JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education

Journal of Educational Measurement
Summer 2019, Vol. 56, No. 2, pp. 302–330

Standard Errors of IRT Parameter Scale Transformation Coefficients: Comparison of Bootstrap Method, Delta Method, and Multiple Imputation Method

Zhonghua Zhang
The University of Melbourne
Mingren Zhao
Shenzhen University

The present study evaluated the multiple imputation method, a procedure that is similar to the one suggested by Li and Lissitz (2004), and compared the performance of this method with that of the bootstrap method and the delta method in obtaining the standard errors for the estimates of the parameter scale transformation coefficients in item response theory (IRT) equating in the context of the common-item nonequivalent groups design. Two different estimation procedures for the variance-covariance matrix of the IRT item parameter estimates, which were used in both the delta method and the multiple imputation method, were considered: empirical cross-product (XPD) *and* supplemented expectation maximization (SEM). *The results of the analyses with simulated and real data indicate that the multiple imputation method generally produced very similar results to the bootstrap method and the delta method in most of the conditions. The differences between the estimated standard errors obtained by the methods using the XPD matrices and the SEM matrices were very small when the sample size was reasonably large. When the sample size was small, the methods using the XPD matrices appeared to yield slight upward bias for the standard errors of the IRT parameter scale transformation coefficients.*

The standard error of equating measures the amount of equating error when the examinee scores or item parameters on two or more different test forms are to be linked for comparison (Kolen & Brennan, 2004). Standard 5.13 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) states that "Standard errors of equating functions should be estimated and reported whenever possible" (p. 105). There are at least two types of widely applied procedures for estimating the standard errors of equating: *the bootstrap method* and *the delta method*. The bootstrap method is a resampling procedure which can become very computationally intensive. The delta method is an analytic method which relies on the derivation of complicated equations and therefore can be time-consuming and intractable. The purpose of this study is to evaluate a multiple imputation method, which is less computationally demanding and does not depend on the development of complicated equations, and to compare the performance of this method with that of the bootstrap and delta methods in estimating the standard errors for the estimates of the item response theory (IRT) (Hambleton, Swaminathan, & Rogers, 1991) parameter scale transformation coefficients in IRT test equating under the common-item nonequivalent groups (CINEG) design (Kolen & Brennan, 2004).

Test Equating

Test equating refers to linking scores on alternate test forms that are developed to measure a common construct and are built to the same statistical specifications (Kolen & Brennan, 2004). It is often employed in standardized testing programs to produce comparable scores across different administrations of tests. IRT equating becomes necessary when the item and person parameters that are developed using IRT in different tests need to be placed onto a common measurement scale. IRT equating typically contains two steps: (1) placing the estimates of the IRT item/person parameters onto a common measurement scale, (2) equating observed or true scores. In practice, the IRT item and person parameters that are estimated from two tests which share a number of common items can be placed onto a common scale through linear parameter scale transformation functions (Hanson & Béguin, 2002; S. H. Kim & Cohen, 1998; Kolen & Brennan, 2004). The standard errors for the estimates of the coefficients of the linear transformation functions are of interest to the current study.

In the study, we focused on the CINEG equating design (Kolen & Brennan, 2004). This design assumes that two tests, which contain a set of internal common items, are administered to two independent groups of examinees, which are nonequivalent with respect to the ability distributions. To be specific, we followed a similar general design to that of Ogasawara (2001a) and Wong (2015) by assuming that two test forms, Test 1 and Test 2, which shared a set of common items, were administered to two independent nonequivalent groups of examinees, Group 1 and Group 2. IRT calibrations were conducted separately for each of the two tests. The IRT parameter estimates were assumed to be equated from the scale of Test 2 to that of Test 1.

IRT Parameter Scale Transformation

The two-parameter logistic IRT (2PL IRT; Birnbaum, 1968) model was adopted in the study. For the aforementioned CINEG design, the probability of the *i*th examinee in Group 1 correctly answering the *j*th item in Test 1 can be represented by the following 2PL IRT model:

$$P_1(X_{1ij} = 1|\theta_{1i}; a_{1j}, b_{1j}) = \frac{e^{a_{1j}(\theta_{1i}-b_{1j})}}{1+e^{a_{1j}(\theta_{1i}-b_{1j})}}\,, \tag{1}$$

where $X_{1ij} = 1$ denotes a correct response of examinee *i* in Group 1 to item *j* in Test 1, θ_{1i} is the level of the latent trait for examinee *i* in Group 1, a_{1j} and b_{1j} are the discrimination and difficulty parameters for item *j* in Test 1, respectively. Analogously, the corresponding probability that the *i*th examinee in Group 2 correctly answer the *j*th item in Test 2 can be defined in the following equation:

$$P_2(X_{2ij} = 1|\theta_{2i}; a_{2j}, b_{2j}) = \frac{e^{a_{2j}(\theta_{2i}-b_{2j})}}{1+e^{a_{2j}(\theta_{2i}-b_{2j})}}\,. \tag{2}$$

To resolve the indeterminacy of the location and scale for the parameter estimates in 2PL IRT calibration, it is common in practice to constrain the person ability parameters (θ) to have a mean of zero and variance of one. For the CINEG design, if the IRT parameters are obtained by conducting two separate 2PL IRT calibrations for Test 1 and Test 2, the estimates of the item discrimination and difficulty parameters for the common items between the two tests are generally different because they are built on different measurement scales. The item parameter estimates as well as the person ability estimates obtained from the two separate IRT calibrations on Test 1 and Test 2 are comparable only if the two sets of parameter estimates are placed onto a common scale. This can be fulfilled via conducting linear IRT parameter scale transformations using the transformation coefficients *A* and *B* (Kolen & Brennan, 2004).

To be specific, the estimates of the item parameters in Test 2 can be transformed to the scale of Test 1 through the following two linear transformation functions:

$$a^*_{2j} = \frac{a_{2j}}{A}, \tag{3}$$

$$b^*_{2j} = Ab_{2j} + B, \tag{4}$$

where a^*_{2j} and b^*_{2j} are the transformed values of the item discrimination and difficulty parameters for item j on the scale of Test 1, respectively; and a_{2j} and b_{2j} are the estimates of item discrimination and difficulty parameters for this item on the scale of Test 2, respectively. *A* and *B* are the transformation coefficients.

Methods for Estimating the IRT Parameter Scale Transformation Coefficients *A* and *B*

There are at least two methods that can be used to estimate the transformation coefficients *A* and *B* in IRT equating: *moment methods* and *characteristic curve methods* (Kolen & Brennan, 2004). Both methods involve using the IRT parameter estimates of the common items to obtain the two linear transformation coefficients. The moment methods utilize the moments (mainly the mean and the standard deviation) of the item parameter estimates of the common items to calculate *A* and *B*. One of the moment methods is the mean/sigma approach (Marco, 1977). This approach uses the means and the standard deviations of the item difficulty parameter estimates of the common items. Another moment method, which is called the mean/mean approach (Loyd & Hoover, 1980), uses the means of the item discrimination parameter estimates and the means of the item difficulty parameter estimates for the common items to calculate the two linking coefficients. The characteristic curve methods make use of the item or test characteristic curves which are based on the IRT probability functions for the common items to find *A* and *B*. The method that is based on the item characteristic curves for the common items is the Haebara approach (Haebara, 1980). The Haebara approach solves for A and B by finding the optimized values that minimize a loss function which describes the summation of the squared differences between the item characteristic curves for the common items in the tests that need to be equated over a range of person ability parameters. The method based on the test characteristic curves is the Stocking-Lord approach (Stocking & Lord, 1983). The Stocking-Lord approach determines the two transformation coefficients by finding the optimized values that minimize a loss function which describes the summation of the squared differences between the test characteristic curves based on the common items in the tests that need to be equated over a range of person ability parameters. Detailed descriptions of applying the moment and characteristic curve methods to obtain *A* and *B* can be found in Hanson and Béguin (2002) and Kolen and Brennan (2004).

Standard Errors of the IRT Parameter Scale Transformation Coefficients

Given that the transformation coefficients *A* and *B* are functions of the estimated parameters of the common items, they are subject to errors carried over from item calibrations due to sampling variation (Battauz, 2015a). The amount of random error in the estimates of the two coefficients associated with sampling variability can be quantified by the standard errors of the estimates of the coefficients. There are primarily two dominant methods that have been extensively used in research and practical settings to obtain the standard errors of the two coefficients: *the bootstrap method* (Efron & Tibshirani, 1993; Kolen & Brennan, 2004; Tsai, Hanson, Kolen, & Forsyth, 2001) and *the delta method* (Battauz, 2013; Ogasawara, 2000, 2001a, 2001b; Wong, 2015). In this study, building on the work of Li and Lissitz (2004), we evaluated the performance of another approach, a multiple imputation method which is based on the imputed values of the item parameter estimates, in obtaining the standard errors for the estimates of the IRT parameter scale transformation coefficients. After providing a review of the bootstrap method and the delta method, the multiple imputation method is introduced.

