MULTIPLE EQUATING OF SEPARATE IRT CALIBRATIONS

Abstract

When test forms are calibrated separately, item response theory parameters are not comparable because they are expressed on different measurement scales. The equating process includes in the conversion of item parameter estimates on a common scale and the determination of comparable test scores. Various statistical methods have been proposed to perform equating between two test forms. This paper provides a generalization to multiple test forms of the mean-geometric mean, the mean-mean, the Haebara and the Stocking-Lord methods. The proposed methods estimate simultaneously the equating coefficients that permit the scale transformation of the parameters of all forms to the scale of the base form. Asymptotic standard errors of the equating coefficients are derived. A simulation study is presented to illustrate the performance of the methods.

Key words: equating coefficients, Haebara, item response theory, linking, mean-geometric mean, mean-mean, standard errors, Stocking-Lord.

1. Introduction

When test forms are calibrated separately, item response theory (IRT) parameters are not comparable because they are expressed on different measurement scales. The equating process includes the conversion of item parameter estimates on a common scale and the determination of comparable test scores (Kolen and Brennan, 2014). Various statistical methods have been proposed to perform equating between two test forms. IRT equating methods are generally divided in two classes. The first class of methods, which are based on moments of item parameters, includes the mean-mean (Loyd and Hoover, 1980) and the mean-geometric mean (Mislevy and Bock, 1990) methods. The second class of methods, based on the characteristic curve, includes the Haebara (Haebara, 1980) and the Stocking-Lord (Stocking and Lord, 1983) methods. However, many testing programs use several forms of a test and require the comparability of the scores of every form. To this end, Haberman (2009) developed a regression procedure that generalizes the mean-geometric mean method to the case of multiple test forms. Instead, this paper provides the generalization to multiple test forms of the mean-mean, the Haebara and the Stocking-Lord methods. Furthermore, the asymptotic standard errors of the equating coefficients will be derived for all the methods, including the procedure proposed in the Haberman (2009) paper, where standard errors were not considered. It is worth noting that in this paper standard errors are derived under the assumption of invariance of item parameters across different administrations. As noted by Haberman, Lee and Qian (2009) and by Michaelides and Haertel (2014), when IRT models do not hold perfectly a further source of error is the selection of the set of common items. This issue will be further treated in the paper.

A different approach to multiple equating of forms calibrated separately was proposed in Battauz (2013). In that work, two forms are equated through direct or chain equating coefficients, depending on the connections that can be established between two forms on the basis of the common items. In some cases, two forms can be linked through more than one path, thus yielding a different scale conversion for every path. These transformations can then be averaged in order to obtain a single scale conversion. Chain and average equating coefficients are a function of direct equating coefficients, thus the IRT equating method that is used is chosen only for the computation of direct equating coefficients.

The approach of this paper is rather different and follows the proposal of Haberman (2009). In this work, all the equating coefficients that permit the scale transformation of the IRT parameters of all forms to the scale of the base form are estimated simultaneously. So, for every form, there is only one pair of equating coefficients (the intercept and the slope) without distinction between direct, chain or average equating coefficients.

An alternative to equating forms calibrated separately is given by concurrent calibration. In this case, item parameters of all forms are estimated simultaneously, thus yielding item parameters already on a common scale. As noted by Haberman (2009) and Battauz (2013) this approach is computationally demanding and it could become challenging with thousands of items. When forms are calibrated separately, the full data matrix containing the responses given to the items by every person is not required in the equating process, that is achieved using only the results of the IRT model estimation. This makes approaches based on separate calibrations more manageable. Furthermore, separate calibrations may be preferable because this approach permits a better control of item parameter drift.

The paper is structured as follows. Section 2 presents the methods, including the method proposed by Haberman (2009). In this section, the derivation of the standard errors of the equating coefficients will be given. A procedure to evaluate the variability of the equating coefficients with respect to the choice of common items is presented in Subsection 2.5. Subsection 2.6 briefly illustrates the methods proposed in Battauz (2013), which are used for comparison of the results of the simulation study presented in Section 3. Finally, Section 4 contains the discussion and some concluding remarks.

2. Multiple IRT Equating Methods

In IRT models the probability of a positive response to item j is a function of the latent trait under investigation, denoted by θ, and some item parameters that are related to the characteristics of the items (for a broad review see van der Linden and Hambleton, 1997). The three-parameter logistic model specifies the probability of a positive response as

$$P(\theta; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp\{Da_j(\theta - b_j)\}}{1 + \exp\{Da_j(\theta - b_j)\}}, \tag{1}$$

where a_j, b_j and c_j are item parameters called discrimination, difficulty and guessing, and D is a known constant, typically set to 1.7. The parameters of the model are estimated using the marginal maximum likelihood method (Bock & Aitkin, 1981).

Let t be the index of the form, $t = 1, \dots, T$, while a_{jt} and b_{jt} denote the item discrimination parameter and the item difficulty parameter of item j in the scale of form t. The set of all item parameters is denoted by J, while the set of item parameters administered in form t is denoted by J_t. The number of elements of J is v, and the number of elements of J_t is v_t.

Let a_j^* and b_j^*, $j = 1, \dots, v$, be the item discrimination and difficulty parameters expressed on the scale of the base form. The conversion to the scale of the base form is obtained by applying the following linear transformations

$$a_j^* = \frac{a_{jt}}{A_t} \tag{2}$$

and

$$b_j^* = b_{jt} A_t + B_t, \tag{3}$$

where A_t and B_t are the equating coefficients related to form t. In the following, without loss of generality, Form 1 will be taken as base form. Thus, $A_1 = 1$ and $B_1 = 0$.

In the next subsection the procedure proposed by Haberman (2009) will be introduced. This method will be called the multiple mean-geometric mean (MM-GM) method in this paper, because it is a generalization to several forms of the mean-geometric mean method (also known as log-mean mean method) for two forms.

2.1. Multiple Mean-Geometric Mean

Haberman (2009) proposed to employ Equations (2) and (3) to specify the regression models

$$\log \hat{a}_{jt} = \log \hat{A}_t + \log \hat{a}_j^* + e_{1jt} \tag{4}$$

and

$$\hat{b}_{jt}\hat{A}_t = -\hat{B}_t + \hat{b}_j^* + e_{2jt}, \tag{5}$$

where e_{1jt} and e_{2jt} are the residuals that should be introduced because Equations (2) and (3) hold only approximately in samples. In the first stage, the estimates $\log \hat{A}_t$ and $\log \hat{a}_j^*$ are obtained using the least squares method. In the second stage, the estimates $\hat{A}_t = \exp(\log \hat{A}_t)$ are used to compute the responses $\hat{b}_{jt}\hat{A}_t$ of the regression model (5) and the estimates $\hat{B}_t$ and $\hat{b}_j^*$ are obtained by means of the least square method. The equating coefficients $\hat{A}_1$ and $\hat{B}_1$ are constrained to 1 and 0. As noted by Haberman (2009), the regression analysis corresponds to an analysis of variance when an incomplete two-way layout is considered. The author provides also the equations to be solved for finding in an efficient way the parameter estimates. Here, the regression models will be expressed in matrix form that will be exploited to obtain the asymptotic standard errors of the parameter estimates. Let $\mathbf{x} = (x_i)_{i=1,\dots,n}$ be a vector with elements x_i with $i = 1, \dots, n$ and let $\log(\mathbf{x}) = (\log(x_i))_{i=1,\dots,n}$ be the vector containing the logarithm of x_i. Model (4) is written as

$$\log \hat{\mathbf{a}} = \mathbf{X}_1 \log \hat{\boldsymbol{\beta}}_1 + \mathbf{e}_1, \tag{6}$$

where $\hat{\mathbf{a}} = (\hat{a}_{jt})_{j=1,\dots,v_t,t=1\dots,T}$ is a vector of length $n = \sum_t v_t$ containing the elements $\hat{a}_{jt}$ with $j = 1, \dots, v_t$ and $t = 1, \dots, T$, $\mathbf{X}_1$ is a design matrix with dimension $n \times q$, $q = T + v - 1$, composed by a set of $T - 1$ dummy variables that indicate in which form t was included the item, and a set of v dummy variables that indicate which item j is considered, $\hat{\boldsymbol{\beta}}_1$ is a vector of length q containing the regression coefficients and that is composed by $\hat{\mathbf{A}} = (\hat{A}_2, \ldots, \hat{A}_T)^\top$ and $\hat{\mathbf{a}}^* = (\hat{a}_1^*, \ldots, \hat{a}_v^*)^\top$, and $\mathbf{e}_1$ is a vector of length n containing the residuals. Let $\mathbf{T}$ be a matrix with dimension $n \times (T-1)$, composed by $T-1$ dummy variables that indicate in which form t was administered the item. Let $\hat{\mathbf{A}}_n$=$\mathbf{T}\hat{\mathbf{A}}$ be a vector of length n containing the equating coefficients $\hat{A}_2, \ldots, \hat{A}_T$, each replicated v_t times. Model (5) can then be written as

$$\mathrm{diag}(\hat{\mathbf{A}}_n)\hat{\mathbf{b}} = \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + \mathbf{e}_2, \tag{7}$$

where diag($\cdot$) denotes a diagonal matrix, $\hat{\mathbf{b}} = (\hat{b}_{jt})_{j=1,\ldots,v_t,t=1\ldots,T}$ is a vector of length n containing the elements $\hat{b}_{jt}$ with $j = 1, \ldots, v_t$ and $t = 1, \ldots, T$, $\mathbf{X}_2$ is a design matrix with dimension $n \times q$, composed by a set of $T-1$ dummy variables multiplied by -1 that indicate in which form t was included the item, and a set of v dummy variables that indicate which item j is considered, $\hat{\boldsymbol{\beta}}_2$ is a vector of length q containing the regression coefficients, composed by $\hat{\mathbf{B}} = (\hat{B}_2, \ldots, \hat{B}_T)^\top$ and $\hat{\mathbf{b}}^* = (\hat{b}_1^*, \ldots, \hat{b}_v^*)^\top$, and $\mathbf{e}_2$ is a vector of length n containing the residuals.

