^*(\theta_i) = 1 / [1 + \exp(-Da_j(\theta_i - b_j))]$ is the item response function for the two-parameter logistic model (Birnbbaum, 1968). Then the log-likelihood Equation 3 becomes

$$\sum_{j=1}^L \left[a_j \left(y_{ij} - P_j(\theta_i) \frac{P_j^*(\theta_i)}{P_j(\theta_i)} \right) \right] = 0. \quad (4)$$

Lord (1983) investigated the statistical properties of the ML estimate $\hat{\theta}_i$ in IRT, and provided the bias function for the ML estimate $\hat{\theta}_i$ when the item parameters are known:

$$B_L(\theta_i) = \frac{1}{I^2(\theta_i)} \sum_{j=1}^L \left[a_j I_j(\theta_i) \left(P_j^*(\theta_i) - \frac{1}{2} \right) \right], \quad (5)$$

where $I(\theta_i)$ is the Fisher test information function evaluated at θ_i , and $I_j(\theta_i)$ is the item information function of item j ($j = 1, 2, \dots, L$).

When all the items are modeled by 3PLM, it is easy to derive

$$I(\theta_i) = \sum_{j=1}^L \left[\frac{(P'_j(\theta_i))^2}{P_j(\theta_i)Q_j(\theta_i)} \right] = \sum_{j=1}^L \left[a_j^2 (1 - c_j) Q_j^*(\theta_i) \frac{(P_j^*(\theta_i))^2}{P_j(\theta_i)} \right] \quad (6)$$

and

$$I_j(\theta_i) = a_j^2 (1 - c_j) Q_j^*(\theta_i) \frac{(P_j^*(\theta_i))^2}{P_j(\theta_i)}, \quad (7)$$

where $Q_j^*(\theta_i) = 1 - P_j^*(\theta_i)$.

Replace true θ_i with its estimate value $\hat{\theta}_i$ when calculating $B_L(\theta_i)$ in Equation 5, $\hat{\theta}_i$ with Lord's bias-correction is thus given by the following equation:

$$\hat{\theta}_{ic} = \hat{\theta}_i - B_L(\hat{\theta}_i). \quad (8)$$

It should be noted that, when the item parameters are known, the bias of $\hat{\theta}_{ic}$ is $o(L^{-1})$ (i.e., $\lim_{L \rightarrow \infty} (L \times \text{Bias}(\hat{\theta}_{ic})) = 0$) while $\text{Bias}(\hat{\theta}_i)$ is $O(L^{-1})$ (i.e., $L \times \text{Bias}(\hat{\theta}_i)$ are bounded for all L) (Zhang, Xie, Song, & Lu, 2011).

Because $B_L(\hat{\theta}_i)$, instead of $B_L(\theta_i)$, is used in Equation 8, in this article, an iterative process is conducted to improve the accuracy of Lord's bias-correction. The iterative formula is as follows:

$$\hat{\theta}_{ic}^{(t+1)} = \hat{\theta}_i - B_L\left(\hat{\theta}_{ic}^{(t)}\right), \quad (9)$$

where the superscripts $t + 1$ and t denote the number of iterations, and the initial iterative value $\hat{\theta}_{ic}^{(0)}$ is set to be $\hat{\theta}_i$. The above iterative process is performed repeatedly until some termination rule is satisfied, for example, the distance between $\hat{\theta}_{ic}^{(t+1)}$ and $\hat{\theta}_{ic}^{(t)}$ is small enough. In addition, this bias-correction method is referred to as *maximum likelihood estimation-Lord's bias-correction with iteration* (MLE-LBCI).

The New Online Calibration Method

Suppose that there are m new items to be calibrated, n_j test takers have answered the j th ($j = 1, 2, \dots, m$) new item, and L_i operational items have been answered by the i th ($i = 1, 2, \dots, n_j$) test taker after the CAT administration. In addition, $\mathbf{u}_j = (u_{1j}, u_{2j}, \dots, u_{n_{jj}})^T$ denotes the response vector of the n_j test takers to the j th new item, and $\mathbf{v}_i = (v_{i1}, v_{i2}, \dots, v_{iL_i})^T$ represents the response vector of the i th test taker to the L_i operational items.

Method A

Before introducing the new online calibration method, the original Method A is first briefly reviewed here. According to Stocking (1988), Method A proceeds with two stages: (a) new items are randomly or adaptively assigned to test takers during the adaptive testing sessions; and (b) the new items are calibrated after all response data are collected. In this article, the first stage is referred to as CAT-stage and the second stage is named as calibration-stage. In addition, based on previous studies (e.g., Han, 2016; Zheng, 2014), MLE is combined with EAP to update the test takers' ability estimates in the CAT-stage.

Following the logic of Method A, MLE is used to estimate the test takers' abilities before calibrating new items, and the ML ability estimate $\hat{\theta}_i$ has bias (see Equation 5), and as a result, the deviation of $\hat{\theta}_i$ will be propagated in the ensuing calibration process for new items. To overcome the theoretical limitation of Method A, it is necessary to correct for the deviation of $\hat{\theta}_i$ whenever Method A is used to calibrate new items.