Bootstrap Method

The bootstrap technique offers a viable alternative to estimate the standard error of an estimator when the standard error of this estimator is mathematically intractable (Efron & Tibshirani, 1993). It is a resampling approach involving repeated analyses of a large number of bootstrap samples that are generated in either a parametric fashion or a nonparametric fashion. In this study, we only focus on the nonparametric bootstrap method. This method relies on randomly drawing multiple bootstrap samples from the original data with replacement. The inference of the bootstrap standard error for an estimator can be made by examining the sampling distribution of the statistics relevant to the estimator calculated over these bootstrap samples (Patton, Cheng, Yuan, & Diao, 2014). The bootstrap technique has been applied to obtain the standard errors of IRT equating under the CINEG design (e.g., Tsai et al., 2001).

In the current study, the steps for obtaining the standard errors for the estimates of the two IRT parameter scale transformation coefficients using the bootstrap method are outlined below.

1. Two bootstrap data sets for Test 1 and Test 2 respectively were generated by randomly drawing samples from the original data of Test 1 and Test 2 with replacement.
2. IRT calibrations were produced separately on each of the two bootstrap data sets generated at Step 1 to obtain the item parameter estimates for Test 1 and Test 2.

3. Based on the two sets of item parameter estimates for the common items of Test 1 and Test 2 obtained at Step 2, IRT equating was conducted to obtain the estimates of the two transformation coefficients (i.e., $\hat{A}$ and $\hat{B}$) using the two moment methods (i.e., the mean/mean approach and the mean/sigma approach) and the two characteristic curve methods (i.e., the Haebara approach and the Stocking-Lord approach), respectively.
4. The above steps (steps 1 to 3) were repeated R times, where R was a large number which was 1,000 in the study (i.e., $R = 1,000$).
5. The standard deviations of the estimated values of the two transformation coefficients (i.e., $\hat{A}$ and $\hat{B}$) were calculated over the R replications to obtain the corresponding bootstrap standard errors of the two coefficients.

Delta Method

Another method that can be used in place of the bootstrap method to estimate the standard errors for the estimates of the two transformation coefficients is the delta method (Battauz, 2013; Ogasawara, 2000, 2001a,b; Wong, 2015). The delta method is an analytical approach that makes use of Taylor expansion. It can provide good approximation to the standard errors of IRT test equating coefficients if the estimated parameters have a high probability of being close enough to their true values and the equating functions involving the item parameters can be differentiated around their true values (Wong, 2015). Provided that the asymptotic variance-covariance matrices of the item parameter estimates are available, the standard errors of the estimates of A and B can be obtained using the delta approach. Ogasawara (2000, 2001a) derived the equations by the delta method to calculate the standard errors for $\hat{A}$ and $\hat{B}$ that were obtained using the moment methods and the characteristic curve methods under the CINEG equating design involving the 2PL IRT model.

Following the notations used in the study of Ogasawara (2001a), let $\gamma = ({r_1}', {r_2}')'$ denote the vector of the item parameters of the common items for Test 1 and Test 2, where $\gamma_1 = (a_{11}, b_{11}, a_{12}, b_{12}, \ldots, a_{1k}, b_{1k})'$ and $\gamma_2 = (a_{21}, b_{21}, a_{22}, b_{22}, \ldots, a_{2k}, b_{2k})'$. The variance-covariance matrix of $\hat{A}$ and $\hat{B}$, by using the delta method, can be estimated as

$$\mathrm{acov}[(\hat{A}, \hat{B})'] = \frac{\partial(A, B)'}{\partial\gamma'} \mathrm{acov}(\hat{\gamma}) \frac{\partial\,(A, B)}{\partial\gamma}, \tag{5}$$

where $\frac{\partial(A,B)'}{\partial\gamma'}$ is a matrix with elements that are the partial derivatives of A and B with respect to the item parameters (γ), with $[\frac{\partial(A,B)'}{\partial\gamma'}]' = \frac{\partial(A,B)}{\partial\gamma}$, and $\mathrm{acov}(\hat{\gamma})$ is the asymptotic variance-covariance matrix of the item parameter estimates for the common items. The squared root values of the elements on the diagonal of the matrix $\mathrm{acov}[(\hat{A}, \hat{B})']$ are the standard errors of $\hat{A}$ and $\hat{B}$.

The delta method has been widely used to develop equations for estimating the standard errors of IRT equating. Battauz (2013) utilized the delta method to develop an approach to estimate the standard errors for the IRT parameter scale transformation coefficients in the case of test equating with a complex linkage plan (i.e., chain equating). Based on the delta method, Ogasawara (2001b) derived the equations for calculating the standard errors of the IRT true score equating coefficients.

Wong (2015) recently extended the derivations of Ogasawara (2001b) to formulate a delta approach to estimate the standard errors of IRT true score equating coefficients involving the polytomous IRT models including the generalized partial credit model (GPCM; Muraki, 1992) and the graded response model (GRM; Samejima, 1969). Andersson (2016) applied the delta method to derive the equations to estimate the standard errors of the IRT observed score kernel equating with polytomous IRT models.

Multiple Imputation Method: An Approach Based on Multiple Imputation of the Item Parameter Estimates

If the IRT item parameters are estimated using the maximum likelihood estimation method (e.g., Bock & Aitkin, 1981), the estimators of the item parameters are, under suitable regularity conditions, asymptotically multivariate normally distributed in large samples (Andersson & Wiberg, 2017; Mislevy & Sheehan, 1989; Mislevy, Wingersky, & Sheehan, 1994; Tsutakawa, 1984; Yang, Hansen, & Cai, 2012). More specifically, for a 2PL IRT model, the limiting distribution of the MLE estimators of the item discrimination and difficulty parameters is a multivariate normal distribution with the mean vector γ_0 and variance-covariance matrix $\text{acov}(\gamma_0)$, where γ_0 is the vector of the population item discrimination and difficulty parameter values and $\text{acov}(\gamma_0)$ is the inverse of the Fisher information matrix. In large samples, with the knowledge of the MLE estimates of the item parameters ($\hat{\gamma}$) and the variance-covariance matrix of the estimates ($\text{acov}(\hat{\gamma})$)), the true distribution of the item parameters can be reasonably approximated by $\text{MVN}(\hat{\gamma},\ \text{acov}(\hat{\gamma}))$, where **MVN** denotes multivariate normal distribution (Mislevy et al., 1994; Thissen & Wainer, 1990; Yang et al., 2012). Multiple sets of plausible values that are randomly drawn from $\text{MVN}(\hat{\gamma},\ \text{acov}(\hat{\gamma}))$ can be used to examine the effects of the uncertainty of item parameters on the estimands or statistics of interest that are derived based on the item parameter estimates. For example, Yang et al. (2012) applied the multiple imputation method to examine how the item parameter uncertainty contributed to the variability in IRT person score estimates under various conditions. Raju et al. (2009) used a similar approach, which was called the item parameter replication method in their differential item functioning (DIF) study, to assess the statistical significance of the noncompensatory DIF index within the differential functioning of items and tests framework.

This multiple imputation method can also be used in IRT test equating to obtain the standard errors of IRT equating. Li and Lissitz (2004) suggested a similar approach to analytically derive the simulation-based standard errors of IRT equating coefficients through repeatedly sampling item parameter estimates based on the analytical variance-covariance matrix of the item parameter estimates. According to their approach, IRT equating could be conducted on each of the replicated sets of item parameter estimates. The standard errors of the IRT equating coefficients were determined by the standard deviations of the estimates of IRT equating coefficients calculated over a large number of replications.

Building on the approach suggested by Li and Lissitz (2004), in this study, we attempted to compare the performance of the multiple imputation method with that of the bootstrap method and the delta method in obtaining the standard errors for the estimates of the IRT parameter scale transformation coefficients. In contrast with the procedure suggested by Li and Lissitz (2004), the multiple imputation method tested in the current study used a full rather than a block-diagonal variance-covariance matrix to draw the plausible values of the item parameter estimates. The covariance of the item parameter estimates should not be ignored in the analysis of consequences of the uncertainty of item parameters when IRT calibration is conducted with any finite sample (Patton et al., 2014; Yang et al., 2012). The steps of the multiple imputation method evaluated in the current study are outlined below.