Let $\exp(\mathbf{x}) = (\exp(x_i))_{i=1,\ldots,n}$ be the vector containing the exponential of x_i. The estimators of the parameters are given by

$$\hat{\boldsymbol{\beta}}_1 = \exp\left[(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top \log \hat{\mathbf{a}}\right] \tag{8}$$

and

$$\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2^\top\mathbf{X}_2)^{-1}\mathbf{X}_2^\top \mathrm{diag}(\hat{\mathbf{A}}_n)\hat{\mathbf{b}}. \tag{9}$$

Note that this method not only provides estimates of the equating coefficients but also yields an estimate of the item parameters a_j^* and b_j^*, $j = 1, \ldots, v$. This estimate synthesizes the estimates obtained for the same item in different calibrations.

Since the estimates of the equating coefficients $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ and the item parameter estimates $\hat{\mathbf{a}}^*$ and $\hat{\mathbf{b}}^*$ are a function of the item parameter estimates obtained by separate calibrations $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$, the asymptotic standard errors can be derived using the delta method. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$ be the vector containing all the regression coefficients, and $\hat{\boldsymbol{\gamma}} = (\hat{\mathbf{a}}^\top, \hat{\mathbf{b}}^\top)^\top$ be the vector containing all the estimates of discrimination and difficulty parameters. The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is then given by

$$\mathrm{acov}(\hat{\boldsymbol{\beta}}) = \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\gamma}}^\top} \mathrm{acov}(\hat{\boldsymbol{\gamma}}) \frac{\partial \hat{\boldsymbol{\beta}}^\top}{\partial \hat{\boldsymbol{\gamma}}}. \tag{10}$$

The derivatives are given in Appendix A.1.

2.2. Multiple Mean-Mean

From equation (2) it follows that

$$A_t = \frac{a_{jt}}{a_j^*}. \tag{11}$$

If a_j^* were known, the mean-mean estimator of the equating coefficient A_t would be

$$\hat{A}_t = \frac{\sum_{j \in J_t} \hat{a}_{jt}}{\sum_{j \in J_t} \hat{a}_j^*}. \tag{12}$$

The proposal of this paper is to replace $\hat{a}_j^*$ in (12) with

$$\hat{a}_j^* = \frac{\sum_{s \in U_j} \hat{a}_{js}}{\sum_{s \in U_j} \hat{A}_s}, \tag{13}$$

where U_j is the set of forms such that item j is in J_t. Substituting equation (13) in equation (12) it is possible to obtain

$$\hat{A}_t = \frac{\sum_{j \in J_t} \hat{a}_{jt}}{\sum_{j \in J_t} \frac{\sum_{s \in U_j} \hat{a}_{js}}{\sum_{s \in U_j} \hat{A}_s}}, \quad t = 2, \dots, T. \tag{14}$$

This defines a set of $T-1$ nonlinear equations, whose root can be found numerically, after setting

$$\hat{A}_1 = \frac{\sum_{j \in J_1} \hat{a}_{j1}}{\sum_{j \in J_1} \hat{a}_j^*} = 1.$$

Equations (14) can be solved by the Newton-Raphson method. However, another algorithm that can be exploited to accomplish this task is the iterative proportional fitting procedure (Deming and Stephan, 1940). Iterative proportional fitting is generally used to estimate cell probabilities in a contingency table so that row and column classifications satisfy the condition of independence

$$p_{ij} = a_i b_j, \quad \forall i, j, \tag{15}$$

where p_{ij} is the proportion of individuals that fall in the ith row and jth column of the table, while a_i and b_j are positive constants (see for example Goodman, 1968). Writing equation (11) as follows

$$a_{jt} = a_j^* A_t,$$

shows the similarity of the case under study with the condition of independence. However, in the present case, the matrix containing the values $\hat{a}_{jt}$ has missing entries because not all items are included in every form. In order to handle the missing entries, it is necessary to resort to the concept of quasi-independence (see Goodman, 1968 and the references therein), which requires that relation (15) holds for the non-missing cells. It follows that

$$\sum_j p_{ij} = a_i \sum_j \delta_{ij} b_j, \qquad \sum_i p_{ij} = b_j \sum_i \delta_{ij} a_i, \tag{16}$$

where $\delta_{ij} = 1$ if the entry in the ith row and jth column if not missing. In the present case, equation (16) implies

$$\sum_{s \in U_j} a_{js} = a_j^* \sum_{s \in U_j} A_s, \qquad \sum_{j \in J_t} a_{jt} = A_t \sum_{j \in J_t} a_j^*,$$

which leads to the estimators (12) and (13). The algorithm proposed by Goodman (1968), adapted to the case under investigation, is as follows. The starting points are given by

$$\hat{a}_j^{*0} = \frac{\sum_{s \in U_j} \hat{a}_{js}}{u_j}, \quad \text{for } j = 1, \ldots, v,$$

where u_j is the number of elements of U_j. For $m \geq 1$, compute the following equations until convergence

$$\hat{A}_t^{(2m-1)} = \frac{\sum_{j \in J_t} \hat{a}_{jt}}{\sum_{j \in J_t} \hat{a}_j^{*(2m-2)}}, \quad \text{for } t = 1, \ldots, T,$$

and

$$\hat{a}_j^{*(2m)} = \frac{\sum_{s \in U_j} \hat{a}_{js}}{\sum_{s \in U_j} \hat{A}_s^{(2m-1)}}, \quad \text{for } j = 1, \ldots, v.$$

Finally, the following step is required to impose $\hat{A}_1 = 1$:

$$\hat{A}_t = \frac{\hat{A}_t^{(2m-1)}}{\hat{A}_1^{(2m-1)}}, \qquad \hat{a}_j^* = \hat{a}_j^{*(2m)} \hat{A}_1^{(2m-1)}.$$

Once the estimates $\hat{A}_2, \ldots, \hat{A}_T$ are obtained, the estimates of the equating coefficients $B_2, \ldots, B_T$ can be obtained following the procedure of the MM-GM method, explained in Subsection 2.1.

When $T = 2$, this method is equivalent to the mean-mean method. For this reason, this method will be called the multiple mean-mean (MM-M) method in this paper. The proof is given in Appendix B.

Also in this case, asymptotic standard errors of both the equating coefficients and the synthetic item parameters can be obtained using the delta method, as in equation (10). The derivatives necessary to compute the covariance matrix are given in Appendix A.2.

2.3. Multiple Item Response Function

The multiple item response function (MIRF) method is a generalization of the Haebara method to the case of multiple forms. The proposal of this paper is to find the equating coefficients by minimizing the following function

$$f^*_{IR} = \sum_{t=1}^{T} \int_{-\infty}^{\infty} \sum_{j \in J_t} \left(P_{jt} - P^*_{jt}\right)^2 h(\theta) d\theta, \tag{17}$$

where $h(\cdot)$ is the density of a standard normal distribution and

$$P_{jt} = P(\theta; \hat{a}_{jt}, \hat{b}_{jt}, \hat{c}_{jt}) \tag{18}$$

is the probability of a positive response to item j using the item parameters estimated for administration t, while

$$P^*_{jt} = P(\theta; \hat{a}^*_{jt}, \hat{b}^*_{jt}, \hat{c}_{jt}), \tag{19}$$

is the probability of a positive response to item j using the synthetic discrimination and difficulty parameters. These parameters are converted on the scale of Form t using the following equations

$$\hat{a}^*_{jt} = \hat{a}^*_j \hat{A}_t \quad \text{and} \quad \hat{b}^*_{jt} = \frac{\hat{b}^*_j - \hat{B}_t}{\hat{A}_t}, \tag{20}$$

where

$$\hat{a}^*_j = \frac{1}{u_j} \sum_{s \in U_j} \frac{\hat{a}_{js}}{\hat{A}_s} \quad \text{and} \quad \hat{b}^*_j = \frac{1}{u_j} \sum_{s \in U_j} (\hat{b}_{js} \hat{A}_s + \hat{B}_s), \tag{21}$$

thus yielding

$$\hat{a}_{jt}^{*} = \frac{1}{u_j} \sum_{s \in U_j} \frac{\hat{a}_{js}}{\hat{A}_s} \hat{A}_t \quad \text{and} \quad \hat{b}_{jt}^{*} = \frac{\frac{1}{u_j} \sum_{s \in U_j} (\hat{b}_{js} \hat{A}_s + \hat{B}_s) - \hat{B}_t}{\hat{A}_t}. \tag{22}$$

In order to obtain the conversion to the scale of Form 1, the constraints $A_1 = 1$ and $B_1 = 0$ are imposed. The MIRF method here proposed, satisfies the symmetry property (Kolen and Brennan, 2014, p. 9). The proof is given in Appendix C.

Since the integrals in Equation (17) do not have a closed-form solution, function f_{IR}^{*} is approximated using Gaussian quadrature

$$f_{IR} = \sum_{t=1}^{T} \sum_{m=1}^{r} \sum_{j \in J_t} \left(P_{mjt} - P_{mjt}^{*} \right)^2 H(y_m), \tag{23}$$

where

$$P_{mjt} = P(y_m; \hat{a}_{jt}, \hat{b}_{jt}, \hat{c}_{jt}), \quad P_{mjt}^{*} = P(y_m; \hat{a}_{jt}^{*}, \hat{b}_{jt}^{*}, \hat{c}_{jt}), \tag{24}$$

y_m, $m = 1, \ldots, r$, are quadrature points and $H(y_m)$ are appropriate weights. The minimization is performed using numerical methods (see for example Kim and Kolen, 2007, for the case $T = 2$). Once the equating coefficients are obtained, the synthetic item parameters can be computed using equations (21).