MLE-LBCI-Method A

MLE-LBCI-Method A is an improved method of Method A, and it corrects for the estimated values of ability via MLE-LBCI before calibrating new items. Similar to Method A, MLE-LBCI-Method A can also be divided into CAT-stage and calibration-stage.

CAT-stage of MLE-LBCI-Method A. The general procedures of the CAT-stage for MLE-LBCI-Method A can be summarized through the following four steps:

Step 1: A test taker's provisional ability estimate (denoted as $\hat{\theta}_i$) is initialized to zero.

Step 2: According to the online calibration design used, if the current examinee reaches a seeding location (i.e., the location where the new item is embedded in the operational tests), select a new item and then assign it to the test taker and collect the test taker's response to it; else, an operational item, which is selected from the item bank based on $\hat{\theta}_i$ according to

the item selection strategy, is presented to the test taker and his or her response on it is recorded.

Step 3: $\hat{\theta}_t$ is sequentially updated based on the responses to the operational items. If the current test length is less than some pre-determined value (e.g., 5) or the response patterns are nonmixed (i.e., all incorrect or all correct), EAP is used; otherwise, MLE is used.

Step 4: Steps 2 and 3 are repeated until the termination criterion is satisfied.

Calibration-stage of MLE-LBCI-Method A. The calibration-stage of MLE-LBCI-Method A is described as follows:

Step 1: ML ability estimate $\hat{\theta}_i$ ($i=1, 2, \dots, n_j$) is obtained by solving the following log-likelihood equation:

$$\sum_{k=1}^{L_i} \left[a_k \left(v_{ik} - P_k(\theta_i) \frac{P_k^*(\theta_i)}{P_k(\theta_i)} \right) \right] = 0. \quad (10)$$

Step 2: Correct for $\hat{\theta}_i$ ($i=1, 2, \dots, n_j$) based on Equation 9, and obtain the corrected ability estimate $\hat{\theta}_{ic}$ ($i=1, 2, \dots, n_j$).

Step 3: Treat the corrected estimates $\hat{\theta}_{ic}$ ($i=1, 2, \dots, n_j$) as “true” ability values and the parameter estimate $\hat{\Delta}_j = (\hat{a}_j, \hat{b}_j, \hat{c}_j)^T$ of the j th ($j=1, 2, \dots, m$) new item is obtained by solving log-likelihood Equation 11:

$$\begin{aligned} \frac{\partial}{\partial \Delta_j} \ln L(\mathbf{u}_j | \hat{\theta}_{1c}, \dots, \hat{\theta}_{n_j c}) &= \sum_{i=1}^{n_j} u_{ij} \frac{1}{P_j(\hat{\theta}_{ic})} \frac{\partial P_j(\hat{\theta}_{ic})}{\partial \Delta_j} \\ &+ \sum_{i=1}^{n_j} (1 - u_{ij}) \frac{1}{1 - P_j(\hat{\theta}_{ic})} \frac{\partial [1 - P_j(\hat{\theta}_{ic})]}{\partial \Delta_j} = 0, \end{aligned} \quad (11)$$

where $\Delta_j = (a_j, b_j, c_j)^T$ denotes the item parameter vector of new item j .

In summary, the main difference between MLE-LBCI-Method A and Method A is that in MLE-LBCI-Method A, ability estimates $\hat{\theta}_i$ ($i=1, 2, \dots, n_j$) are first corrected before they are used to calibrate the new items. Notice that in the CAT-stage of MLE-LBCI-Method A, if available, MLE is used to update the provisional ability estimates. Thus, the provisional ability estimates are also characterized by “outward deviation” due to the use of MLE. Then it is of interest to know whether this deviation produced by MLE in the process of adaptive testing (rather than at the end of CAT) is also worth to be corrected via Equation 9.

Simulation Studies

Research Objectives

Two simulation studies were conducted using programs written in Matlab R2013a. The main purpose of Study 1 is twofold: one is to evaluate the performance of MLE-LBCI and MLE in terms of person-parameter recovery, and the other one focuses on whether it is worth to correct for the provisional ability estimates during CAT. In Study 1, a total number of 100 examinees are simulated, and for each examinee, the person parameter is estimated in the context of CAT under three levels of test length (i.e., $L = 10, 20$, and 30). Besides, the entire CAT simulation process, which includes selecting operational items, simulating the examinees’ responses to the

operational items, and updating provisional ability estimates, is repeated 100 times ($\text{rep}_1 = 100$) for reducing random errors.

In Study 2, the new calibration method, MLE-LBCI-Method A, is thoroughly compared with the original Method A and two other existing representative methods (i.e., OEM and MEM) under three levels of sample size ($N = 1,000, 2,000$, and $3,000$) and three levels of test length ($L = 10, 20$, and 30). The three levels of sample size and three levels of test length constitute nine experimental conditions in total. In addition, both the item parameters and the latent weights are estimated/updated concurrently when using MEM, because MEM with *multiple weights updating* (i.e., MWU-MEM) showed better results than NWU-MEM according to Kim (2006). Thus, MEM in this study actually refers to MWU-MEM.