1. IRT calibrations were produced separately on Test 1 and Test 2 to obtain the item parameter estimates ($\hat{\gamma}$) as well as their corresponding variance-covariance matrix (acov($\hat{\gamma}$)) for the two tests. Two procedures which were introduced below were used for estimating the variance-covariance matrices: *empirical cross-product* and *supplemented expectation maximization.*
2. Based on the item parameter estimates ($\hat{\gamma}$) and the variance-covariance matrices (acov($\hat{\gamma}$)) obtained at Step 1, one set of plausible values of the item parameter estimates were drawn from a multivariate normal distribution with mean vector $\hat{\gamma}$ and variance-covariance matrix acov($\hat{\gamma}$) (i.e., MVN($\hat{\gamma}$, acov($\hat{\gamma}$))) for Test 1 and Test 2, respectively.
3. Based on the two sets of imputed item parameter estimates obtained at Step 2, IRT equating was conducted to get the estimates of the two transformation coefficients (i.e., $\hat{A}$ and $\hat{B}$) using the two moment methods (i.e., the mean/mean approach and the mean/sigma approach) and the two characteristic curve methods (i.e., the Haebara approach and the Stocking-Lord approach), respectively.
4. Step 2 and step 3 were repeated R times, where R was a large number which was 1,000 in the current study (i.e., $R = 1,000$).
5. The standard errors of the two transformation coefficients were calculated from the standard deviations of the estimated values of A and B derived from the R replications.

Variance-Covariance Matrix for IRT Item Parameter Estimates (acov($\hat{\gamma}$))

The variance-covariance matrix of the item parameter estimates, which is required in both the delta method and the multiple imputation method, can be obtained from the IRT calibration by inverting the information matrix (i.e., acov($\hat{\gamma}$) $= I^{-1}$ ($\hat{\gamma}$), where $I(\hat{\gamma})$ is the information matrix) (Bock & Aitkin, 1981; Bock & Lieberman, 1970; Ogasawara, 2000). There are at least three procedures which can be used to obtain the information matrix: *Fisher information (FIS), empirical cross-product (XPD)*, and *supplemented expectation maximization (SEM)* (Paek & Cai, 2014). Other available estimators of the variance-covariance matrix for the IRT parameter estimates, such as the sandwich estimator (Yuan, Cheng, & Patton, 2014) which can be used when the model is misspecified, are not considered in this investigation.

The calculation of the exact expected FIS information matrix requires the marginalization over the sample space of responses to all the items. Therefore, the computational burden of deriving FIS increases exponentially as the test length increases (Paek & Cai, 2014). The derivation will become practically infeasible when the test length is large (Bock & Aitkin, 1981; Bock & Lieberman, 1970). Thus, this approach was not considered in this study. The XPD information matrix, which is based on observed response patterns, has been widely used in practice to approximate the FIS matrix. It is also called the observed information matrix (Efron & Hinkley, 1978; Yuan et al., 2014). The XPD information matrix indicates the precision with which the item parameters have been estimated from the realized sample (Efron & Hinkley, 1978; Mislevy & Sheehan, 1989). Another approach, which was introduced by Cai (2008) in IRT calibration to compute the information matrix for deriving the variance-covariance matrix of the item parameter estimates, is based on the supplemented expectation maximization algorithm (SEM; Meng & Rubin, 1991). SEM is a supplementary procedure to the expectation maximization (EM) algorithm to compute the asymptotic variance-covariance matrix for the maximum likelihood estimates of the parameters. The SEM procedure derives the observed information matrix from the difference between the information matrix obtained from the complete data and the information matrix obtained from the missing data (Paek & Cai, 2014). In a simulation study, Paek and Cai (2014) compared the SEM procedure with the FIS and XPD procedures for estimating the standard errors of the IRT item parameter estimates. Their results indicated that all three procedures produced similar results with respect to the bias in the standard error estimates for most of the simulated conditions. However, the SEM procedure was preferred to the XPD procedure when the number of items was large and the sample size was small. In the current study, both the XPD and SEM information matrices were used to derive the variance-covariance matrix of the item parameter estimates. For more details of XPD and SEM, refer to Cai (2008), and Paek and Cai (2014).

Research Purpose

The purpose of this article is twofold. The primary goal of the investigation is to evaluate the multiple imputation method and compare its performance with that of the bootstrap method and the delta method in estimating the standard errors of the IRT parameter scale transformation coefficients. Although a similar approach to the multiple imputation method has been suggested by Li and Lissitz (2004), to our knowledge there is a lack of systematic investigation of the performance of this method as well as the comparison between this approach and other existing popular methods such as the bootstrap method and the delta method in IRT equating studies. In this research, by using both simulated and real data, we attempted to provide an initial evaluation of the performance of the multiple imputation method in obtaining the standard errors of the IRT parameter scale transformation coefficients in the context of the CINEG equating design involving the 2PL IRT model. The second purpose of this research is to examine the effects of different estimation procedures for the variance-covariance matrix of the item parameter estimates on the estimation of the standard errors of the transformation coefficients. Despite the availability of different procedures for obtaining the variance-covariance matrix of the IRT item parameter estimates, less is known about how different types of variance-covariance matrices of the item parameter estimates, which are required in both the delta method and the multiple imputation method, affect the estimates of the standard errors for the estimates of the two linking coefficients. In this investigation, two types of variance-covariance matrices, the XPD matrix and the SEM matrix, were compared with respect to their performances in obtaining the standard errors for the estimates of A and B in IRT test equating.

Simulation Study

Method

In this simulated design, as stated in the introduction, two tests (Test 1 and Test 2), each of which was composed of 40 dichotomously scored items, were assumed to be administered to two independent and nonequivalent groups (Group 1 and Group 2). The population item discrimination parameters for Test 1 and Test 2 were both randomly drawn from a uniform distribution with the range (.7, 1.5) (i.e., $a \sim U(0.7, 1.5)$). The item difficulty parameters for the two tests were both randomly generated from a standard normal distribution (i.e., $b \sim N(0, 1)$). These values were considered based on the realistic item parameter estimates that were obtained from operational assessment programs in practice. It was assumed that the estimates of the IRT parameters of Test 2 were equated to the scale of Test 1. The manipulated factors for data generation included the number of common items (CI), sample size (N), and the differences in ability distributions between the two groups, all of which have been identified as important factors affecting the standard errors for the estimates of the IRT equating coefficients (Battauz, 2015a; Ogasawara, 2000, 2001a,b). Two levels of the number of common items were used: $CI = 10$ and $CI = 20$. These two levels have been used in Battauz's (2015a) study which also assumed that the test length was equal to 40. The first level ($CI = 10$) was simulated by following the rule of thumb which suggested that at least 20% of total items should be used as common items in test equating (Cook & Eignor, 1991; Kolen & Brennan, 2004). The second level with the large number of common items ($CI = 20$) simulated a condition in which the ratio of the number of common items could be up to 50% of the full-length test (e.g., Kaskowitz & De Ayala, 2001; S. H. Kim & Cohen, 1998; J. Kim, Lee, Kim, & Kelly, 2009; Ogasawara, 2000). The population discrimination and difficulty parameters for the common items were drawn from the same distributions as those for the entire tests. This ensures that the statistical properties of the set of the common items could mirror those of the entire tests to be equated (Cook & Eignor, 1991). Regarding the sample size, two levels were simulated to investigate the effect of the calibration sample size on the estimation of the standard errors for the estimates of the linking coefficients: $N = 500$ and $N = 1,000$. In terms of the ability distributions, the ability parameters for the examinees in Group 1 for Test 1 were randomly drawn from a standard normal distribution (i.e., $\theta_1 \sim N(0, 1)$). Two levels of ability distributions for examinees in Group 2 for Test 2 were considered: $\theta_2 \sim N(0.1, 1.1^2)$ and $\theta_2 \sim N(0.5, 1.2^2)$. The first level ($N(0.1, 1.1^2)$) was used to examine the case of a CINEG design in which the samples administered the two tests differed somewhat in ability. While the second level ($N(0.5, 1.2^2)$), which has been extensively used in previous simulation studies of test equating (e.g., Andersson, 2016; Ogasawara, 2000, 2001a,b; Wong, 2015), was used to simulate a vertical equating situation in which the samples administered the two tests differed significantly in ability. The two levels of group ability differences determine that the population values for the IRT parameter scale transformation coefficients are $A = 1.1$ and $B = 0.1$ for $\theta_2 \sim N(0.1, 1.1^2)$ and $A = 1.2$ and $B = 0.5$ for $\theta_2 \sim N(0.5, 1.2^2)$, respectively. A combination of the different levels under each of three manipulated factors resulted in a total of eight simulated conditions ($2 \times 2 \times 2$). All the item responses under each of the conditions were generated by using the 2PL IRT model. Each of the eight simulated conditions was replicated 100 times.

The simulated response data were calibrated with the 2PL IRT model using the marginal maximum likelihood estimation method (MML; Bock & Aitkin, 1981; Bock & Lieberman, 1970), which were performed with the computer software FlexMIRT 3.0 (Cai, 2015). The convergence criteria were the same as those used in the study of Paek and Cai (2014). More specifically, the maximum allowed numbers of E-steps and M-steps were 1,000 and 500, respectively. The E-step, the M-step, and the SEM convergence tolerances were 10^{-6}, 10^{-9}, and 10^{-3}, respectively. For each of the IRT calibrations, both the XPD and SEM variance-covariance matrices for the item parameter estimates were estimated and stored in external files for subsequent estimation of the standard errors for the estimates of A and B.