The covariance matrix of the equating coefficients and the synthetic item parameters are again obtained using the delta method. The partial derivatives of the equating coefficients with respect to the estimated item parameters, required to apply the delta method, are obtained using implicit differentiation as in Ogasawara (2001b)

$$\frac{\partial (\hat{\mathbf{A}}^\top, \hat{\mathbf{B}}^\top)^\top}{\partial \hat{\boldsymbol{\gamma}}^\top} = - \left[\frac{\partial \mathbf{S}_{IR}}{\partial (\hat{\mathbf{A}}^\top, \hat{\mathbf{B}}^\top)} \right]^{-1} \frac{\partial \mathbf{S}_{IR}}{\partial \hat{\boldsymbol{\gamma}}^\top}, \tag{25}$$

where $\mathbf{S}_{IR}$ is the vector containing the partial derivatives of f_{IR} with respect to the vectors of equating coefficients $\mathbf{A}$ and $\mathbf{B}$. The partial derivative with respect to the equating coefficients $\hat{A}_k$, $k = 1, \ldots, T$, is given by

$$\frac{\partial f_{IR}}{\partial \hat{A}_k} = -\sum_{t=1}^{T} \sum_{m=1}^{r} \sum_{j \in J_t} \left(P_{mjt} - P^*_{mjt}\right) \frac{\partial P^*_{mjt}}{\partial \hat{A}_k} H(y_m), \tag{26}$$

and the partial derivatives with respect to the equating coefficients $\hat{B}_k$, $k = 1, \ldots, T$, are obtained by substituting $\hat{A}_k$ with $\hat{B}_k$ in equation (26).

The components of $\frac{\partial \mathbf{S}_{IR}}{\partial (\hat{\mathbf{A}}^\top, \hat{\mathbf{B}}^\top)^\top}$ are the second derivatives of f_{IR} with respect to couples of equating coefficients. For the equating coefficients $\hat{A}_k$ and $\hat{B}_h$ they are given by

$$\frac{\partial^2 f_{IR}}{\partial \hat{A}_k \partial \hat{B}_h} = \sum_{t=1}^{T} \sum_{m=1}^{r} \sum_{j \in J_t} \left[\frac{\partial P^*_{mjt}}{\partial \hat{A}_k} \frac{\partial P^*_{mjt}}{\partial \hat{B}_h} - \left(P_{mjt} - P^*_{mjt}\right) \frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{B}_h} \right] H(y_m). \tag{27}$$

The derivatives with respect to other couples of equating coefficients are obtained by substituting $\hat{A}_k$ and $\hat{B}_k$ with other equating coefficients in equation (27).

The components of $\frac{\partial \mathbf{S}_{IR}}{\partial \hat{\gamma}^\top}$ are the second derivatives of f_{IR} with respect to the equating coefficients and the item parameters. For the equating coefficient $\hat{A}_k$ and the difficulty parameter $\hat{b}_{ih}$ they are

$$\frac{\partial^2 f_{IR}}{\partial \hat{A}_k \partial \hat{b}_{ih}} = -\sum_{t=1}^{T} \sum_{m=1}^{r} \sum_{j \in J_t} \left[\left(\frac{\partial P_{mjt}}{\partial \hat{b}_{ih}} - \frac{\partial P^*_{mjt}}{\partial \hat{b}_{ih}} \right) \frac{\partial P^*_{mjt}}{\partial \hat{A}_k} + \left(P_{mjt} - P^*_{mjt}\right) \frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{b}_{ih}} \right] H(y_m). \tag{28}$$

Note that $\frac{\partial P_{mjt}}{\partial \hat{b}_{ih}}$, $\frac{\partial P^*_{mjt}}{\partial \hat{b}_{ih}}$ and $\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{b}_{ih}}$ are 0 when $i \neq j$. The other derivatives are obtained by substituting $\hat{A}_k$ with other equating coefficients and $\hat{b}_{ih}$ with other item parameters in equation (28). All the derivatives entering in equations (26), (27) and (28) are given in Appendix A.3 (see Equations from (A9) to (A54)).

In order to obtain the asymptotic standard errors of the synthetic item parameters, the derivatives of the synthetic item parameters with respect to the estimates of the item parameters obtained from separate calibrations are necessary. These derivatives are given in Appendix A.3 (see Equations from (A55) to (A60)).

2.4. Multiple Test Response Function

The multiple test response function (MTRF) method proposed here is a generalization of the Stocking-Lord method to the case of multiple forms, and requires the minimization of the following objective function

$$f_{TR}^* = \sum_{t=1}^{T} \int \left(\sum_{j \in J_t} P_{jt} - P_{jt}^* \right)^2 h(\theta) d\theta. \tag{29}$$

The response functions P_{jt} and P_{jt}^* are defined in Equations from (18) to (22). The symmetry property is satisfied also by the MTRF method, as proven in Appendix C.

Also in this case, the integrals in equation (29) do not have a closed-form solution and they are approximated using Gaussian quadrature with r points

$$f_{TR} = \sum_{t=1}^{T} \sum_{m=1}^{r} \left(\sum_{j \in J_t} P_{mjt} - P_{mjt}^* \right)^2 H(y_m), \tag{30}$$

where P_{mjt} and P_{mjt}^* are defined in Equation (24). After the estimation of the equating coefficients by means of numerical methods, synthetic item parameters can be computed using Equations (21).

Similarly to the MIRF method, the partial derivatives for obtaining the asymptotic covariance matrix with the delta method are computed as follows

$$\frac{\partial(\hat{\mathbf{A}}^\top, \hat{\mathbf{B}}^\top)^\top}{\partial \hat{\boldsymbol{\gamma}}^\top} = - \left[\frac{\partial \mathbf{S}_{TR}}{\partial(\hat{\mathbf{A}}^\top, \hat{\mathbf{B}}^\top)} \right]^{-1} \frac{\partial \mathbf{S}_{TR}}{\partial \hat{\boldsymbol{\gamma}}^\top}, \tag{31}$$

where the elements of $\mathbf{S}_{TR}$ are

$$\frac{\partial f_{TR}}{\partial \hat{A}_k} = - \sum_{t=1}^{T} \sum_{m=1}^{r} \sum_{j \in J_t} \left(P_{mjt} - P_{mjt}^* \right) \sum_{j \in J_t} \frac{\partial P_{mjt}^*}{\partial \hat{A}_k} H(y_m) \tag{32}$$

for $k = 1, \ldots, T$, and $\partial f_{TR}/\partial \hat{B}_k$ for $k = 1, \ldots, T$, which are obtained analogously.

The components of $\frac{\partial \mathbf{S}_{TR}}{\partial (\hat{\mathbf{A}}^\top, \hat{\mathbf{B}}^\top)^\top}$ for the equating coefficients $\hat{A}_k$ and $\hat{B}_h$ are

$$\frac{\partial^2 f_{TR}}{\partial \hat{A}_k \partial \hat{B}_h} = \sum_{t=1}^{T} \sum_{m=1}^{r} \left[\sum_{j \in J_t} \frac{\partial P^*_{mjt}}{\partial \hat{A}_k} \sum_{j \in J_t} \frac{\partial P^*_{mjt}}{\partial \hat{B}_h} - \sum_{j \in J_t} \left(P_{mjt} - P^*_{mjt} \right) \sum_{j \in J_t} \frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{B}_h} \right] H(y_m), \tag{33}$$

while the derivatives with respect to other couples of equating coefficients are obtained analogously.

The components of $\frac{\partial \mathbf{S}_{TR}}{\partial \hat{\gamma}^\top}$ for the equating coefficient $\hat{A}_k$ and the difficulty parameter $\hat{b}_{ih}$ are

$$\frac{\partial^2 f_{TR}}{\partial \hat{A}_k \partial \hat{b}_{ih}} = -\sum_{t=1}^{T} \sum_{m=1}^{r} \left[\sum_{j \in J_t} \left(\frac{\partial P_{mjt}}{\partial \hat{b}_{ih}} - \frac{\partial P^*_{mjt}}{\partial \hat{b}_{ih}} \right) \sum_{j \in J_t} \frac{\partial P^*_{mjt}}{\partial \hat{A}_k} + \sum_{j \in J_t} \left(P_{mjt} - P^*_{mjt} \right) \sum_{j \in J_t} \frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{b}_{ih}} \right] H(y_m). \tag{34}$$

The derivatives for the other equating coefficients and the other item parameters are obtained analogously. All the derivatives entering in equations (32), (33) and (34) are the same given for the MIRF method, and they are given in Appendix A.3 (see Equations from (A9) to (A54)). The derivatives of the synthetic item parameters with respect to the item parameters separately calibrated are the same provided for the MIRF method and they are given in Appendix A.3 (see Equations from (A55) to (A60)).

Appendix D provides the formulas for the computation of the reliability index and the standard error of estimated abilities after the scale transformation with all the multiple equating methods proposed in this paper.

2.5. Stability of Equating with Respect to the Choice of Common Items

The asymptotic covariance matrices of the equating coefficients derived in this paper are obtained under the assumption of invariance of item parameters across different administrations. However, Haberman et al. (2009) and Michaelides and Haertel (2014) noted that when Equations (2) and (3) do not hold perfectly, the set of common items selected constitutes a further source of error in the equating process. In order to examine the stability of the equating process with respect to the choice of common items, a procedure for the case of multiple test forms based on the proposal of Haberman et al. (2009) and Michaelides and Haertel (2014) will be presented here. Both these works consider two test forms and make use of resampling techniques to evaluate the variability of an equating result, which can be an equating coefficient or an equated score. While the proposal of Haberman et al. (2009) is based on the jackknife method, Michaelides and Haertel (2014) proposed to use the bootstrap method. Both these articles apply resampling techniques to examinees to estimate the sample variability. Furthermore, resampling techniques are also applied to common items in order to quantify the variability of an equating result with respect to the choice of the set of common items. It is worth remarking that, when IRT model assumptions hold perfectly, the choice of the common items would not add variance to the equating transformation (Michaelides and Haertel, 2014). Suppose instead that invariance of item parameters does not hold. So, true equating coefficients depend on the set of common items selected. Let A_m be the true A equating coefficient for the set of common items m. Similarly to Haberman et al. (2009), the variability of A_m can be measured using the sample variance

$$\sigma_A^2 = \frac{1}{M-1}\sum_{m=1}^{M}(A_m - A.)^2,$$

where $A. = M^{-1}\sum_{m=1}^{M} A_m$. An estimate of σ_A^2 is

$$\hat{\sigma}_A^2 = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{A}_m - \hat{A}.)^2,$$

where $\hat{A}. = M^{-1}\sum_{m=1}^{M} \hat{A}_m$. As noted by Haberman et al. (2009, Equation (18)) this estimate is biased. In fact

$$\mathrm{E}(\hat{\sigma}_A^2) = \sigma_A^2 + \Delta_A,$$

where

$$\Delta_A = \frac{1}{M-1}\sum_{m=1}^{M}\mathrm{var}(\hat{A}_m - \hat{A}.).$$

While Haberman et al. (2009) estimate these quantities by means of the jackknife method, the bootstrap method is employed in this paper. Differently from Haberman et al. (2009) and Michaelides and Haertel (2014), here multiple forms should be considered. So, bootstrap samples of examinees are obtained by randomly sampling with replacement from each population of examinees, while bootstrap samples of common items are obtained by randomly sampling with replacement from each set of common items between different forms.