Generation of Items and Examinees

An operational item bank which consists of 1000 items is simulated. For the operational items, the pseudo-guessing parameter c is randomly drawn from a uniform distribution $U(0, 0.2)$, and the item parameter vector $\eta = (a, b)^T$ is randomly drawn from a multivariate normal distribution with a mean vector of $\mu_\eta = (\mu_a, \mu_b)^T$ and a covariance matrix of $\Sigma_\eta = \begin{pmatrix} \text{var}(a) & \text{cov}(a, b) \\ \text{cov}(b, a) & \text{var}(b) \end{pmatrix}$. In addition, the correlation coefficient ρ between a and b is set to be .25 due to the certain degree of positive correlation between a and b (H.-H. Chang, Qian, & Ying, 2001). Besides, because a is generally truncated by an interval ($a \in (L_a, U_a)$) and is supposed to follow a log-normal distribution (Baker & Kim, 2004), it is of interest to know the expectation and variance of a when it is bounded by L_a and U_a . According to Lien (1985), the r th moment of the truncated a is as follows:

$$E(a^r | L_a < a < U_a) = E(a^r) \times \frac{\Phi\left(r\sigma - \frac{\ln(L_a) - \mu}{\sigma}\right) - \Phi\left(r\sigma - \frac{\ln(U_a) - \mu}{\sigma}\right)}{\Phi\left(\frac{\ln(U_a) - \mu}{\sigma}\right) - \Phi\left(\frac{\ln(L_a) - \mu}{\sigma}\right)},$$

where $E(a^r) = \exp(r\mu + (r^2\sigma^2/2))$ and $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution (P. Chen & Wang 2016; Lien, 1985). Thus, $\mu_a \triangleq E(a | 0.2 < a < 2) = 0.8804$ and $\text{var}(a) \triangleq \text{var}(a | 0.2 < a < 2) = 0.2296$ can be obtained by setting $L_a = 0.2$, $U_a = 2$, $\mu = 0$, and $\sigma = 1$. Finally, $\mu_\eta = (0.8804, 0)^T$ and $\Sigma_\eta = \begin{pmatrix} 0.2296 & 0.1198 \\ 0.1198 & 1 \end{pmatrix}$ can be obtained by setting $\mu_b = 0$, and $\text{var}(b) = 1$. In addition, the generated a is truncated between 0.2 and 2, and b is truncated between -3 and 3 in this study.

The examinees' true abilities are randomly drawn from the standard normal distribution $N(0, 1)$ and are truncated between -3 and 3 . A total of 100 examinees are generated in Study 1, whereas three samples of 1,000, 2,000, and 3,000 examinees are simulated in Study 2. Due to space constraints, the descriptive statistics of the simulated true abilities and operational items in the bank are shown in Table A1 in the Online Appendix A.

In Study 2, for each of the nine experimental conditions, a total number of 20 new items ($m = 20$) are generated in the same manner with the operational items. Moreover, following the suggestions in Y. Chen, Liu, and Ying (2015) and P. Chen and Wang (2016), the entire process of generating the new items, simulating the examinees' responses to the new items and then calibrating the new items are replicated 100 times ($\text{rep}_2 = 100$) with a goal to reduce random errors.

Simulation Details

Some simulation details merit illustration. In the CAT-stage, the maximum Fisher information method (Lord, 1980) is chosen as the item selection method, and fixed-length ($L = 10, 20$, and 30) stopping rule is used for simplicity. Besides, random online calibration design is adopted to assign the new items to the test takers during CAT due to its convenient implementation and acceptable calibration precision (e.g., Ban et al., 2001; P. Chen et al., 2012; Wainer & Mislevy, 1990). In addition, to fairly compare all new items, each new item is controlled to be answered by exactly the same number of examinees (i.e., $(N \times w)/m$) as in previous studies (e.g., P. Chen et al., 2012; P. Chen & Wang, 2016; Zheng, 2016), where $N = 1,000, 2,000$ or $3,000$ is the total number of test takers, $w = 5$ is the number of new items each test taker answers, and $m = 20$ is the number of the new items that need to be calibrated. Therefore, the number of responses to each new item is 250, 500 and 750 for sample size $N = 1,000, 2,000$, and $3,000$, respectively.

Evaluation Criteria

The estimation accuracy of test taker's ability is evaluated by root mean square error (RMSE) and bias, and the calibration precision of the new items is measured by RMSE, bias, and the average weighted area difference (AWD) between the true and the estimated item characteristic curves (Zheng, 2014). Smaller RMSE indicates higher estimation accuracy or calibration precision; and if the bias is close to 0, then the estimation or calibration can be regarded as unbiased. Smaller AWD implies better overall recovery of item parameters. And the detailed calculation formulas for the evaluation criteria are given in the Online Appendix A.

Wilcoxon Test

A nonparametric test, Wilcoxon signed-rank test (Rosner, Glynn, & Lee, 2006; Rey & Neuhauser, 2011), is used to explore whether MLE-LBCI can significantly improve the estimation accuracy of MLE in Study 1 and whether MLE-LBCI-Method A is significantly better than the original Method A in terms of item-parameter recovery in Study 2. It should be noted that the Wilcoxon signed-rank test is calculated based on the absolute bias values (see Online Appendix A for details).