The transformation coefficients A and B were estimated using the two moment methods and the two characteristic curve methods, respectively, which were all conducted in the R programming language with the package equateIRT (Battauz, 2015b). The standard errors of the estimates of A and B were obtained using three different types of approaches: *the bootstrap method, the delta method,* and *the multiple imputation method.* For the delta method and the multiple imputation method, two different variance-covariance matrices for the item parameter estimates were used: *XPD* and *SEM*. Therefore, a total of five different methods were applied to obtain the standard errors for the estimates of A and B: *the bootstrap method, the delta method using XPD, the delta method using SEM, the multiple imputation method using XPD,* and *the multiple imputation method using SEM.*

Because the true standard errors for the estimates of A and B are not available, the empirical standard errors (SE_e), which were the empirical standard deviations of the estimated values of the two coefficients calculated over 10,000 replications (Paek & Cai, 2014), were used as criterion standard errors in the study. To evaluate the performances of the five different standard error estimation approaches, bias and root mean square difference (RMSD) were calculated for the estimated standard errors of the two linking coefficients A and B against their corresponding empirical standard errors for each simulated condition across all the 100 replications:

$$\text{Bias}_A = \frac{\sum_{i=1}^{R}(SE_i(\hat{A}) - SE_e(A))}{R}, \tag{6}$$

$$\text{Bias}_B = \frac{\sum_{i=1}^{R}(SE_i(\hat{B}) - SE_e(B))}{R}, \tag{7}$$

$$\text{RMSD}_A = \sqrt{\frac{\sum_{i=1}^{R}(SE_i(\hat{A}) - SE_e(A))^2}{R}}, \tag{8}$$

$$\mathrm{RMSD}_B = \sqrt{\frac{\sum_{i=1}^{R} \left(SE_i(\hat{B}) - SE_e(B)\right)^2}{R}}\ , \tag{9}$$

where $SE_i(\hat{A})$ and $SE_i(\hat{B})$are the estimated standard errors of A and B obtained from a given approach in the ith replication, respectively, $SE_e(A)$ and $SE_e(B)$ are the empirical standard errors of A and B, respectively, and $R = 100$ is the number of replications.

Results

The means and standard deviations of the estimated transformation coefficients (i.e., $\hat{A}$ and $\hat{B}$) calculated over the 100 replications are shown in Table 1. Generally, the results indicated that the means of $\hat{A}$ and $\hat{B}$ were very close to their corresponding population values under each of the simulated conditions.

Table 2 provides a summary of the empirical standard errors as well as the means and standard deviations of the estimated standard errors for the simulated data conditions relating to small differences in group ability distributions (i.e., $\theta_2 \sim N(0.1, 1.1^2)$). Table 3 does the same for the simulated data conditions relating to large differences in group ability distributions (i.e., $\theta_2 \sim N(0.5, 1.2^2)$). Results are also graphically depicted in Figures 1 and 2. The results of bias and RMSD for the estimated standard errors are presented in Tables 4 and 5, respectively.

When the sample size was large, all the results consistently indicate that the standard errors produced by the multiple imputation methods were very close to the empirical standard errors as well as those obtained using the bootstrap method and the delta methods for the estimates of A and B that were derived from the mean/mean method and the two characteristic curve methods. However, for the estimates of the linking coefficient A that were calculated from the mean/sigma method, the multiple imputation method tended to produce slightly larger standard errors than the delta methods using the same type of variance-covariance matrix (i.e., XPD or SEM).

When the sample size was small, the multiple imputation methods and the delta methods that used the same type of variance-covariance matrix produced very similar or nearly identical standard errors for the estimates of A and B that were derived from the mean/mean method and the two characteristic curve methods. In line with the findings with the larger sample size, for the estimates of A calculated using the mean/sigma method, the multiple imputation methods appeared to yield larger standard errors than the delta method. In addition, the standard errors estimated by the multiple imputation methods using the SEM variance-covariance matrices were very close to the empirical standard errors as well as those estimated from the bootstrap methods. The multiple imputation methods using the XPD variance-covariance matrices tended to overestimate the standard errors and exhibited slight upward bias.

The differences between the estimated standard errors obtained from the methods (i.e., the delta method and the multiple imputation method) using the XPD variance-covariance matrices and the SEM variance-covariance matrices under each of the simulated conditions were very small when the sample size was large. However, when the sample size was small, the methods using the SEM variance-covariance

Table 1

Descriptive Statistics of the Estimates of the IRT Parameter Scale Transformation Coefficients Calculated Over the 100 Replications

	$\theta_2 \sim N(0.1, 1.1^2)$								$\theta_2 \sim N(0.5, 1.2^2)$							
	$N = 500$				$N = 1{,}000$				$N = 500$				$N = 1{,}000$			
	CI = 10		CI = 20		CI = 10		CI = 20		CI = 10		CI = 20		CI = 10		CI = 20	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Mean/Mean																
A	1.108	.085	1.100	.073	1.106	.052	1.099	.044	1.200	.091	1.198	.079	1.202	.066	1.201	.055
B	.109	.097	.100	.091	.109	.059	.098	.055	.499	.091	.501	.084	.507	.072	.503	.060
Mean/Sigma																
A	1.104	.117	1.106	.085	1.105	.078	1.102	.053	1.211	.120	1.208	.104	1.195	.098	1.193	.068
B	.107	.094	.100	.090	.109	.060	.099	.055	.503	.094	.504	.082	.503	.070	.500	.060
Haebara																
A	1.108	.082	1.107	.071	1.106	.050	1.103	.043	1.206	.086	1.206	.077	1.202	.066	1.199	.056
B	.095	.083	.104	.078	.109	.051	.098	.052	.499	.086	.502	.083	.501	.064	.497	.055
Stocking-Lord																
A	1.104	.082	1.104	.072	1.105	.048	1.101	.043	1.203	.087	1.201	.077	1.201	.064	1.199	.056
B	.093	.085	.104	.079	.108	.051	.099	.052	.499	.087	.504	.083	.502	.065	.497	.055

Note. CI = number of common items.

Table 2

Descriptive Statistics of the Standard Errors for the Estimates of the IRT Parameter Scale Transformation Coefficients Calculated Over the 100 Replications ($\theta_2 \sim N(0.1, 1.1^2)$)

			A								B							
			MM		MS		HA		SL		MM		MS		HA		SL	
N	CI	Approach	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
500	10	Empirical	.081	–	.116	–	.078	–	.078	–	.093	–	.092	–	.080	–	.081	–
		Bootstrap	.083	.008	.125	.023	.080	.007	.080	.007	.097	.008	.095	.008	.081	.004	.081	.004
		Delta (XPD)	.088	.008	.115	.017	.085	.007	.085	.007	.101	.007	.099	.008	.086	.004	.086	.004
		MI (XPD)	.089	.008	.145	.030	.087	.008	.086	.008	.107	.009	.105	.010	.087	.004	.087	.004
		Delta (SEM)	.081	.007	.103	.015	.078	.006	.078	.006	.091	.006	.090	.006	.079	.004	.079	.004
		MI (SEM)	.082	.007	.126	.026	.080	.007	.079	.007	.096	.008	.094	.008	.080	.004	.080	.004
	20	Empirical	.068	–	.088	–	.066	–	.066	–	.081	–	.081	–	.073	–	.074	–
		Bootstrap	.068	.005	.097	.014	.067	.005	.066	.005	.084	.004	.083	.005	.074	.003	.074	.003
		Delta (XPD)	.073	.006	.092	.010	.071	.006	.071	.005	.088	.005	.087	.005	.079	.003	.078	.003
		MI (XPD)	.073	.006	.114	.023	.072	.006	.072	.006	.093	.007	.092	.007	.079	.003	.079	.003
		Delta (SEM)	.067	.005	.083	.009	.066	.004	.065	.004	.080	.004	.079	.004	.073	.003	.073	.003
		MI (SEM)	.068	.005	.097	.014	.067	.005	.066	.005	.084	.011	.082	.005	.073	.003	.074	.003
1,000	10	Empirical	.057	–	.079	–	.055	–	.055	–	.064	–	.064	–	.056	–	.057	–
		Bootstrap	.058	.003	.082	.009	.056	.003	.056	.003	.066	.003	.065	.004	.057	.002	.057	.002
		Delta (XPD)	.060	.003	.075	.008	.057	.003	.057	.003	.067	.003	.066	.003	.059	.002	.059	.002
		MI (XPD)	.060	.004	.086	.010	.058	.003	.058	.003	.069	.003	.068	.004	.059	.002	.059	.002
		Delta (SEM)	.057	.003	.071	.007	.055	.003	.055	.003	.064	.002	.063	.003	.056	.002	.056	.001
		MI (SEM)	.058	.003	.082	.009	.056	.003	.056	.003	.066	.003	.065	.003	.056	.002	.057	.002