For each t, $t = 1, \dots, T$, let $\hat{A}_{t(m)}$ be an estimate of A_t using the bootstrap sample of common items m, $m = 1, \dots, M$, and $\hat{A}_{t(m,b)}$ be an estimate of A_t using the bootstrap sample of common items m and the bootstrap sample of examinees b, $b = 1, \dots, B$. In order to speed up the computational time, in this paper item parameters are estimated only after resampling of examinees. Resampling of common items is carried out by resampling of estimated item parameters, which are kept constant for each b.

The bootstrap estimate of $\sigma^2_{A_t}$ is then

$$\hat{\sigma}^2_{boot}(A_t) = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{A}_{t(m)} - \hat{A}_{t(.)})^2 - \hat{\Delta}_{boot}(A_t),$$

where

$$\hat{\Delta}_{boot}(A_t) = \frac{1}{M-1}\frac{1}{B-1}\sum_{m=1}^{M}\sum_{b=1}^{B}(\hat{A}_{t(m,b)} - \hat{A}_{t(\cdot,b)} - \hat{A}_{t(m,\cdot)} + \hat{A}_{t(\cdot,\cdot)})^2,$$

and $\hat{A}_{t(\cdot)} = M^{-1}\sum_{m=1}^{M}\hat{A}_{t(m)}$, $\hat{A}_{t(\cdot,b)} = M^{-1}\sum_{m=1}^{M}\hat{A}_{t(m,b)}$, $\hat{A}_{t(m,\cdot)} = B^{-1}\sum_{b=1}^{B}\hat{A}_{t(m,b)}$ $\hat{A}_{t(\cdot,\cdot)} = M^{-1}B^{-1}\sum_{m=1}^{M}\sum_{b=1}^{B}\hat{A}_{t(m,b)}$.

The difference here with respect to the work of Michaelides and Haertel (2014) is that in their paper there is not the correction $\hat{\Delta}_{boot}(A_t)$, so the estimated variability is positive even if the IRT model assumptions hold perfectly, due to the sample variability of $\hat{A}_{t(m)}$.

The bootstrap estimate of $\sigma^2_{B_t}$ can be obtained analogously.

2.6. Chain and Bisector Equating Coefficients

A different approach to equate multiple forms is given in Battauz (2013). Since the methods proposed in this paper will be compared by means of simulations to the bisector and the weighted bisector methods proposed in Battauz (2013), here these methods will be briefly described.

Suppose that two forms are linked through a chain of forms that presents common items in pairs. Define the path from Form 0 to Form l as $p = \{0, 1, \ldots, l\}$. Chain equating coefficients transforming the scale of Form 0 to that of Form l can be obtained as a function of the direct equating coefficients

$$A_p = A_{0,1,\dots,l} = \prod_{g=1}^{l} A_{g-1,\,g}$$

and

$$B_p = B_{0,1,\dots,l} = \sum_{g=1}^{l} B_{g-1,\,g}\; A_{g,\dots,l}\,,$$

where $A_{g,\dots,l} = \prod_{h=g+1}^{l} A_{h-1,\,h}$ is the coefficient that links Form g to Form l, while $A_{g-1,\,g}$ and $B_{g-1,\,g}$ are direct equating coefficients between Forms $g-1$ and g. When two forms can be linked through different paths, the transformations provided by each path can be averaged. Define the set of paths that link Forms 0 and l as $\mathcal{P}_{0l}$ and the linking coefficients related to path p as A_p and B_p, $p \in \mathcal{P}_{0l}$. In order to average the transformations provided by each path, the bisector method proposed by Battauz (2013) yields the equating coefficients

$$A^*_{0l} = \sum_{p\in\mathcal{P}_{0l}} A_p w_p \quad \text{and} \quad B^*_{0l} = \sum_{p\in\mathcal{P}_{0l}} B_p w_p,$$

where

$$w_p = \frac{n_p(1+A_p^2)^{-1/2}}{\sum_{b\in\mathcal{P}_{0l}} n_b(1+A_b^2)^{-1/2}},$$

and n_p are optional weights. The weighted bisector method is obtained when the weights n_p are determined by minimizing the average variance of θ^*_l, namely

$$\mathrm{E}_{\theta_0}\left[\mathrm{var}(\hat{A}^*_{0l}\,\theta_0 + \hat{B}^*_{0l}|\theta_0)\right] = \mathrm{var}(\hat{A}^*_{0l}) + \mathrm{var}(\hat{B}^*_{0l}), \tag{35}$$

assuming that θ_0 has zero mean and variance equal to one.

3. A Simulation Study

The performance of the methods proposed was assessed by means of a simulation study. Six administrations per year for nine years are considered, resulting in a total of 54 administrations. In order to simulate seasonality and a slight trend in mean ability levels, the mean ability is determined by the following equation

$$0.05\cos(2\pi\, y) - 0.2\sin(2\pi\, y) + 0.002\, y, \tag{36}$$

where y denotes the year. The function used to simulate seasonality was proposed in Lee and Haberman (2013) for modeling mean scores. Figure 1 represents mean ability levels over time. The points represent the test forms, which are administered in months 3, 5, 6, 10, 11 and 12 of each year. Dotted lines represent the links between forms that share common items. Each form is linked to two old forms, one administered one year prior in the same month of the year, and the other administered two years prior in a different month of the year.

[Figure 1 about here.]

Each form is composed of 40 items and the number of common items between two forms is 5. For every form, 2000 abilities have been generated independently from a normal distribution. The mean of the distribution is given by Equation (36), while the standard deviation was generated from a uniform distribution with range $[0.9, 1.2]$. Item responses were simulated using the two-parameter logistic model. Item difficulties are generated from a standard normal distribution, while discrimination parameters were generated from a normal distribution with mean 0.9 and standard deviation 0.3, truncated at 0.3 and 1.8. Results are based on 500 simulated data sets. All computations were performed using the `R` statistical software (R Development Core Team, 2015). Item parameters were estimated using the `ltm` function of the `ltm` package (Rizopoulos, 2006) with 41 quadrature points. The `ltm` package estimates item parameters by means of the marginal maximum likelihood method, hence assuming a standard normal distribution for the abilities. The code developed to implement the methods proposed in this paper was partly written in C language to speed up computational time. The minimization of equations (23) and (30) was performed using the `R` function `nlminb`. All forms were equated to Form 1 using all methods presented in this paper. On a PC with Intel Core i5-3210M at 2.50 GHz the MM-GM and the M-MM methods take just a few seconds to compute the equating coefficients for one data set. Instead, the MIRF and the MTRF take about 2 minutes for the computation of the equating coefficients. The computation of standard errors requires a bit more time. Approximatively, it takes 2 minutes for the MM-GM method, 9 minutes for the M-MM method and 2 minutes for the MIRF and the MTRF methods. These times can be reduced by improving the efficiency of the code and making use of parallel computation. Bisector and weighted bisector equating coefficients (Battauz, 2013) were also calculated using the `equateIRT` package (Battauz, 2015). Chain equating coefficients, which are used in the computation of bisector equating coefficients, were calculated for all possible chains with length from 3 to 9. In order to limit the number of chains constructed, only links from newer forms to older forms were used in the computation of chain equating coefficients. For each group of chain equating coefficients linking the same couple of forms, bisector equating coefficients were calculated using the mean-mean, mean-geometric mean, Haebara and Stocking-Lord methods for direct equating coefficients.

In order to evaluate the properties of the methods proposed, the mean and the standard deviation of the estimates of the equating coefficients were calculated for each form. Table 1 reports the absolute value of the difference between the mean estimates and the true values of the equating coefficients. Since there is a value for each of the 54 forms, in the table only the mean and the maximum values are reported. The table shows that the differences are very small for all the multiple equating methods, thus indicating that these methods are nearly unbiased.

[Table 1 about here.]

Table 2 reports mean and maximum values (across different forms) of the absolute value of the difference between mean standard errors and standard deviations of the equating coefficients. The small values shown in the table indicate that the calculated standard errors are nearly unbiased. The standard deviations of the standard errors calculated for each simulated dataset are instead summarized in Table 3. The table reports for each method minimum, mean and maximum values of the standard deviations of the standard errors across different forms and shows that the standard errors exhibit little variability.

[Table 2 about here.]

[Table 3 about here.]

The multiple equating methods yield similar equating coefficients (see Figure 2). In particular, the MM-GM and the MM-M methods have the smallest differences, especially for the B equating coefficients. Also the MIRF and the MTRF methods tend to produce very similar results for the B equating coefficient.

[Figure 2 about here.]

A comparison of the standard deviations of the estimated equating coefficients obtained with the various methods is given in Figure 3. The figure shows that the standard deviations of the A equating coefficient are similar between the various methods, although the MM/M method presents slightly smaller values than the other methods, while the TRF method produces standard deviations slightly higher than the other methods. Instead, the MIRF and the MTRF methods produce lower standard deviations for the B equating coefficient, compared to the MM-GM and the MM-M methods. In particular, the MIRF method yields standard deviation slightly smaller that the MTRF method.

[Figure 3 about here.]

The performance of the multiple equating methods proposed in this paper was then compared with the bisector and the weighted bisector methods. The results obtained with the bisector and the weighted bisector methods are rather similar, so only the weighted bisector method in shown in figures. Figure 4 compares the estimates of the A equating coefficient obtained with the various multiple equating methods with the estimates obtained with the weighted bisector method and shows an high similarity between the two methods. Results for the B equating coefficient are very similar and they are not shown.

[Figure 4 about here.]

A comparison of the standard deviations of the estimates of the A equating coefficient obtained with the methods presented in this paper with the weighted bisector method is given in Figure 5. The standard deviations of the weighted bisector method are equal or slightly greater than the multiple equating methods. The difference can be due to the fact that in the computation of the bisector equating coefficients only links from newer forms to older forms have been considered, thus not exploiting all the links present in the network of forms. As expected, the standard deviations of the bisector method (not shown here) are slightly larger than the weighted bisector method. The B equating coefficient presents very similar results.