Results and Conclusion

Study 1

The modified bias-correction method (i.e., MLE-LBCI) and MLE are thoroughly compared in Study 1. And MLE-LBCI is carried out in two ways: (a) during CAT process when updating test takers' ability estimates via MLE or (b) just at the end of CAT. Denote $\hat{\theta}$ as the ML estimate for an examinee with true ability θ , and the corrected value for $\hat{\theta}$ is denoted as $\hat{\theta}_{Fc}$ if $\hat{\theta}$ is only corrected at the end of CAT; and if examinee's provisional ML ability estimate is sequentially corrected during CAT administration, then the final corrected value of $\hat{\theta}$ is referred to as $\hat{\theta}_{Ic}$.

Estimation accuracy of ability and Wilcoxon signed-rank test. The estimation accuracy of ability and the standardized Wilcoxon signed-rank statistics W^* s for testing $\hat{\theta}_{Fc}$ (or $\hat{\theta}_{Ic}$) versus $\hat{\theta}$, and $\hat{\theta}_{Fc}$ versus $\hat{\theta}_{Ic}$ based on the absolute bias values are reported in Table 1. In addition, the significance level α is chosen to be .05 (i.e., $\alpha = .05$), one-tailed test is used to test $\hat{\theta}_{Fc}$ (or $\hat{\theta}_{Ic}$) versus $\hat{\theta}$, and two-tailed test is used to test $\hat{\theta}_{Fc}$ versus $\hat{\theta}_{Ic}$. Thus, $\hat{\theta}_{Fc}$ (or $\hat{\theta}_{Ic}$) is better than $\hat{\theta}$ if W^* is less than $-Z_{\alpha} = -1.64$, and there is no significant difference between $\hat{\theta}_{Fc}$ and $\hat{\theta}_{Ic}$ if W^* falls in the range of $[-Z_{\alpha/2}, Z_{\alpha/2}] = [-1.96, 1.96]$.

Table 1. Estimation Accuracy of Ability and W^* s Under Different Conditions.

Test length	RMSE			bias			W^*		
	$\hat{\theta}$	$\hat{\theta}_{Fc}$	$\hat{\theta}_{Ic}$	$\hat{\theta}$	$\hat{\theta}_{Fc}$	$\hat{\theta}_{Ic}$	$\hat{\theta}_{Fc}$ vs. $\hat{\theta}$	$\hat{\theta}_{Ic}$ vs. $\hat{\theta}$	$\hat{\theta}_{Fc}$ vs. $\hat{\theta}_{Ic}$
10	0.4582	0.4290	0.4198	0.0016	-0.0025	-0.0027	-22.37	-12.76	-1.16
20	0.3149	0.3091	0.3052	0.0003	-0.0028	-0.0029	-19.47	-8.75	1.43
30	0.2602	0.2574	0.2570	0.0001	-0.0019	-0.0021	-18.32	-9.11	0.20

Note. W^* s are computed based on the absolute bias values. RMSE = root mean square error.

As seen from Table 1, all average bias values are close to zero at all levels of test length. Moreover, the average RMSE values of both $\hat{\theta}_{Fc}$ and $\hat{\theta}_{Ic}$ are smaller than those of $\hat{\theta}$, indicating that the MLE-LBCI method can indeed improve the estimation accuracy of ability compared to the traditional MLE method. In addition, the average RMSE values of $\hat{\theta}_{Fc}$ (or $\hat{\theta}_{Ic}$) are closer to those of $\hat{\theta}$ as the test length increases from 10 to 30, which indicates that the advantage of correcting for the biases is less pronounced for conditions with longer test length. One possible explanation is that $\hat{\theta}$ is gradually closer to its true value as more and more items are answered by this test taker (large sample property of MLE). Thus, it is expected that MLE-LBCI-Method A, which treats the corrected ability estimate $\hat{\theta}_{Fc}$ s or $\hat{\theta}_{Ic}$ s as true abilities, will improve the performance of Method A especially for conditions with shorter test length.

The W^* values in Table 1 confirm that MLE-LBCI is superior to MLE and there is no significant difference between $\hat{\theta}_{Fc}$ and $\hat{\theta}_{Ic}$. More specifically, the W^* values for testing $\hat{\theta}_{Fc}$ (or $\hat{\theta}_{Ic}$) versus $\hat{\theta}$ are less than -1.64 for all levels of test length, indicating that the absolute bias values of $\hat{\theta}_{Fc}$ (or $\hat{\theta}_{Ic}$) are significantly smaller than those of $\hat{\theta}$. Besides, it is not necessary to correct for the provisional ability estimates during the adaptive testing sessions, because all the W^* s for testing $\hat{\theta}_{Fc}$ versus $\hat{\theta}_{Ic}$ fall in the interval $[-1.96, 1.96]$ for all levels of test length. Thus, a correction has to be made for the ML ability estimates just at the end of CAT.