(*Continued*)

Table 2
Continued

			A								B							
			MM		MS		HA		SL		MM		MS		HA		SL	
N	CI	Approach	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
	20	Empirical	.048	–	.060	–	.046	–	.046	–	.057	–	.056	–	.052	–	.053	–
		Bootstrap	.048	.003	.062	.004	.047	.002	.046	.002	.057	.002	.057	.002	.052	.002	.053	.002
		Delta (XPD)	.049	.002	.059	.004	.048	.002	.048	.002	.059	.002	.058	.002	.054	.001	.054	.001
		MI (XPD)	.049	.002	.065	.005	.048	.002	.048	.002	.060	.002	.060	.002	.054	.002	.054	.002
		Delta (SEM)	.047	.002	.057	.004	.046	.002	.046	.002	.056	.001	.056	.002	.051	.001	.051	.001
		MI (SEM)	.047	.002	.062	.005	.046	.002	.046	.002	.057	.002	.056	.002	.052	.002	.052	.002

Note. XPD = empirical cross-product; SEM = supplemented expectation maximization; MI = multiple imputation method; MM = mean/mean approach, MS = mean/sigma approach; HA = Haebara approach; SL = Stocking-Lord approach; CI = number of common items; Empirical = empirical standard errors.

Table 3

Descriptive Statistics of the Standard Errors for the Estimates of the IRT Parameter Scale Transformation Coefficients Calculated Over the 100 Replications ($\theta_2 \sim N(0.5, 1.2^2)$)

			A								B							
			MM		MS		HA		SL		MM		MS		HA		SL	
N	CI	Approach	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
500	10	Empirical	.088		.128		.086		.085		.099		.099		.088		.088	
		Bootstrap	.089	.008	.135	.022	.088	.008	.087	.008	.101	.007	.101	.009	.088	.005	.088	.005
		Delta (XPD)	.095	.008	.126	.019	.093	.008	.092	.008	.106	.007	.103	.008	.093	.005	.093	.005
		MI (XPD)	.096	.009	.159	.039	.095	.008	.094	.008	.113	.010	.111	.010	.094	.005	.094	.005
		Delta (SEM)	.087	.007	.113	.016	.085	.007	.085	.007	.097	.006	.095	.007	.087	.005	.087	.005
		MI (SEM)	.088	.007	.137	.024	.086	.007	.086	.007	.101	.007	.101	.009	.087	.005	.088	.005
	20	Empirical	.073		.098		.071		.071		.088		.086		.080		.080	
		Bootstrap	.074	.006	.108	.015	.073	.006	.072	.006	.089	.005	.088	.006	.080	.004	.081	.004
		Delta (XPD)	.079	.006	.103	.012	.078	.005	.077	.005	.094	.005	.092	.006	.085	.004	.085	.004
		MI (XPD)	.079	.006	.129	.030	.079	.006	.078	.005	.099	.008	.097	.008	.085	.004	.085	.004
		Delta (SEM)	.073	.005	.093	.010	.072	.005	.071	.005	.086	.004	.085	.005	.080	.004	.080	.004
		MI (SEM)	.073	.005	.110	.017	.073	.005	.072	.005	.089	.005	.088	.006	.080	.004	.080	.004
1,000	10	Empirical	.061		.089		.060		.059		.069		.068		.061		.061	
		Bootstrap	.062	.004	.090	.009	.061	.004	.060	.004	.070	.003	.069	.004	.061	.003	.062	.003
		Delta (XPD)	.064	.004	.083	.008	.063	.004	.062	.003	.072	.003	.070	.004	.063	.002	.063	.002
		MI (XPD)	.065	.004	.096	.011	.063	.004	.063	.004	.073	.004	.072	.005	.063	.002	.064	.002
		Delta (SEM)	.062	.004	.078	.007	.060	.004	.060	.003	.069	.003	.067	.004	.061	.002	.061	.002
		MI (SEM)	.062	.004	.091	.009	.061	.004	.060	.004	.070	.003	.069	.004	.061	.002	.061	.002

(*Continued*)

Table 3
Continued

			A								B							
			MM		MS		HA		SL		MM		MS		HA		SL	
N	CI	Approach	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
	20	Empirical	.052		.066		.050		.050		.062		.061		.057		.057	
		Bootstrap	.052	.003	.068	.005	.051	.003	.050	.003	.061	.003	.060	.003	.057	.002	.057	.002
		Delta (XPD)	.053	.003	.066	.005	.052	.003	.052	.003	.063	.002	.062	.003	.058	.002	.058	.002
		MI (XPD)	.053	.003	.073	.006	.052	.003	.052	.003	.064	.003	.063	.003	.058	.002	.059	.002
		Delta (SEM)	.051	.002	.063	.004	.050	.003	.050	.002	.060	.002	.059	.002	.056	.002	.056	.002
		MI (SEM)	.052	.003	.069	.005	.051	.003	.050	.003	.061	.003	.060	.003	.056	.002	.056	.002

Note. XPD = empirical cross-product; SEM = supplemented expectation maximization; MI = multiple imputation method; MM = mean/mean approach, MS = mean/sigma approach; HA = Haebara approach; SL = Stocking-Lord approach; CI = number of common items; Empirical = empirical standard errors.

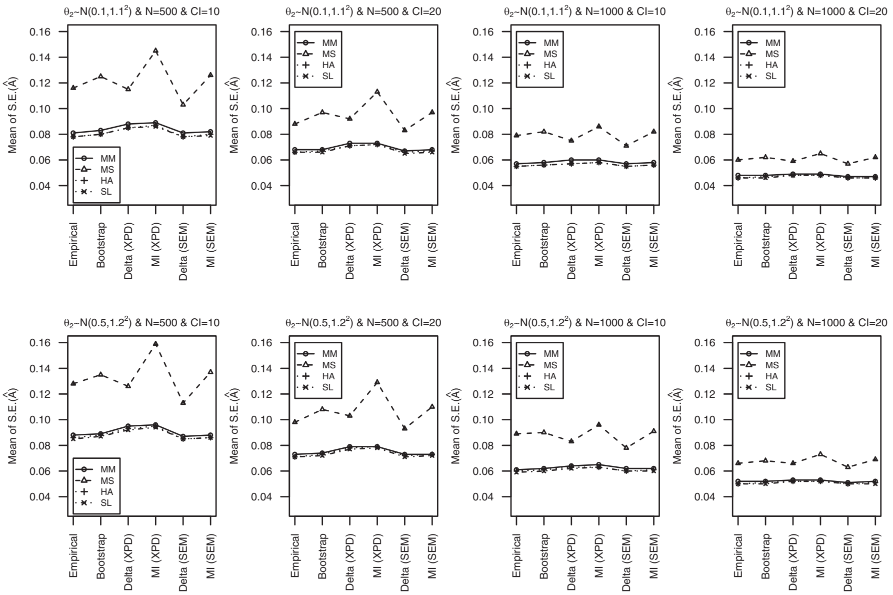

Figure 1. Means of the standard errors for the estimates of the IRT parameter scale transformation coefficients ($\hat{A}$).
Note. MI = multiple imputation method; XPD = empirical cross-product; SEM = supplemented expectation maximization; MM = mean/mean approach; MS = mean/sigma approach; HA = Haeraba approach; SL = Stocking-Lord approach; CI = number of common items; Empirical = empirical standard errors.

matrices seemed to perform better than the methods using the XPD variance-covariance matrices across all the simulated conditions. As shown by the results, both the multiple imputation methods and the delta methods using the SEM variance-covariance matrices produced very similar results to the empirical standard errors as well as that of the bootstrap methods. But the methods using the XPD variance-covariance matrices yielded standard errors that were slightly larger than the empirical standard errors and exhibited more upward bias than the methods using the SEM variance-covariance matrices.

The results also confirm the effects of sample size and the number of common items, as well as the group ability differences on the estimation of the standard errors. With the increase of the sample size and the number of common items, the estimated standard errors for the estimates of *A* and *B* tended to decrease, which is consistent for all the standard error estimation procedures. In addition, larger differences in the ability distributions between the two groups led to greater standard errors.