[Figure 5 about here.]

In order to explore the effect of violations of the assumption of invariance of item parameters across different administrations, a data set with perturbed item parameters was also generated. Only the item parameters of Form 54, the last one, were perturbed. Form 54 shares same items with Forms 37 and 48. The difficulty item parameters in common with these forms were modified by adding values generated from a normal distribution with zero mean and standard deviation equal to 0.3. Instead, the discrimination item parameters were modified by adding values generated from a normal distribution with zero mean and standard deviation equal to 0.2. The values thus obtained were then truncated at 0.3 and 1.8. Only the MM-M method was used to estimate the equating coefficients and the quantities $\hat{\sigma}^2_{boot}(A_t)$ and $\hat{\sigma}^2_{boot}(B_t)$ as explained in Subsection 2.5 have then been calculated with $M = 300$ and $B = 300$. Despite in real applications disturbances are likely to involve numerous items, here only the items of one form were perturbed in order to observe the behavior of $\hat{\sigma}^2_{boot}(A_t)$ and $\hat{\sigma}^2_{boot}(B_t)$ for forms with non-perturbed item parameters. Similarly to Michaelides and Haertel (2014), item parameters were estimated after resampling of examinees, while resampling of common items did not required the estimation the IRT model. Figure 6 represents the estimated variability of the equating coefficients and shows that $\hat{\sigma}^2_{boot}(A_t)$ and $\hat{\sigma}^2_{boot}(B_t)$ are all near zero excepted Form 54. For this form, $\hat{\sigma}^2_{boot}(A_t) = 0.010$ and $\hat{\sigma}^2_{boot}(B_t) = 0.014$. Since the squared standard errors of the equating coefficients were 0.006 for the A equating coefficient and 0.011 for the B equating coefficient, the variability of the equating coefficients due to the choice of common items is not negligible for this form.

[Figure 6 about here.]

4. Discussion and Conclusions

This paper proposes a generalization to the case of a network of forms of the methods proposed in the literature to equate two test forms. Specifically, the methods considered are the mean-geometric mean, the mean-mean, the Haebara and the

Stocking-Lord methods. The mean-sigma method (Marco, 1977) was instead not considered in this paper because this method produces biased estimators of the equating coefficients (Baldwin, 2013) and simulation studies not presented here showed that a generalization of the method to multiple forms leads to non negligible bias.

This work was inspired by the illuminating paper of Haberman (2009), who proposed a generalization of the mean-geometric mean method to a large number of test forms by formalizing the problem as a regression model. A contribution of the present paper is the derivation of the asymptotic standard errors of the equating coefficients obtained with the procedure described by Haberman (2009), along with the standard errors of the equating coefficients obtained with the other methods presented in this paper. Standard errors of the equating coefficients are an important tool for the assessment of the accuracy of the equating process. The derivation of analytic standard errors of the equating coefficients has received attention in the literature (see Ogasawara 2000, 2001b, for direct equating coefficients and Battauz, 2013, for chain and average equating coefficients). Determining the asymptotic covariance matrix of the equating coefficients is also important because this matrix is necessary to obtain the standard errors of the equated scores as in Ogasawara (2001a, 2003). These standard errors are obtained under the assumption of invariance of item parameters. As remarked in the paper, when this assumption is not satisfied, the selection of the set of common items constitutes a further source of variability of the equating coefficients. This paper provides also an adaptation to the case of multiple forms of the procedures described in Haberman et al. (2009) and Michaelides and Haertel (2014) to estimate this variability.

The impact of sample error of the equating coefficients can be evaluated by considering the change in the reliability index of the estimated abilities after the scale transformation (as explained in Appendix D). The values relative to Form 3 of one of the simulated data sets have been taken as an example. Suppose that the estimated standard error of $\hat{\theta}_3$ for one person is 0.33 (which is the median of the estimated standard errors of Form 3). So, the reliability of $\hat{\theta}_3$ is equal to $1/(1 + 0.33^2) = 0.9$. Since $\hat{A}_3 = 1.11$, $\hat{se}(\hat{A}_3) = 0.078$ and $\hat{se}(\hat{B}_3) = 0.097$, the reliability of $\hat{\theta}^*$ is $1.11^2/(0.078^2 + 0.097^2 + (1 + 0.33^2)1.11^2) = 0.89$. The estimated standard error of $\hat{\theta}^*$ is $(0.078^2 + 0.097^2 + 0.33^2 1.11^2)^{1/2} = 0.39$. So, the reliability of the measure of ability of this person is just slightly reduced after the conversion to the base scale, indicating that the equating process is rather accurate.

The MIRF and MTRF methods are based on the minimization of a loss function given in Equations (17) and (29) that depends on the difference between the response function evaluated using the item parameters estimated separately for each form and the synthetic item parameters converted on the scale of that form. Alternatively, a possible choice would have been to convert all item parameters to the scale of the base form and leave unchanged the synthetic item parameters. However, this approach gives different results than the one adopted in this paper. In particular, the approach chosen in this paper assures the symmetry property, which implies that the results are independent of the choice of the base form. So, the equating coefficients for a different base form can be derived just by applying a linear transformation to those obtained with the MIRF or the MTRF methods. Instead, following the alternative approach, this property is not satisfied.

While the MM-GM and the MM-M methods with two forms produce the same results as the classical mean-geometric mean and mean-mean methods, the MIRF and MTRF methods with two forms do not correspond to the Haebara and Stocking-Lord methods. However, a simulation study, not reported here, showed that the estimates of the equating coefficients obtained with the MIRF and MTRF methods are extremely similar to those obtained with the Haebara and Stocking-Lord methods. The mean and the standard deviation of the equating coefficients were also very close. Then, the characteristic curve methods proposed here to equate multiple test forms constitute a new method to equate two test forms, which also returns the synthetic item parameters as a byproduct.

Potentially, the methods proposed in this paper do not have limits on the number of forms equated. A restriction is given by the amount of memory available on the computer, but the code implementation has an important role in this respect. The time required for the computation of the equating coefficients and the standard errors is expected to increase when the total number of common items between all the forms becomes larger. Instead, the number of examinees has an effect only on the estimation of the IRT model and it is not relevant to the time required for the estimation of the equating coefficients.

Though it should be remarked that simulated data are not real data and that simulation studies have limits in terms of generalizability, the simulation study conducted here showed the good performance of all the methods proposed to equate simultaneously multiple test forms and showed also that the methods give similar results. The simulation study provided also a comparison with the bisector method (Battauz, 2013), which was also developed to address equating of multiple test forms. This study revealed that the methods presented here give similar results to the bisector and weighted bisector methods, both considering the estimated values and the standard deviations of the equating coefficients. The standard deviations were slightly higher for the weighted bisector method under the settings chosen here. However, this result can change when the dimension of the groups of examinees varies across different administrations, since the multiple equating methods proposed in this paper do not account for the dimension of the samples or the standard error of the item parameter estimates. Instead, the weighted bisector method is constructed in order to attain an efficient estimator. The two approaches present both advantages and drawbacks. Linking simultaneously all the forms is certainly more straightforward and seems preferable when the network of connections between forms is very intricate. On the other hand, the bisector method can deal more easily with the case of a new form that needs to be added to a network of forms previously equated. In fact, with the bisector method it is not necessary to compute again all the equating coefficients, but only those involved by this new form. Furthermore, the bisector method performs well also using the mean-sigma method. Large differences in the equating coefficients obtained with the two approaches can reveal problems with the equating process. So, computing the equating coefficients with both the approaches and comparing the results could be a convenient strategy.

Appendix A. Partial Derivatives necessary to obtain the asymptotic standard errors

A.1. MM-GM Method

The derivatives in equation (10), necessary to compute the covariance matrix of the equating coefficients and the synthetic item parameters with the MM-GM method, are given by

$$\frac{\partial \hat{\boldsymbol{\beta}}_1}{\partial \hat{\mathbf{a}}^\top} = \mathrm{diag}(\hat{\boldsymbol{\beta}}_1)(\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top \left(\mathrm{diag}(\hat{\mathbf{a}})\right)^{-1}, \tag{A1}$$

$$\frac{\partial \hat{\boldsymbol{\beta}}_1}{\partial \hat{\mathbf{b}}^\top} = \mathbf{0}, \tag{A2}$$

$$\frac{\partial \hat{\boldsymbol{\beta}}_2}{\partial \hat{\mathbf{a}}^\top} = \frac{\partial \hat{\boldsymbol{\beta}}_2}{\partial \hat{\mathbf{A}}^\top}\frac{\partial \hat{\mathbf{A}}}{\partial \hat{\mathbf{a}}^\top} = (\mathbf{X}_2^\top \mathbf{X}_2)^{-1}\mathbf{X}_2^\top \mathrm{diag}(\hat{\mathbf{b}})\mathbf{T}^\top \frac{\partial \hat{\mathbf{A}}}{\partial \hat{\mathbf{a}}^\top}, \tag{A3}$$

$$\frac{\partial \hat{\boldsymbol{\beta}}_2}{\partial \hat{\mathbf{b}}^\top} = (\mathbf{X}_2^\top \mathbf{X}_2)^{-1}\mathbf{X}_2^\top \mathrm{diag}(\hat{\mathbf{A}}_n), \tag{A4}$$

where $\frac{\partial \hat{\mathbf{A}}}{\partial \hat{\mathbf{a}}^\top}$ is given in the first $T-1$ rows of $\frac{\partial \hat{\beta}_1}{\partial \hat{\mathbf{a}}^\top}$.