Bias plotted against true abilities. Bias for $\hat{\theta}$, $\hat{\theta}_{Fc}$, and $\hat{\theta}_{Ic}$ against the true abilities for different test lengths are plotted in Figures 1 to 3, respectively. As can be clearly seen from the figures, the average bias values generated by $\hat{\theta}_{Fc}$ and $\hat{\theta}_{Ic}$ are closer to zero than those yielded by $\hat{\theta}$, especially for extreme true ability values. Besides, the biases for $\hat{\theta}_{Fc}$ is very similar to those for $\hat{\theta}_{Ic}$, which is consistent with the results of Wilcoxon signed-rank test.

Study 2

As concluded in Study 1, it is not necessary to correct for the provisional ability estimates during CAT administrations. Thus, only $\hat{\theta}_{Fc}$ values are used in this study when MLE-LBCI-Method A is used to calibrate the new items.

Results regarding the estimation accuracy of ability. The estimation accuracy of $\hat{\theta}$ s and $\hat{\theta}_{Fc}$ s under the nine conditions are provided in Table B1 in the Online Appendix B. As expected, the results presented in Table B1 are consistent with those in Study 1. To be specific, $\hat{\theta}_{Fc}$ s consistently generate smaller average RMSE values than $\hat{\theta}$ s and are less biased than $\hat{\theta}$ s. In addition, Wilcoxon signed-rank test is also conducted based on absolute bias values, and the resulting W^* s values are shown in Table B2 in the Online Appendix B. As seen from Table B2, for all of the nine experimental conditions, the W^* values are less than -1.64 , implying that the corrected ability values, $\hat{\theta}_{Fc}$ s, perform significantly better than $\hat{\theta}$ s.

Results regarding the calibration precision. The calibration results for different online calibration methods under different conditions are summarized in Table 2, where MA and MLMA are short

Figure 1. Bias against true trait for test length 10.
Note. MLE = maximum likelihood estimation.

Figure 2. Bias against true trait for test length 20.
Note. MLE = maximum likelihood estimation.

for Method A and MLE-LBCI-Method A, respectively. And the best results among the four calibration methods are in bold type.

As can be seen from Table 2, the calibration precision for the new items increases as the calibration sample size increases under different test lengths. Specifically, for each calibration method, both the average RMSE values and the average AWD values are strictly decreasing with the increasing of sample size under all test lengths. Besides, the test length does have an impact on calibration precision; for example, when the new items are calibrated by Method A, the average RMSE values of $\hat{\gamma}$ ($\hat{\gamma} = \hat{a}$, $\hat{b}$, or $\hat{c}$) decrease as the test length increases for all levels of sample size (the only exception is the RMSE value for $\hat{b}$ at sample size 2000).

By browsing Table 2, one can notice that the RMSE values on parameter a are smaller than those on parameter b , which seems different from the “experiential knowledge” that b is

Figure 3. Bias against true trait for test length 30.

Note. MLE = maximum likelihood estimation.

typically estimated more accurately than a . One possible explanation for this result is that in this article, parameters a and b are, respectively, truncated between 0.2 and 2 and between -3 and 3 to mimic realistic scenarios. Thus, the narrower range might cause less variation for a ; on the contrary, because the mean squared error (MSE) equals to the sum of variation and the square of bias, less variation of parameter a may lead to smaller RMSE values. In addition, this type of result is also seen in Ban et al. (2001), C. Wang and Xu (2015), and so on. For example, in the study of Ban et al. (2001), except for the MEM method at sample size 1,000, all of the calibration methods (i.e., OEM, MEM, Method A, Method B, and BILOG/Prior) generate smaller RMSE values on parameter a than on parameter b under all conditions. For more rigorous evidence for this explanation, more studies are welcome in the future.

Compared with the original Method A, MLE-LBCI-Method A improves the calibration precision in almost all experimental conditions. To be specific, MLE-LBCI-Method A generates smaller RMSE values, AWD values, and absolute bias values than Method A. For example, for all experimental conditions (except for the condition with test length of 30 and sample size of 2000), the AWD values generated by MLE-LBCI-Method A are strictly smaller than those produced by Method A, meaning that MLE-LBCI-Method A outperforms Method A in terms of AWD as a whole. Moreover, for test lengths of 10 and 20, MLE-LBCI-Method A consistently generated smaller RMSE values than Method A for all sample sizes. However, as the test length increases to 30, it seems that MLE-LBCI-Method A is not well behaved based on RMSE values, for example, MLE-LBCI-Method A generates slightly larger RMSE values on parameters b and c for sample sizes of 2,000 and 3,000. One possible explanation for this is that the ML ability estimates are closer to their true values when the test length becomes larger, which may leave less room for MLE-LBCI to make an improvement for MLE.

By comparing the performance of the four calibration methods, one can find that MEM generates the smallest AWD values in almost all conditions. However, MEM yields the worst performance with the lowest calibration precision on parameter c , because the average RMSE values of c for MEM are the largest ones under all conditions. Also, MEM exhibits unsatisfactory performance on the discrimination parameter a when sample size is small (i.e., 1,000). Contrary to MEM, OEM generates the smallest average RMSE values on parameter c under all conditions (except for the condition with test length of 20 and sample size of 3,000) and shows

Table 2. Results of Different Online Calibration Methods Under Different Conditions.