The estimated standard errors for the estimates of the linking coefficient *A* that were calculated from the mean/sigma method were uniformly larger than those for the estimates that were obtained from the mean/mean method and the two characteristic curve methods under each of the eight manipulated conditions, regardless of the standard error estimation procedures. Moreover, the standard errors for the estimates of the coefficient *B* that were calculated from the two moment methods were

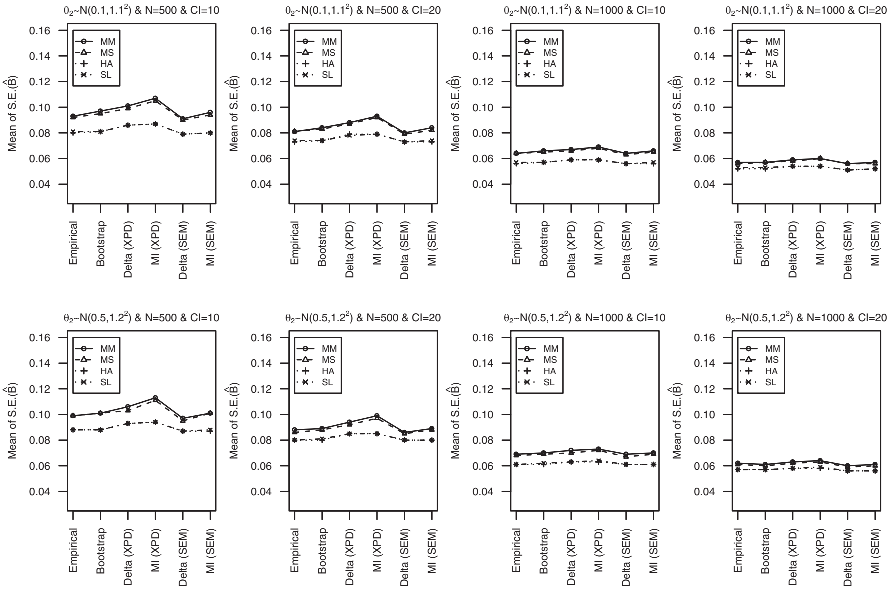

Figure 2. Means of the standard errors for the estimates of the IRT parameter scale transformation coefficients ($\hat{B}$).
Note. MI = multiple imputation method; XPD = empirical cross-product; SEM = supplemented expectation maximization; MM = mean/mean approach; MS = mean/sigma approach; HA = Haeraba approach; SL = Stocking-Lord approach; CI = number of common items; Empirical = empirical standard errors.

slightly greater than those for the estimates that were derived from the two characteristic curve methods, irrespective of the approaches being used for estimating the standard errors.

Empirical Illustration

Method

The empirical data consisted of examinees' responses to two mathematics tests, Test 1 and Test 2, each of which was composed of 30 dichotomously scored items. There were 11 common items between the two tests. The sample sizes for Test 1 data and Test 2 data were 2,757 and 3,057, respectively. The item parameters were estimated with the 2PL IRT model using the MML estimation method (Bock & Aitkin, 1981; Bock & Lieberman, 1970), which were all performed with the computer software FlexMIRT 3.0 (Cai, 2015). The convergence criteria were the same as those used in the simulation study. The descriptive statistics of the unequated item parameter estimates for the two tests as well as for the common items are presented in Table 6. IRT equating was conducted to place the IRT parameters estimated from Test 2 onto the scale of Test 1. The estimates of *A* and *B* were obtained by using the two moment methods and the two characteristic curve methods, which were all performed with the R package equateIRT (Battauz, 2015b). The standard errors for

Table 4

Bias for the Standard Errors of the Estimates of the IRT Parameter Scale Transformation Coefficients

	$\theta_2 \sim N(0.1, 1.1^2)$								$\theta_2 \sim N(0.5, 1.2^2)$							
	A				B				A				B			
	MM	MS	HA	SL	MM	MS	HA	SL	MM	MS	HA	SL	MM	MS	HA	SL
$N = 500$ and CI = 10																
Bootstrap	.002	.009	.002	.002	.003	.003	.001	.001	.002	.007	.001	.002	.002	.002	−.001	.000
Delta (XPD)	.007	−.001	.007	.007	.008	.007	.005	.006	.007	−.002	.007	.007	.007	.005	.004	.005
MI (XPD)	.008	.030	.009	.009	.014	.013	.006	.006	.008	.031	.009	.009	.014	.013	.006	.006
Delta (SEM)	.000	−.012	.000	.000	−.002	−.002	−.001	−.001	.000	−.015	.000	.000	−.002	−.003	−.001	−.001
MI (SEM)	.001	.011	.001	.002	.003	.002	−.001	−.001	.000	.009	.001	.001	.002	.002	−.001	−.001
$N = 500$ and CI = 20																
Bootstrap	.000	.009	.001	.001	.002	.002	.001	.001	.001	.010	.001	.002	.001	.002	.000	.001
Delta (XPD)	.005	.004	.005	.005	.006	.006	.005	.005	.006	.005	.006	.007	.006	.006	.005	.005
MI (XPD)	.005	.026	.006	.006	.012	.011	.006	.006	.006	.031	.007	.008	.012	.012	.005	.005
Delta (SEM)	.000	−.005	.000	.000	−.001	−.001	−.001	−.001	.000	−.005	.000	.001	−.002	−.001	−.001	.000
MI (SEM)	.000	.010	.001	.001	.003	.002	.000	.000	.000	.012	.002	.002	.002	.002	.000	.000
$N = 1{,}000$ and CI = 10																
Bootstrap	.001	.003	.001	.001	.002	.001	.001	.001	.001	.002	.001	.001	.001	.001	.001	.001
Delta (XPD)	.003	−.004	.002	.002	.003	.003	.002	.002	.003	−.006	.003	.003	.003	.001	.002	.002
(XPD)	.003	.008	.003	.003	.005	.004	.002	.003	.004	.007	.004	.003	.005	.004	.003	.003
Delta (SEM)	.000	−.007	.000	.000	.000	.000	.000	.000	.000	−.010	.000	.000	.000	−.002	.000	.000
MI (SEM)	.001	.004	.001	.001	.001	.001	.000	.000	.001	.002	.001	.001	.001	.001	.000	.000

(*Continued*)

Table 4
Continued

	$\theta_2 \sim N(0.1, 1.1^2)$								$\theta_2 \sim N(0.5, 1.2^2)$							
	A				B				A				B			
	MM	MS	HA	SL	MM	MS	HA	SL	MM	MS	HA	SL	MM	MS	HA	SL
$N = 1{,}000$ and CI = 20																
Bootstrap	.000	.002	.000	.000	.000	.000	.000	.000	.000	.002	.000	.000	.000	.000	.000	.000
Delta (XPD)	.001	−.001	.001	.001	.002	.002	.001	.001	.002	.000	.001	.002	.001	.001	.001	.001
MI (XPD)	.001	.005	.002	.002	.003	.003	.001	.001	.002	.006	.002	.002	.003	.003	.001	.001
Delta (SEM)	.000	−.003	.000	.000	−.001	−.001	−.001	−.001	.000	−.003	.000	.000	−.001	−.002	−.001	−.001
MI (SEM)	.000	.002	.000	.000	.000	.000	−.001	−.001	.000	.003	.000	.000	.000	−.001	−.001	−.001

Note. XPD = empirical cross-product; SEM = supplemented expectation maximization; MI = multiple imputation method; MM = mean/mean approach, MS = mean/sigma approach; HA = Haebara approach; SL = Stocking-Lord approach; CI = number of common items.

Table 5

RMSD for the Standard Errors of the Estimates of the IRT Parameter Scale Transformation Coefficients

	$\theta_2 \sim N(0.1, 1.1^2)$								$\theta_2 \sim N(0.5, 1.2^2)$							
	A				B				A				B			
	MM	MS	HA	SL	MM	MS	HA	SL	MM	MS	HA	SL	MM	MS	HA	SL
$N = 500$ and CI = 10																
Bootstrap	.008	.024	.007	.008	.008	.009	.004	.004	.008	.023	.008	.008	.008	.009	.005	.005
(XPD)	.011	.017	.010	.010	.010	.010	.006	.007	.011	.019	.010	.011	.010	.009	.006	.007
MI (XPD)	.012	.042	.012	.012	.017	.016	.007	.008	.012	.050	.012	.012	.017	.016	.008	.008
Delta (SEM)	.007	.020	.006	.006	.006	.007	.004	.004	.007	.022	.007	.007	.006	.008	.005	.005
MI (SEM)	.007	.028	.007	.007	.008	.009	.004	.004	.007	.025	.007	.007	.008	.009	.005	.005
$N = 500$ and CI = 20																
Bootstrap	.005	.016	.005	.005	.005	.005	.003	.003	.006	.018	.006	.006	.005	.006	.004	.004
Delta (XPD)	.008	.011	.007	.008	.008	.008	.006	.006	.008	.013	.008	.009	.008	.009	.006	.007
MI (XPD)	.008	.035	.008	.008	.014	.013	.007	.007	.008	.043	.009	.010	.014	.014	.007	.007
Delta (SEM)	.005	.010	.004	.004	.004	.004	.003	.003	.005	.011	.005	.005	.004	.005	.004	.004
MI (SEM)	.005	.017	.005	.005	.012	.006	.003	.003	.005	.021	.005	.005	.005	.007	.004	.004
$N = 1{,}000$ and CI = 10																
Bootstrap	.003	.009	.003	.003	.004	.004	.002	.002	.004	.009	.004	.004	.003	.004	.003	.003
Delta (XPD)	.004	.008	.004	.004	.004	.004	.003	.003	.005	.010	.005	.005	.004	.004	.003	.003
MI (XPD)	.004	.013	.004	.004	.006	.006	.003	.003	.005	.013	.005	.005	.006	.006	.003	.004
Delta (SEM)	.003	.010	.003	.003	.002	.003	.001	.002	.004	.012	.003	.004	.003	.004	.002	.002
MI (SEM)	.003	.010	.003	.003	.003	.004	.002	.002	.004	.010	.004	.004	.003	.004	.002	.002