A.2. MM-M Method

The derivatives of the equating coefficients A_t, $t = 2, \dots, T$, with respect to the estimated item discrimination parameters $\hat{a}_{js}$, $j = 1, \dots, v$, $s = 1, \dots, T$, namely $\frac{\partial \hat{A}_t}{\partial \hat{a}_{js}}$, can not be found in closed form, but can instead be determined numerically. These derivatives compose the matrix $\frac{\partial \hat{\mathbf{A}}}{\partial \hat{\mathbf{a}}^\top}$, that corresponds to the first $T-1$ rows of $\frac{\partial \hat{\beta}_1}{\partial \hat{\mathbf{a}}^\top}$. The derivatives of the synthetic discrimination parameters $\hat{a}_j^*$ with respect to the discrimination parameter estimates obtained from each calibration can be then found as follows:

$$\frac{\partial \hat{a}_j^*}{\partial \hat{a}_{jt}} = \frac{1}{\sum_{s \in U_j} \hat{A}_s} - \frac{\sum_{s \in U_j} \hat{a}_{js}}{\left(\sum_{s \in U_j} \hat{A}_s\right)^2} \sum_{s \in U_j} \frac{\partial \hat{A}_s}{\partial \hat{a}_{jt}}, \tag{A5}$$

$$\frac{\partial \hat{a}_j^*}{\partial \hat{a}_{it}} = -\frac{\sum_{s \in U_j} \hat{a}_{js}}{\left(\sum_{s \in U_j} \hat{A}_s\right)^2} \sum_{s \in U_j} \frac{\partial \hat{A}_s}{\partial \hat{a}_{it}}, \quad \forall i \neq j. \tag{A6}$$

These derivatives form the matrix $\frac{\partial \hat{\mathbf{a}}^*}{\partial \hat{\mathbf{a}}^\top}$, that corresponds to the last v rows of $\frac{\partial \hat{\beta}_1}{\partial \hat{\mathbf{a}}^\top}$. The derivatives $\frac{\partial \hat{\beta}_1}{\partial \hat{\mathbf{b}}^\top}$, $\frac{\partial \hat{\beta}_2}{\partial \hat{\mathbf{a}}^\top}$ and $\frac{\partial \hat{\beta}_2}{\partial \hat{\mathbf{b}}^\top}$ can then be determined as explained in Appendix A.1 for the MM-GM method, using the appropriate matrices $\frac{\partial \hat{\mathbf{A}}}{\partial \hat{\mathbf{a}}^\top}$ and $\hat{\mathbf{A}}_n$.

A.3. MIRF and MTRF Methods

In order to obtain the partial derivatives necessary to compute the asymptotic standard errors of the equating coefficients, P^*_{mjt} will be written as follows:

$$P^*_{mjt} = \hat{c}_{jt} + (1 - \hat{c}_{jt}) \frac{\exp(LP_{mjt})}{1 + \exp(LP_{mjt})}, \tag{A7}$$

where

$$LP_{mjt} = D \frac{1}{u_j} \left(\sum_{\substack{s \in U_j \\ s \neq t}} \frac{\hat{a}_{js}}{\hat{A}_s} \hat{A}_t + \hat{a}_{jt} \right) y_m - D \frac{1}{u_j} \sum_{s \in U_j} \frac{\hat{a}_{js}}{\hat{A}_s} \left(\frac{1}{u_j} \sum_{s \in U_j} (\hat{b}_{js} \hat{A}_s + \hat{B}_s) - \hat{B}_t \right). \tag{A8}$$

In the following, all the derivatives entering in Equations (26), (27) (28), (32), (33) and (34) will be given.

$$\frac{\partial P^*_{mjt}}{\partial \hat{A}_t} = \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \cdot \frac{\partial LP_{mjt}}{\partial \hat{A}_t}, \tag{A9}$$

where

$$\frac{\partial P^*_{mjt}}{\partial LP_{mjt}} = (P^*_{mjt} - \hat{c}_{jt}) \left(1 - \frac{P^*_{mjt} - \hat{c}_{jt}}{1 - \hat{c}_{jt}} \right), \tag{A10}$$

$$\frac{\partial LP_{mjt}}{\partial \hat{A}_t} = \frac{D}{u_j}\left[\sum_{\substack{s\in U_j\\ s\neq t}} \frac{\hat{a}_{js}}{\hat{A}_s} y_m + \frac{\hat{a}_{jt}}{\hat{A}_t^2}(\hat{b}_j^* - \hat{B}_t) I_{U_j}(t) - \hat{a}_j^* \hat{b}_{jt} I_{U_j}(t)\right], \tag{A11}$$

and $I_{U_j}(t)$ is an indicator function, which is 1 if $t \in U_j$.

$$\frac{\partial P^*_{mjt}}{\partial \hat{A}_k} = \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \cdot \frac{\partial LP_{mjt}}{\partial \hat{A}_k}, \quad \forall k \neq t, \tag{A12}$$

where

$$\frac{\partial LP_{mjt}}{\partial \hat{A}_k} = \frac{D}{u_j}\left[-\frac{\hat{a}_{jk}}{\hat{A}_k^2}\hat{A}_t y_m + \frac{\hat{a}_{jk}}{\hat{A}_k^2}(\hat{b}_j^* - \hat{B}_t) - \hat{a}_j^* \hat{b}_{jk}\right] I_{U_j}(k); \tag{A13}$$

$$\frac{\partial P^*_{mjt}}{\partial \hat{B}_t} = \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \cdot \frac{\partial LP_{mjt}}{\partial \hat{B}_t}, \tag{A14}$$

where

$$\frac{\partial LP_{mjt}}{\partial \hat{B}_t} = D\hat{a}_j^* \left(1 - \frac{1}{u_j} I_{U_j}(t)\right); \tag{A15}$$

$$\frac{\partial P^*_{mjt}}{\partial \hat{B}_k} = \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \cdot \frac{\partial LP_{mjt}}{\partial \hat{B}_k}, \quad \forall k \neq t, \tag{A16}$$

where

$$\frac{\partial LP_{mjt}}{\partial \hat{B}_k} = -D\hat{a}_j^* \frac{1}{u_j} I_{U_j}(k); \tag{A17}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{B}_t \partial \hat{B}_k} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{B}_k} \frac{\partial LP_{mjt}}{\partial \hat{B}_t}, \quad \forall k, \tag{A18}$$

where

$$\frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{B}_k} = \frac{\partial P^*_{mjt}}{\partial \hat{B}_k} - 2\frac{P^*_{mjt} - \hat{c}_{jt}}{1 - \hat{c}_{jt}} \frac{\partial P^*_{mjt}}{\partial \hat{B}_k}. \tag{A19}$$

All other second derivatives of P^*_{mjt} with respect to LP_{mjt} and one of these variables $\hat{B}_h, \hat{B}_t, \hat{A}_k, \hat{A}_h, \hat{A}_t, \hat{a}_{jk}, \hat{a}_{jh}, \hat{a}_{jt}, \hat{b}_{jk}, \hat{b}_{jh}, \hat{b}_{jt}$ are analogous, and can be obtained by substituting $\hat{B}_k$ with the appropriate variable in (A19). The other derivatives entering in Equations (27), (28), (33) and (34) are

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{B}_t \partial \hat{A}_k} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{A}_k} \frac{\partial LP_{mjt}}{\partial \hat{B}_t} - \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \frac{\hat{a}_{jk}}{\hat{A}_k^2} \left(1 - \frac{1}{u_j} I_{U_j}(t)\right) I_{U_j}(k), \quad \forall k, \tag{A20}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{B}_k \partial \hat{B}_h} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{B}_h} \frac{\partial LP_{mjt}}{\partial \hat{B}_k}, \quad \forall k \neq t,\ \forall h, \tag{A21}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{B}_k \partial \hat{A}_h} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{A}_h} \frac{\partial LP_{mjt}}{\partial \hat{B}_k} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j^2} \frac{\hat{a}_{jh}}{\hat{A}_h^2} I_{U_j}(k) I_{U_j}(h), \quad \forall k \neq t,\ \forall h, \tag{A22}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_t^2} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{A}_t} \frac{\partial LP_{mjt}}{\partial \hat{A}_t} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \left[-2 \frac{\hat{a}_{jt}}{\hat{A}_t^3} (\hat{b}_j^* - \hat{B}_t) + 2 \frac{\hat{a}_{jt} \hat{b}_{jt}}{u_j \hat{A}_t^2} \right] I_{U_j}(t), \tag{A23}$$

$$\begin{aligned} \frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_t \partial \hat{A}_k} = {} & \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{A}_k} \frac{\partial LP_{mjt}}{\partial \hat{A}_t} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \\ & \left[-\frac{\hat{a}_{jk}}{\hat{A}_k^2} y_m + \frac{\hat{a}_{jt} \hat{b}_{jk}}{u_j \hat{A}_t^2} I_{U_j}(t) + \frac{\hat{a}_{jk} \hat{b}_{jt}}{u_j \hat{A}_k^2} I_{U_j}(t) \right] I_{U_j}(k), \quad \forall k \neq t, \end{aligned} \tag{A24}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_t \partial \hat{B}_t} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{B}_t} \frac{\partial LP_{mjt}}{\partial \hat{A}_t} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \frac{\hat{a}_{jt}}{\hat{A}_t^2} \left(\frac{1}{u_j} - 1 \right) I_{U_j}(t), \tag{A25}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_t \partial \hat{B}_k} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{B}_k} \frac{\partial LP_{mjt}}{\partial \hat{A}_t} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j^2} \frac{\hat{a}_{jt}}{\hat{A}_t^2} I_{U_j}(t) I_{U_j}(k), \quad \forall k \neq t, \tag{A26}$$

$$\begin{aligned} \frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k^2} = {} & \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{A}_k} \frac{\partial LP_{mjt}}{\partial \hat{A}_k} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \\ & \left[2 \frac{\hat{a}_{jk}}{\hat{A}_k^3} \hat{A}_t y_m - 2 \frac{\hat{a}_{jk}}{\hat{A}_k^3} (\hat{b}_j^* - \hat{B}_t) + 2 \frac{\hat{a}_{jk} \hat{b}_{jk}}{u_j \hat{A}_k^2} \right] I_{U_j}(k), \quad \forall k \neq t, \end{aligned} \tag{A27}$$