Test length	Sample size	Method	RMSE			bias			AWD
			<i>a</i>	<i>b</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>c</i>	
10	1,000	MA	0.3765	0.6410	0.0781	−0.1577	0.0210	−0.0014	0.0397
		MLMA	0.3725	0.6376	0.0771	−0.1397	0.0289	−0.0075	0.0384
		OEM	0.3601	0.6260	0.0727	−0.1794	0.0372	−0.0034	0.0388
		MEM	0.3982	0.5763	0.0944	0.0217	−0.0238	−0.0240	0.0361
	2,000	MA	0.3401	0.5505	0.0716	−0.1685	0.0161	0.0028	0.0325
		MLMA	0.3260	0.5234	0.0715	−0.1527	−0.0027	−0.0022	0.0305
		OEM	0.3283	0.5211	0.0678	−0.1946	0.0006	−0.0011	0.0320
		MEM	0.2889	0.4358	0.0847	0.0084	−0.0424	−0.0178	0.0254
	3,000	MA	0.3369	0.5082	0.0700	−0.1650	0.0186	0.0032	0.0290
		MLMA	0.3033	0.4769	0.0696	−0.1485	0.0070	−0.0014	0.0267
		OEM	0.3241	0.5050	0.0677	−0.1881	0.0110	0.0013	0.0289
		MEM	0.2461	0.4289	0.0803	−0.0041	−0.0264	−0.0121	0.0215
20	1,000	MA	0.3614	0.6334	0.0775	−0.0937	0.0398	−0.0069	0.0366
		MLMA	0.3562	0.6012	0.0766	−0.0780	0.0449	−0.0052	0.0356
		OEM	0.3324	0.5859	0.0744	−0.1017	0.0327	−0.0047	0.0351
		MEM	0.3782	0.5678	0.0939	0.0147	−0.0355	−0.0244	0.0351
	2,000	MA	0.2860	0.4879	0.0713	−0.1033	0.0016	−0.0049	0.0271
		MLMA	0.2831	0.4750	0.0705	−0.0905	0.0053	−0.0038	0.0267
		OEM	0.2843	0.4836	0.0695	−0.1091	0.0091	−0.0020	0.0269
		MEM	0.2941	0.4476	0.0841	0.0030	−0.0349	−0.0186	0.0254
	3,000	MA	0.2821	0.4866	0.0683	−0.0975	0.0266	−0.0016	0.0233
		MLMA	0.2773	0.4571	0.0674	−0.0872	0.0118	−0.0026	0.0227
		OEM	0.2746	0.4470	0.0675	−0.1037	0.0051	−0.0018	0.0229
		MEM	0.2502	0.4290	0.0810	−0.0015	−0.0349	−0.0167	0.0210
30	1,000	MA	0.3504	0.5900	0.0769	−0.0665	0.0245	−0.0101	0.0350
		MLMA	0.3481	0.5594	0.0758	−0.0477	0.0295	−0.0075	0.0346
		OEM	0.3464	0.5846	0.0756	−0.0667	0.0309	−0.0064	0.0345
		MEM	0.3813	0.5723	0.0932	0.0111	−0.0322	−0.0257	0.0348
	2,000	MA	0.2822	0.5079	0.0706	−0.0747	0.0213	−0.0045	0.0262
		MLMA	0.2741	0.5221	0.0715	−0.0714	0.0217	−0.0048	0.0262
		OEM	0.2556	0.4900	0.0686	−0.0913	0.0133	−0.0046	0.0258
		MEM	0.2552	0.4613	0.0841	−0.0110	−0.0379	−0.0215	0.0251
	3,000	MA	0.2749	0.4632	0.0676	−0.0730	0.0152	−0.0034	0.0221
		MLMA	0.2619	0.4676	0.0686	−0.0712	0.0094	−0.0048	0.0218
		OEM	0.2500	0.4723	0.0664	−0.0884	0.0252	−0.0014	0.0222
		MEM	0.2460	0.4597	0.0780	−0.0113	−0.0082	−0.0136	0.0213

Note. RMSE = root mean square error; AWD = average weighted area difference; MA = Method A, MLMA = MLE-LBCI-Method A; MLE-LBCI = maximum likelihood estimation-Lord's bias-correction with iteration; OEM = one EM cycle method; MEM = multiple EM cycle method; EM = Expectation Maximization.

better recovery on parameter *a* than MEM for small sample size. When comparing MLE-LBCI-Method A with OEM, an interesting finding is that MLE-LBCI-Method A shows better performance than OEM in terms of AWD index under most conditions. Compared with MEM, although MLE-LBCI-Method A is not as good as this classical calibration method, it does generate smaller average RMSE values for *c* under all conditions and smaller average RMSE values for *a* when the sample size is small. Furthermore, MLE-LBCI-Method A improves the calibration precision of the original Method A under most conditions and is relatively simpler and more straightforward conceptually and computation wise compared with MEM. What's more, MLE-LBCI-Method A is not as time-consuming as MEM and calibration efficiency is a

great concern when the new items have to be calibrated timely required by some online calibration designs such as the optimal Bayesian adaptive design (OBAD; van der Linden & Ren, 2015) (i.e., the new items are sequentially updated each time they receives a fixed number of new responses), thus, in this sense, MLE-LBCI-Method A may be a better choice than MEM. In addition, the average times for MLE-LBCI-Method A and MEM to calibrate a new item are about 0.26 and 2.89 (s), respectively.