(*Continued*)

Table 5
Continued

	$\theta_2 \sim N(0.1, 1.1^2)$								$\theta_2 \sim N(0.5, 1.2^2)$							
	A				B				A				B			
	MM	MS	HA	SL	MM	MS	HA	SL	MM	MS	HA	SL	MM	MS	HA	SL
$N = 1{,}000$ and CI = 20																
Bootstrap	.003	.005	.002	.002	.002	.002	.002	.002	.003	.006	.003	.003	.003	.003	.002	.002
Delta (XPD)	.003	.004	.002	.003	.003	.003	.002	.002	.003	.005	.003	.003	.003	.003	.002	.002
MI (XPD)	.003	.007	.003	.003	.004	.004	.002	.002	.003	.009	.003	.003	.004	.004	.003	.003
Delta (SEM)	.002	.005	.002	.002	.002	.002	.002	.002	.002	.005	.002	.003	.002	.003	.002	.002
MI (SEM)	.002	.005	.002	.002	.002	.002	.002	.002	.003	.006	.003	.003	.003	.003	.003	.003

Note. XPD = empirical cross-product; SEM = supplemented expectation maximization; MI = multiple imputation method; MM = mean/mean approach, MS = mean/sigma approach; HA = Haebara approach; SL = Stocking-Lord approach; CI = number of common items.

Table 6
Descriptive Statistics of the Unequated Item Parameter Estimates for the Real Data

	Item Discrimination (a)		Item Difficulty (b)	
	Mean	*SD*	Mean	*SD*
Test 1 (30 items)	1.056	.431	−.003	1.162
Test 2 (30 items)	1.061	.438	.183	1.098
Test 1 common items (11 items)	1.167	.505	−.003	1.109
Test 2 common items (11 items)	1.153	.485	−.085	1.100

Table 7
Estimates and Standard Errors of the IRT Parameter Scale Transformation Coefficients for the Real Data

		Standard Error				
	Estimate	Bootstrap	Delta (XPD)	MI (XPD)	Delta (SEM)	MI (SEM)
Mean/Mean						
A	.988	.033	.032	.031	.032	.031
B	.081	.046	.044	.046	.043	.043
Mean/Sigma						
A	1.008	.059	.051	.059	.051	.058
B	.083	.049	.047	.049	.046	.046
Haebara						
A	1.020	.032	.030	.031	.031	.031
B	.132	.033	.034	.034	.033	.032
Stocking-Lord						
A	1.008	.033	.030	.030	.030	.030
B	.127	.032	.034	.035	.032	.032

Note. XPD = empirical cross-product; SEM = supplemented expectation maximization; MI = multiple imputation method.

the estimates of A and B were estimated using the five approaches tested in the simulation study.

Results

The estimates of the IRT parameter scale transformation coefficients as well as their corresponding standard errors derived from the five standard error estimation procedures are presented in Table 7. The results shown in Table 7 are generally in line with the results of the simulated data. The multiple imputation method, using either the XPD variance-covariance matrix or the SEM variance-covariance matrix, produced comparable standard errors to those obtained using the bootstrap method and the delta method. The standard errors for the estimates of the coefficient A calculated from the mean/sigma method were larger than those for the estimates obtained from the mean/mean method and the characteristic curve methods. In addition, the standard errors for the estimates of the coefficient B obtained from the characteristic curve methods were generally less than those for the estimates calculated from the moment methods.

Discussion

In this research, we evaluated the performance of the multiple imputation method, which is similar to the approach suggested by Li and Lissitz (2004), in obtaining the standard errors of the two IRT parameter scale transformation coefficients A and B with 2PL IRT model under the CINEG equating design. The multiple imputation method makes use of the plausible values that are drawn from the distribution of the item parameter estimates to obtain the standard errors of IRT equating coefficients. Generally, the results from the simulated data and the empirical data suggested that the multiple imputation method produced similar results to those of the bootstrap method and the delta method. No bias of practical significance was found between the standard errors yielded by these different standard error estimation approaches and the empirical standard errors.

One of the appealing features of considering the multiple imputation method in estimating the standard errors of IRT equating is that it is relatively simple to apply. Despite the popularity of the bootstrap method, the application of the bootstrap method can become computationally demanding with the increase of the number of bootstrap replications as well as the complexities of test equating design and IRT models (Kolen & Brennan, 2004). The computational intensity of this approach mainly comes from a considerable number of IRT calibrations on a large number of bootstrap samples. For example, in this study which involves equating between two tests, a total of 2,000 IRT calibrations had to be run for 2,000 bootstrap data sets (i.e., 1,000 bootstrap data sets for Test 1 and Test 2, respectively). Therefore, the bootstrap technique can become very computationally intensive in the case where a complex linking plan involving multiple test forms is employed. The delta method seems to be the simplest, in terms of application, among the three different types of approaches. However, the flexibility of the application of this method is practically limited in some cases. One of the advantages of employing the delta method is that the standard errors can be directly obtained based on the item parameter estimates and their asymptotic variance-covariance matrix, both of which can be readily acquired through a single IRT calibration for each of the tests to be equated. However, the application of this method relies on the derivation of equations which can become intractable when the equating design, IRT models, and equating methods are complicated (Kolen & Brennan, 2004). For the delta method, specific formulas which involve the matrices for the partial derivatives normally need to be developed depending on the use of the IRT models, equating methods, and/or equating designs (e.g., Andersson, 2016; Andersson & Wiberg, 2017; Battauz, 2013). That is, equations have to be reformulated if either of the varying relevant factors involved in equating (e.g., IRT models, equating designs, and equating methods) changes. For example, the equations that were derived for IRT equating involving the GPCM could not be directly used for obtaining the standard errors for IRT equating coefficients when modeling with the GRM (Wong, 2015). In contrast, the multiple imputation method is less time-consuming than the bootstrap method and can be more flexibly implemented than the delta method (Li & Lissitz, 2004). Compared with the bootstrap method, the multiple imputation method just needs a single IRT calibration to each of the tests to be equated to get the item parameter estimates and their corresponding variance-covariance matrix for drawing plausible values to obtain the standard errors of IRT equating. This significantly reduces the computational intensity. On the other hand, unlike the delta method, the multiple imputation method does not depend on the derivation of complicated equations. Therefore, the relative advantages of the multiple imputation method over the bootstrap method and the delta method make the multiple imputation method an alternative for researchers and practitioners to determine the standard errors of the IRT equating coefficients. Both the bootstrap method and the multiple imputation method rely on a large number of replications. To get stable standard error estimates for the transformation coefficients, 1,000 bootstrap replications were conducted for each condition in the study, which is common in practice (Cui & Kolen, 2008; Kolen, 1998; Kolen & Brennan, 2004). For the multiple imputation method, 1,000 replications were also conducted for each condition so that the differences between the bootstrap method and the multiple imputation method were not confounded by variable numbers of replications. We would recommend that a large number of replications (e.g., 1,000 or more plausible values) be applied when using the multiple imputation method in future. Given the current capacity of computers, 1,000 or more replications for the multiple imputation method, which are less computationally intensive than the same number of replications for the bootstrap method, can be accomplished efficiently. We also suggest that future simulation studies are conducted to investigate the effect of the number of replications on the results of the multiple imputation method for developing more robust practical guidelines for practitioners. Furthermore, in the current study, the performance of the multiple imputation method was evaluated only for estimating the standard errors for the IRT parameter scale transformation coefficients involving the 2PL IRT model. More investigations should be conducted to examine the performance of this method in (a) deriving standard errors of IRT true-score and observed-score equating and (b) cases involving other IRT models such as the 3PL IRT model and polytomous IRT models.