$$\begin{aligned} \frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{A}_t} = {} & \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{A}_t} \frac{\partial LP_{mjt}}{\partial \hat{A}_k} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \\ & \left[-\frac{\hat{a}_{jk}}{\hat{A}_k^2} y_m + \frac{\hat{a}_{jk} \hat{b}_{jt}}{u_j \hat{A}_k^2} I_{U_j}(t) + \frac{\hat{a}_{jt} \hat{b}_{jk}}{u_j \hat{A}_t^2} I_{U_j}(t) \right] I_{U_j}(k), \quad \forall k \neq t, \end{aligned} \tag{A28}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{B}_h} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{B}_h} \frac{\partial LP_{mjt}}{\partial \hat{A}_k} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j^2} \frac{\hat{a}_{jk}}{\hat{A}_k^2} I_{U_j}(k) I_{U_j}(h), \quad \forall k,\ h \neq t, \tag{A29}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{B}_t} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{B}_t} \frac{\partial LP_{mjt}}{\partial \hat{A}_k} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \frac{\hat{a}_{jk}}{\hat{A}_k^2} \left(\frac{1}{u_j} I_{U_j}(t) - 1 \right) I_{U_j}(k), \quad \forall k \neq t, \tag{A30}$$

$$\begin{aligned} \frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{A}_h} = {} & \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{A}_h} \frac{\partial LP_{mjt}}{\partial \hat{A}_k} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \left[\frac{\hat{a}_{jk} \hat{b}_{jh}}{u_j \hat{A}_k^2} + \frac{\hat{a}_{jh} \hat{b}_{jk}}{u_j \hat{A}_h^2} \right] I_{U_j}(k) I_{U_j}(h), \\ & \forall k,\ h \neq t, \end{aligned} \tag{A31}$$

$$\frac{\partial P_{mjt}}{\partial \hat{c}_{jt}} = 1 - \frac{P_{mjt} - \hat{c}_{jt}}{1 - \hat{c}_{jt}}, \tag{A32}$$

$$\frac{\partial P_{mjt}}{\partial \hat{a}_{jt}} = (P_{mjt} - \hat{c}_{jt}) \left(1 - \frac{P_{mjt} - \hat{c}_{jt}}{1 - \hat{c}_{jt}}\right) D(y_m - \hat{b}_{jt}), \tag{A33}$$

$$\frac{\partial P_{mjt}}{\partial \hat{b}_{jt}} = -(P_{mjt} - \hat{c}_{jt}) \left(1 - \frac{P_{mjt} - \hat{c}_{jt}}{1 - \hat{c}_{jt}}\right) D\hat{a}_{jt}, \tag{A34}$$

$$\frac{\partial P^*_{mjt}}{\partial \hat{c}_{jt}} = 1 - \frac{P^*_{mjt} - \hat{c}_{jt}}{1 - \hat{c}_{jt}}, \tag{A35}$$

$$\frac{\partial P^*_{mjt}}{\partial \hat{a}_{jt}} = \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \left[y_m - \frac{1}{\hat{A}_t}(\hat{b}^*_j - \hat{B}_t) \right] I_{U_j}(t), \tag{A36}$$

$$\frac{\partial P^*_{mjt}}{\partial \hat{a}_{jk}} = \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \left[\frac{\hat{A}_t}{\hat{A}_k} y_m - \frac{1}{\hat{A}_k}(\hat{b}^*_j - \hat{B}_t) \right] I_{U_j}(k), \quad \forall k \neq t, \tag{A37}$$

$$\frac{\partial P^*_{mjt}}{\partial \hat{b}_{jk}} = -\frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \hat{a}^*_j \hat{A}_k I_{U_j}(k), \quad \forall k, \tag{A38}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{B}_t \partial \hat{c}_{jt}} = -\frac{1}{1 - \hat{c}_{jt}} \frac{\partial P^*_{mjt}}{\partial \hat{B}_t}, \tag{A39}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{B}_t \partial \hat{a}_{jk}} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{a}_{jk}} \frac{\partial LP_{mjt}}{\partial \hat{B}_t} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \frac{1}{\hat{A}_k} \left(1 - \frac{1}{u_j} I_{U_j}(t)\right) I_{U_j}(k), \quad \forall k, \tag{A40}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{B}_t \partial \hat{b}_{jk}} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{b}_{jk}} \frac{\partial LP_{mjt}}{\partial \hat{B}_t}, \quad \forall k, \tag{A41}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{B}_k \partial \hat{c}_{jt}} = -\frac{1}{1 - \hat{c}_{jt}} \frac{\partial P^*_{mjt}}{\partial \hat{B}_k}, \tag{A42}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{B}_k \partial \hat{a}_{jh}} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{a}_{jh}} \frac{\partial LP_{mjt}}{\partial \hat{B}_k} - \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} D \frac{1}{u_j^2} \frac{1}{\hat{A}_h} I_{U_j}(k) I_{U_j}(h), \quad \forall k \neq t, \, \forall h, \tag{A43}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{B}_k \partial \hat{b}_{jh}} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{b}_{jh}} \frac{\partial LP_{mjt}}{\partial \hat{B}_k}, \quad \forall k \neq t, \, \forall h, \tag{A44}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_t \partial \hat{c}_{jt}} = -\frac{1}{1 - \hat{c}_{jt}} \frac{\partial P^*_{mjt}}{\partial \hat{A}_t}, \tag{A45}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_t \partial \hat{a}_{jt}} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{a}_{jt}} \frac{\partial LP_{mjt}}{\partial \hat{A}_t} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \left[\frac{1}{\hat{A}_t^2}(\hat{b}^*_j - \hat{B}_t) - \frac{\hat{b}_{jt}}{u_j \hat{A}_t} \right] I_{U_j}(t), \tag{A46}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_t \partial \hat{a}_{jk}} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{a}_{jk}} \frac{\partial LP_{mjt}}{\partial \hat{A}_t} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \left[\frac{1}{\hat{A}_k} y_m - \frac{\hat{b}_{jt}}{u_j \hat{A}_k} I_{U_j}(t) \right] I_{U_j}(k), \quad \forall k \neq t, \tag{A47}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_t \partial \hat{b}_{jt}} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{b}_{jt}} \frac{\partial LP_{mjt}}{\partial \hat{A}_t} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \left[\frac{\hat{a}_{jt}}{u_j \hat{A}_t} - \hat{a}^*_j \right] I_{U_j}(t), \tag{A48}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_t \partial \hat{b}_{jk}} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{b}_{jk}} \frac{\partial LP_{mjt}}{\partial \hat{A}_t} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \left[\frac{\hat{a}_{jt} \hat{A}_k}{u_j \hat{A}_t^2} \right] I_{U_j}(t) I_{U_j}(k), \quad \forall k \neq t, \tag{A49}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{c}_{jt}} = -\frac{1}{1 - \hat{c}_{jt}} \frac{\partial P^*_{mjt}}{\partial \hat{A}_k}, \quad \forall k \neq t. \tag{A50}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{a}_{jk}} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{a}_{jk}} \frac{\partial LP_{mjt}}{\partial \hat{A}_k} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \left[-\frac{\hat{A}_t}{\hat{A}_k^2} y_m + \frac{1}{\hat{A}_k^2} (\hat{b}^*_j - \hat{B}_t) - \frac{\hat{b}_{jk}}{u_j \hat{A}_k} \right] I_{U_j}(k), \quad \forall k \neq t, \tag{A51}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{a}_{jh}} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{a}_{jh}} \frac{\partial LP_{mjt}}{\partial \hat{A}_k} - \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j^2} \frac{\hat{b}_{jk}}{\hat{A}_h} I_{U_j}(k) I_{U_j}(h), \quad \forall k \neq t, \, \forall h \neq k, \tag{A52}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{b}_{jk}} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{b}_{jk}} \frac{\partial LP_{mjt}}{\partial \hat{A}_k} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j} \left(\frac{\hat{a}_{jk}}{u_j \hat{A}_k} - \hat{a}^*_j \right) I_{U_j}(k), \quad \forall k \neq t, \tag{A53}$$

$$\frac{\partial^2 P^*_{mjt}}{\partial \hat{A}_k \partial \hat{b}_{jh}} = \frac{\partial^2 P^*_{mjt}}{\partial LP_{mjt} \partial \hat{b}_{jh}} \frac{\partial LP_{mjt}}{\partial \hat{A}_k} + \frac{\partial P^*_{mjt}}{\partial LP_{mjt}} \frac{D}{u_j^2} \frac{\hat{a}_{jk} \hat{A}_h}{\hat{A}_k^2} I_{U_j}(k) I_{U_j}(h), \quad \forall k \neq t, \, \forall h \neq k, \tag{A54}$$

The derivatives of the synthetic discrimination parameters with respect to the item parameter obtained from separate calibration can be found as follows:

$$\frac{\partial \hat{a}^*_j}{\partial \hat{a}_{jt}} = \frac{1}{u_j} \left(\frac{1}{\hat{A}_t} I_{U_j}(t) - \sum_{s \in U_j} \frac{\hat{a}_{js}}{\hat{A}_s^2} \frac{\partial \hat{A}_s}{\partial \hat{a}_{jt}} \right), \tag{A55}$$

$$\frac{\partial \hat{a}^*_j}{\partial \hat{a}_{it}} = \frac{1}{u_j} \left(- \sum_{s \in U_j} \frac{\hat{a}_{js}}{\hat{A}_s^2} \frac{\partial \hat{A}_s}{\partial \hat{a}_{it}} \right), \quad \forall i \neq j, \tag{A56}$$

$$\frac{\partial \hat{a}^*_j}{\partial \hat{b}_{it}} = \frac{1}{u_j} \left(- \sum_{s \in U_j} \frac{\hat{a}_{js}}{\hat{A}_s^2} \frac{\partial \hat{A}_s}{\partial \hat{b}_{it}} \right), \quad \forall i. \tag{A57}$$

$$\frac{\partial \hat{b}^*_j}{\partial \hat{a}_{it}} = \frac{1}{u_j} \sum_{s \in U_j} \left(\hat{b}_{jt} \frac{\partial \hat{A}_s}{\partial \hat{a}_{it}} + \frac{\partial \hat{B}_s}{\partial \hat{a}_{it}} \right), \quad \forall i. \tag{A58}$$

$$\frac{\partial \hat{b}_j^*}{\partial \hat{b}_{jt}} = \frac{1}{u_j}\left[\hat{A}_t I_{U_j}(t) + \sum_{s\in U_j}\left(\hat{b}_{js}\frac{\partial \hat{A}_s}{\partial \hat{b}_{jt}} + \frac{\partial \hat{B}_s}{\partial \hat{b}_{jt}}\right)\right]. \tag{A59}$$

$$\frac{\partial \hat{b}_j^*}{\partial \hat{b}_{it}} = \frac{1}{u_j}\sum_{s\in U_j}\left(\hat{b}_{js}\frac{\partial \hat{A}_s}{\partial \hat{b}_{it}} + \frac{\partial \hat{B}_s}{\partial \hat{b}_{it}}\right), \quad \forall i \neq j. \tag{A60}$$

Appendix B. Proof of the Correspondence Between the MM-M Method for Two Forms and the Mean-Mean Method.