Because large negative bias and large positive bias may cancel out, the Wilcoxon signed-rank test based on absolute bias values is also conducted for testing MLE-LBCI-Method A versus Method A, and the resulting W^* values are described in Table B3 in the Online Appendix B. As seen from Table B3, for test lengths of 10 and 20 under all sample sizes and for test length 30 under sample size of 1,000, the W^* s for testing both $\hat{a}$ and $\hat{b}$ are less than -1.64 . In other words, MLE-LBCI-Method A significantly improves the calibration precision for both parameters a and b under these conditions. In addition, the reduction percentage points of MLE-LBCI-Method A relative to Method A based on the average RMSE values and AWD values under different conditions are summarized in Table B4 in the Online Appendix B, where the negative values denote the increased percentage points.

Conclusion

Results of this article can be summarized as follows:

(a) Lord's bias-correction with iteration could indeed increase the estimation accuracy of ability; (b) the new calibration method, MLE-LBCI-Method A, outperforms Method A under most experimental conditions; (c) MEM yields the best overall item parameters recovery, especially when the calibration sample size is large, but it is time-consuming; (d) MLE-LBCI-Method A exemplifies better AWD results than OEM and better RMSE values on parameter a or c than MEM under most conditions; (e) the larger the sample size is, the higher the calibration precision is; (f) both Method A and MLE-LBCI-Method A are sensitive to the test length, and longer test length produces more accurate calibration results.

Discussion and Future Directions

Among all available calibration methods in UCAT, Method A is the simplest and most straightforward one (P. Chen & Wang, 2016) in terms of methodology and implementation. However, Method A has a fatal limitation, that is, it treats the ability estimates $\hat{\theta}_s$ as their "true" values and ignores the estimation errors of θ_s in the calibration process of new items. This article aimed to make a remedy for the limitation of Method A and improve its performance, and thus a new online calibration method, MLE-LBCI-Method A, was proposed. In addition, two simulation studies were conducted for the purposes of examining whether the corrected ability estimates are more accurate than the ML ability estimates and comparing the new calibration method with some existing methods (i.e., Method A, OEM, and MEM) in terms of item-parameter recovery under a variety of conditions. Results from Study 1 showed that the corrected ability values were more accurate than the uncorrected ones and the provisional ML ability estimates were not worth to be corrected during the CAT administrations. Results from Study 2 showed that MLE-LBCI-Method A outperformed Method A for test lengths of 10 and 20 under all sample sizes and was less time-consuming than MEM.

The current article can be expanded in a number of directions. First of all, the correction ideas that originate from the new method can be generalized to multidimensional online calibration methods such as M-Method A (P. Chen & Wang, 2016; P. Chen, Wang, Xin, & Chang, 2017) and combined with other correction methods such as full functional maximum likelihood

estimation-multidimensional-Method A (FFMLE-M-Method A; P. Chen & Wang, 2016) to make a comprehensive correction for the errors caused by the uncertainty of ability estimates. Second, due to the growing emphasis on non-multiple-choice items, online calibration methods based on polytomously scored items have attracted much attention from researchers recently (e.g., Zheng, 2016). Thus, how to extend the new method to polytomous IRT models is a future direction. Third, this article has only focused on the ability estimate deviation brought by MLE method, and do not take other possible factors into consideration. Other factors such as the measurement errors (e.g., Carroll, Ruppert, Stefanski, & Crainiceanu, 2006) and position effects (e.g., Hecht, Weirich, Siegle, & Frey, 2015) should also be further considered in future studies. Fourth, assume that the item parameters of the operational items are known and fixed; however, they are also estimated with calibration errors in practical applications. Previous studies such as Patton et al. (2013) and Zhang et al. (2011) have investigated the impact of the uncertainty about operational item parameters on ability estimate, and showed that the item-parameter errors would be transferred to the ensuing scoring process. It is still an open territory whether the uncertainty regarding the operational item parameters could be corrected when calibrating new items, and this is worthy of investigation in the future. Fifth, only random online design is considered in our simulation studies, more complicated design such as OBAD (van der Linden & Ren, 2015) and sequential design (Berger, 1994; Berger, King, & Wong, 2000; Y. C. I. Chang & Lu, 2010; Jones & Jin, 1994) may be taken into consideration in the future to explore the performance of the new calibration method. Finally, as response time could be incorporated as collateral information (e.g., Kang, 2016; van der Linden, Klein Entink, & Fox, 2010; T. Wang & Hanson, 2005), future studies may focus on how to improve calibration accuracy when response time is available.

Acknowledgment

The author is indebted to the editor, associate editor, and two anonymous reviewers for their suggestions and comments on the earlier manuscript.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was partially supported by the National Natural Science Foundation of China (Grant 31300862), the Specialized Research Fund for the Doctoral Program of Higher Education (Grant 20130003120002), and KLAS (Grant 130028614).