Some of the results of this study echo the findings of previous studies. First, in line with the findings of Battauz (2015a), the results of our simulation study indicate that increasing the calibration sample size and the number of common items could reduce the standard errors. Second, our study suggests that the standard errors for the estimates of A and B that were derived from the characteristic curve methods were smaller than those for the estimates that were calculated from the moment methods, which is also consistent with the findings of previous studies (e.g., Ogasawara, 2001a, 2001b). Last, in accordance with the findings from Ogasawara (2000, 2001a,b) and Wong (2015), the standard errors for the estimates of the linking coefficient A that were calculated using the mean/sigma method were significantly greater than those for the estimates that were derived from the mean/mean method and the two characteristic curve methods, regardless of the standard error estimation procedures being used. This might be due to the instability of the standard deviation of the item difficulty parameter estimates which were used for calculating the coefficient *A* (Baker & Al-Karni, 1991; Ogasawara, 2000). It is notable that, compared with the bootstrap method and the delta method, the multiple imputation method appeared to significantly overestimate the standard errors for the estimates of the coefficient *A* calculated using the mean/sigma method when the sample size was small. Although the mean/sigma method was not recommended for calculating the IRT parameter scale transformation coefficients when performing IRT equating (Ogasawara, 2000), in the unusual case when the mean/sigma method has to be applied the multiple imputation method is not recommended as an alternative to the bootstrap method and the delta method to estimate the standard errors for the estimates of the transformation coefficients.

Both the multiple imputation method and the delta method hinge on the availability of the asymptotic variance-covariance matrix of item parameter estimates, which might limit its wide application if the available IRT software fails to produce the variance-covariance matrix for the item parameter estimates. Fortunately, more and more newly developed computer programs like FlexMIRT 3.0 (Cai, 2015) have the capacity to provide the asymptotic variance-covariance matrix for the item parameter estimates. In this study, two different types of variance-covariance matrices for the item parameter estimates, XPD and SEM, were compared in terms of their effects on the estimation of the standard errors for the two transformation coefficients. Generally, our results indicate that the standard errors derived from the approaches using the XPD matrices were approximately at the same levels as those obtained from the methods using the SEM matrices under each of the simulated conditions when the sample size was large. However, when the sample size was small, the methods using the SEM matrices appeared to perform better than the approaches using the XPD matrices. More specifically, the standard errors produced by the approaches using the SEM matrices were very similar to the criterion standard errors as well as the standard errors obtained from the bootstrap method. But the methods using the XPD matrices tended to yield slightly larger standard errors than the methods using the SEM matrices and the bootstrap method. This might be due to the fact that, as found by Paek and Cai (2014), the XPD procedure tended to produce slight upward bias for the standard errors of the item parameter estimates when the test length was long and the sample size was small. Therefore, the SEM matrix rather than the XPD matrix is recommended as the default choice for estimating the standard errors of the IRT parameter scale transformation coefficients when the number of items is large and the sample is small. When the sample size is large, as suggested by Paek and Cai (2014), the XPD procedure, which is more computationally efficient than the SEM procedure, could be taken as a practically viable option to obtain the variance-covariance matrix for the item parameter estimates for estimating the standard errors of the IRT parameter scale transformation coefficients.

In the current study, the models used for IRT calibrations were assumed to be correctly specified. Perhaps another interesting future study would be to compare the performance of the multiple imputation method with that of the bootstrap method and the delta method in obtaining the standard errors for IRT equating coefficients when the IRT model is misspecified. Under the circumstance of model misspecification, the variance-covariance matrices of the item parameter estimates that are used in the study might not be consistent and an alternative is the robust sandwich-type variance-covariance matrix (Freedman, 2006; Yuan et al., 2014). It is interesting to investigate the effects of different estimators of the variance-covariance matrix, including the sandwich estimator, on the estimation of the standard errors of IRT equating when the IRT model is not correctly specified.

All in all, various factors can affect researchers' and practitioners' choices of the approach for obtaining the standard errors of IRT equating in practice. These factors might be related to the features that characterize different aspects of IRT equating such as sample size, number of common items, IRT models, availability of the variance-covariance matrix of item parameter estimates, computing cost, equating design, and equating methods. The results of the current investigation suggest that the multiple imputation method could be considered as a practically viable alternative to the bootstrap method and the delta method to determine the variability of the IRT equating coefficients. The findings of the current study could facilitate the reporting of standard errors of equating, particularly in cases with complicated linking designs (e.g., chain equating; Battauz, 2013; Kolen & Brennan, 2004), which has been advocated as a standard practice (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Acknowledgments

The authors would like to thank the editors, the associate editor, the anonymous reviewers, Dr. Margaret Wu, and Dr. Matthew Courtney for their helpful and constructive comments.

References

American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andersson, B. (2016). Asymptotic standard errors of observed-score equating with polytomous IRT models. *Journal of Educational Measurement*, *53*, 459–477.

Andersson, B., & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrika*, *82*, 48–66.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*, 147–163.

Battauz, M. (2013). IRT test equating in complex linkage plans. *Psychometrika*, *78*, 464–480.

Battauz, M. (2015a). Factors affecting the variability of IRT equating coefficients. *Statistica Neerlandica*, *69*, 85–101.

Battauz, M. (2015b). equateIRT: An R package for IRT test equating. *Journal of Statistical Software*, *68*, 1–22.

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. 397–472). Reading, MA: Addison-Wesley.

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. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*, 309–329.

Cai, L. (2015). *flexMIRT_ 3.0: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice*, *10*(3), 37–45.

Cui, Z., & Kolen, M. J. (2008). Comparison of parametric and nonparametric bootstrap methods for estimating random error in equipercentile equating. *Applied Psychological Measurement*, *32*, 334–347.

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, *65*, 457–482.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall.

Freedman, D. A. (2006). On the so-called "Huber sandwich estimator" and "robust standard errors." *American Statistician*, *60*, 299–302.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144–149.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*, 3–24.

Kaskowitz, G. S., & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement*, *25*, 39–52.

Kim, J., Lee, W. C., Kim, D. I., & Kelly, K. (2009, April). *Investigation of vertical scaling using the Rasch model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*, 131–143.

Kolen, M. J. (1998). Standard errors of Tucker equating. *Applied Psychological Measurement*, *9*, 209–223.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.), New York: Springer-Verlag.

Li, Y. H., & Lissitz, R. W. (2004). Applications of the analytically derived asymptotic standard errors of item response theory parameter estimates. *Journal of Educational Measurement*, *41*, 85–117.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*, 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139–160.

Meng, X.-L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*, 899–909.

Mislevy, R. J., & Sheehan, K. M. (1989). Information matrices in latent-variable models. *Journal of Educational Statistics*, *14*, 335–350.

Mislevy, R. J., Wingersky, K. M., & Sheehan, K. M. (1994). *Dealing with uncertainty about item parameters: Expected response functions* (Research Report No. 94-28). Princeton, NJ: Educational Testing Service.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review (Otaru University of Commerce)*, *51*, 1–23.

Ogasawara, H. (2001a). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement*, *25*, 53–67.

Ogasawara, H. (2001b). Item response theory true score equatings and their standard errors. *Journal of Educational and Behavioral Statistics*, *26*, 31–50.

Paek, I., & Cai, L. (2014). A comparison of item parameter standard error estimation procedures for unimensional and multidimensional item response theory modelling. *Educational and Psychological Measurement*, *74*, 58–76.

Patton, J. M., Cheng, Y., Yuan, K. H., & Diao, Q. (2014). Bootstrap standard errors for maximum likelihood ability estimates when item parameters are unknown. *Educational and Psychological Measurement*, *74*, 697–712.

Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M. L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement*, *33*, 133–147.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Richmond, VA: Psychometric Society.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Thissen, D., & Wainer, H. (1990). Confidence envelopes for item response theory. *Journal of Educational and Behavioral Statistics*, *15*, 113–128.

Tsai, T. H., Hanson, B. A., Kolen, M. J., & Forsyth, A. R. A. (2001). A comparison of bootstrap standard errors of IRT equating methods for the common-item nonequivalent groups design. *Applied Measurement in Education*, *14*, 17–30.

Tsutakawa, R. K. (1984). Estimation of two-parameter logistic item response curves. *Journal of Educational Statistics*, *9*, 263–276.

Wong, C. C. (2015). Asymptotic standard errors for item response theory true score equating of polytomous items. *Journal of Educational Measurement*, *52*, 106–120.

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, *72*, 264–290.

Yuan, K.-H., Cheng, Y., & Patton, J. (2014). Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika*, *79*, 232–254.

Authors

ZHONGHUA ZHANG is Research Fellow at the Melbourne Graduate School of Education, The University of Melbourne, 100 Leicester Street, Carlton, Victoria, 3053, Australia; zhonghua.zhang@unimelb.edu.au. His primary research interests include test equating, IRT, SEM, and multilevel modeling, as well as the application of quantitative methods in educational and psychological studies (e.g., student learning, teacher education, school leadership).

MINGREN ZHAO is Professor at Normal College, Shenzhen University. 3688 Nanhai Avenue, Nanshan District, Shenzhen, Guangdong Province, China, 518061; mrzhao@szu.edu.cn. His primary research interests include teacher education, research methods in educational studies, and curriculum and instruction.