Supplemental Material

The online appendices are available at <http://journals.sagepub.com/doi/suppl/10.1177/0146621617697958>.

References

- Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.
- Ban, J. C., Hanson, B. A., Wang, T. Y., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item—Calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, 38, 191-212.

- Berger, M. P. F. (1994). D-optimal sequential sampling designs for item response theory models. *Journal of Educational Statistics*, 19, 43-56.
- Berger, M. P. F., King, J., & Wong, W. K. (2000). Minimax D-optimal designs for item response theory models. *Psychometrika*, 65, 377-390.
- Birnbaum, A. (1968). Some latent trait 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. 379-479). Reading, MA: Addison-Wesley.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). London, England: Chapman & Hall.
- Chang, H.-H. (2012). Making computerized adaptive testing diagnostic tools for schools. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessments: Recent history and predictions for the future* (pp. 195-226). Information Age.
- Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80, 1-20.
- Chang, H.-H., Qian, J. H., & Ying, Z. L. (2001). a-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, 25, 333-341.
- Chang, H.-H., & Zhang, J. M. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika*, 67, 387-398.
- Chang, Y. C. I., & Lu, H. Y. (2010). Online calibration via variable length computerized adaptive testing. *Psychometrika*, 75, 140-157.
- Chen, P., & Wang, C. (2016). A new online calibration method for multidimensional computerized adaptive testing. *Psychometrika*, 81, 674-701.
- Chen, P., Wang, C., Xin, T., & Chang, H.-H. (2017). Developing new online calibration methods for multidimensional computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 70, 81-117.
- Chen, P., & Xin, T. (2014). Online calibration with cognitive diagnostic assessment. In Y. Cheng & H.-H. Chang (Eds.), *Advancing methodologies to support both summative and formative assessments* (pp. 287-313). Charlotte, NC: Information Age.
- Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, 77, 201-222.
- Chen, Y., Liu, J. C., & Ying, Z. L. (2015). Online item calibration for Q-matrix in CD-CAT. *Applied Psychological Measurement*, 39, 5-15.
- Cheng, Y. (2008). *Computerized adaptive testing—New developments and applications* (Unpublished doctoral thesis). University of Illinois at Urbana-Champaign.
- Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical & Statistical Psychology*, 62, 369-383.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27-38.
- Han, K. T. (2016). Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests. *Applied Psychological Measurement*, 40, 289-301.
- Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Modeling booklet effects for nonequivalent group designs in large-scale assessment. *Educational and Psychological Measurement*, 75, 568-584.
- Jones, D. H., & Jin, Z. Y. (1994). Optimal sequential designs for on-line item estimation. *Psychometrika*, 59, 59-75.
- Kang, H.-A. (2016). *Likelihood estimation for jointly analyzing item responses and response times* (Unpublished doctoral thesis). University of Illinois at Urbana-Champaign.
- Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, 43, 355-381.
- Lien, D.-H. D. (1985). Moments of truncated bivariate log-normal distributions. *Economics Letters*, 19, 243-247.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their Parallel-Forms reliability. *Psychometrika*, 48, 233-245.

- Makransky, G. (2009, June). *An automatic online calibration design in adaptive testing*. Paper presented at the GMAC conference on Computerized Adaptive Testing, McLean, VA.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Parshall, C. G. (1998, September). *Item development and pretesting in a computer-based testing environment*. Paper presented at the colloquium Computer-Based Testing: Building the Foundation for Future Assessments, Philadelphia, PA.
- Patton, J. M., Cheng, Y., Yuan, K. H., & Diao, Q. (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement*, 37, 24-40.
- Rey, D., & Neuhäuser, M. (2011). *Wilcoxon Signed-Rank Test*. Berlin, Germany: Springer Berlin Heidelberg.
- Rosner, B., Glynn, R. J., & Lee, M. L. T. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*, 62, 185-192.
- Stocking, M. L. (1988). *Scale drift in on-line calibration* (Research Report No. 88-28). Princeton, NJ: Educational Testing Service.
- van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, 13, 35-53.
- van der Linden, W. J., Klein Entink, R. H., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, 3, 327-347.
- van der Linden, W. J., & Ren, H. (2015). Optimal Bayesian adaptive design for test-item calibration. *Psychometrika*, 80, 263-288.
- Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 65-102). Hillsdale, NJ: Lawrence Erlbaum.
- Wang, C., Chang, H.-H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, 37, 99-122.
- Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, 68, 456-477.
- Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, 29, 323-339.
- Wang, T., Hanson, B. A., & Lau, C. M. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement*, 23, 263-278.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-450.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473-492.
- Zhang, J. M., Xie, M. G., Song, X. L., & Lu, T. (2011). Investigating the impact of uncertainty about item parameters on ability estimation. *Psychometrika*, 76, 97-118.
- Zheng, Y. (2014). *New methods of online calibration for item bank replenishment* (Unpublished doctoral thesis). University of Illinois at Urbana-Champaign.
- Zheng, Y. (2016). Online calibration of polytomous items under the generalized partial credit model. *Applied Psychological Measurement*, 40, 434-450.