When $T = 2$ the estimator of the equating coefficient A_2 with the MM-M method is given by

$$\hat{A}_2 = \frac{\sum_{j\in J_2}\hat{a}_{j2}}{\sum_{j\in J_2}\hat{a}_j^*} = \frac{\sum_{j\in J_2}\hat{a}_{j2}}{\sum_{j\in J_1\cap J_2}\frac{\hat{a}_{j1}+\hat{a}_{j2}}{1+\hat{A}_2} + \sum_{j\in J_2\setminus J_1}\frac{\hat{a}_{j2}}{\hat{A}_2}}, \tag{A61}$$

from which we obtain

$$\hat{A}_2\sum_{j\in J_1\cap J_2}\frac{\hat{a}_{j1}+\hat{a}_{j2}}{1+\hat{A}_2} + \hat{A}_2\sum_{j\in J_2\setminus J_1}\frac{\hat{a}_{j2}}{\hat{A}_2} = \sum_{j\in J_1\cap J_2}\hat{a}_{j2} + \sum_{j\in J_2\setminus J_1}\hat{a}_{j2}, \tag{A62}$$

and

$$\frac{\hat{A}_2}{1+\hat{A}_2}\sum_{j\in J_1\cap J_2}\hat{a}_{j1}+\hat{a}_{j2} = \sum_{j\in J_1\cap J_2}\hat{a}_{j2}. \tag{A63}$$

We then obtain

$$\hat{A}_2\sum_{j\in J_1\cap J_2}\hat{a}_{j1} = \sum_{j\in J_1\cap J_2}\hat{a}_{j2}. \tag{A64}$$

The estimator of the equating coefficient A_2 is then equal to

$$\hat{A}_2 = \frac{\sum_{j\in J_1\cap J_2}\hat{a}_{j2}}{\sum_{j\in J_1\cap J_2}\hat{a}_{j1}}, \tag{A65}$$

that corresponds to the mean-mean estimator of the equating coefficient A for two forms.

Appendix C. Proof of the Symmetry Property of MIRF and MTRF methods.

In order to convert item parameters on the scale of Form r, the equating coefficients are transformed as follows:

$$\hat{A}'_t = \frac{\hat{A}_t}{\hat{A}_r} \quad \text{and} \quad \hat{B}'_t = \frac{\hat{B}_t - \hat{B}_r}{\hat{A}_r}, \quad \text{for } t = 1, \ldots, T,$$

so that $\hat{A}'_r = 1$ and $\hat{B}'_r = 0$. If $\hat{A}_t$ is replaced with $\hat{A}'_t$ and $\hat{B}_t$ is replaced with $\hat{B}'_t$ in Equation (22), it is simple to verify that $\hat{a}^*_{jt}$ and $\hat{b}^*_{jt}$ do not vary after this substitution. Consequently, Equations (17) and (29) are invariant with respect to changes of the base form, thus proving the symmetry property.

Appendix D. Variability of Estimated Abilities.

The following equation gives the conversion of estimated abilities from the scale of Form t to the scale of the base form

$$\theta^* = \theta_t A_t + B_t.$$

The estimated ability $\hat{\theta}_t$ can be transformed using the estimated equating coefficients

$$\hat{\theta}^* = \hat{\theta}_t \hat{A}_t + \hat{B}_t.$$

The variance of $\hat{\theta}^*$ given $\hat{\theta}_t$ is

$$\mathrm{var}(\hat{\theta}^* | \hat{\theta}_t) = \hat{\theta}_t^2 \mathrm{var}(\hat{A}_t) + \mathrm{var}(\hat{B}_t) + 2\hat{\theta}_t \mathrm{cov}(\hat{A}_t, \hat{B}_t),$$

while the conditional expected value is

$$\mathrm{E}(\hat{\theta}^* | \hat{\theta}_t) = \hat{\theta}_t \mathrm{E}(\hat{A}_t) + \mathrm{E}(\hat{B}_t) = \hat{\theta}_t (A_t + o(1)) + B_t + o(1),$$

provided that the estimators $\hat{A}_t$ and $\hat{B}_t$ are consistent. So, the variance of $\hat{\theta}^*$ is

$$\begin{aligned}\mathrm{var}(\hat{\theta}^*) &= \mathrm{E}\{\mathrm{var}(\hat{\theta}^*|\hat{\theta}_t)\} + \mathrm{var}\{\mathrm{E}(\hat{\theta}^*|\hat{\theta}_t)\} \\ &= \mathrm{var}(\hat{A}_t) + \mathrm{var}(\hat{B}_t) + \mathrm{var}(\hat{\theta}_t)A_t^2 + o(1),\end{aligned}$$

where $\mathrm{E}(\theta_t)$ and $\mathrm{var}(\theta_t)$ are assumed to be 0 and 1 respectively, as usual with the marginal maximum likelihood estimation method. Hence, if the reliability of $\hat{\theta}_t$ is

$$\rho(\hat{\theta}_t) = \frac{\mathrm{var}(\theta_t)}{\mathrm{var}(\hat{\theta}_t)} = \frac{1}{\mathrm{var}(\hat{\theta}_t)}, \tag{A66}$$

the reliability of $\hat{\theta}^*$ is

$$\rho(\hat{\theta}^*) = \frac{\mathrm{var}(\theta^*)}{\mathrm{var}(\hat{\theta}^*)} \simeq \frac{\mathrm{var}(\theta_t)A_t^2}{\mathrm{var}(\hat{A}_t) + \mathrm{var}(\hat{B}_t) + \mathrm{var}(\hat{\theta}_t)A_t^2} = \frac{A_t^2}{\mathrm{var}(\hat{A}_t) + \mathrm{var}(\hat{B}_t) + \mathrm{var}(\hat{\theta}_t)A_t^2}. \tag{A67}$$

The reliability of $\hat{\theta}^*$ is then always greater than the reliability of $\hat{\theta}_t$, due to variability of the estimated equating coefficients. These reliabilities can be estimated by substituting the true values with their estimates in (A66) and (A67). An estimate of $\mathrm{var}(\hat{\theta}_t)$ is $1 + \hat{se}^2(\hat{\theta}_t)$, where $\hat{se}(\hat{\theta}_t)$ is the estimated standard error of $\hat{\theta}_t$.

Another quantity of interest is the standard error of $\hat{\theta}^*$, which can be obtained as follows:

$$se(\hat{\theta}^*) = \{\mathrm{var}(\hat{\theta}^*) - \mathrm{var}(\theta^*)\}^{1/2} \simeq \{\mathrm{var}(\hat{A}_t) + \mathrm{var}(\hat{B}_t) + se^2(\hat{\theta}_t)A_t^2\}^{1/2}.$$

References

Baldwin, P. (2013). On mean-sigma estimators and bias. *British Journal of Mathematical and Statistical Psychology* , 66, 277–289. DOI: 10.1111/j.2044-8317.2012.02048.x

Battauz, M. (2013). IRT test equating in complex linkage plans. *Psychometrika*, 78, 464–480. DOI: 10.1007/s11336-012-9316-y.

Battauz, M. (2015). equateIRT: An R Package for IRT Test Equating. *Journal of Statistical Software*, 68, 1–22. DOI: 10.18637/jss.v068.i07.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459. DOI:10.1007/BF02293801.

Deming, W. E. & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11, 427–444. DOI:10.1214/aoms/1177731829.

Goodman, L. A. (1968). The analysis of cross-classified data: independence, quasi-independence and interactions in contingency tables with or without missing entries. *Journal of the American Statistical Association*, 63, 1091–1131. DOI:10.1080/01621459.1968.10480916.

Haberman, S. J. (2009). Linking parameter estimates derived from an item response model through separate calibrations. *ETS Research Report Series*, 2009, i-9. DOI: 10.1002/j.2333-8504.2009.tb02197.x.

Haberman, S. J., Lee, Y. H. & Qian, J. (2009). Jackknifing techniques for evaluation of equating accuracy . *ETS Research Report Series*, 2009, i-37. DOI: 10.1002/j.2333-8504.2009.tb02196.x.

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

Kim, S. & Kolen, M. J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics*, 32, 371–397. DOI: 10.3102/1076998607302632

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking: methods and practices* (3rd ed.). New York, NY: Springer-Verlag.

Lee, Y.-H. & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika*, 78, 815–829. DOI: 10.1007/S11336-013-9337-1.

Loyd B. H. & Hoover H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement* , 17, 179–193. DOI: 10.1111/j.1745-3984.1980.tb00825.x.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139–160. DOI: 10.1111/j.1745-3984.1977.tb00033.x.

Michaelides, M. P. & Haertel, E. H. (2014). Selection of common items as an unrecognized source of variability in test equating: a bootstrap approximation assuming random sampling of common items *Applied Measurement in Education*, 27, 46–57. DOI:10.1080/08957347.2013.853069.

Mislevy R.J. & Bock R. D. (1990). BILOG 3. Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

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

Ogasawara, H. (2001a). Item response theory true score equatings and their standard errors. *Journal of Educational and Behavioral Statistics*, 26, 31–50. DOI: 10.3102/10769986026001031.

Ogasawara, H. (2001b). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement*, 25, 53–67. DOI: 10.1177/01466216010251004.

Ogasawara, H. (2003). Asymptotic standard errors of IRT observed-score equating methods. *Psychometrika*, 68, 193–211. DOI: 10.1007/BF02294797.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17, 1–25. DOI: 10.18637/jss.v017.i05.

Stocking M. & Lord M. L. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210. DOI: 10.1177/014662168300700208.

van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory.* New York, NY: Springer-Verlag.

List of Figures

Figure 1.
Ability levels and linkage plan for the simulation study.

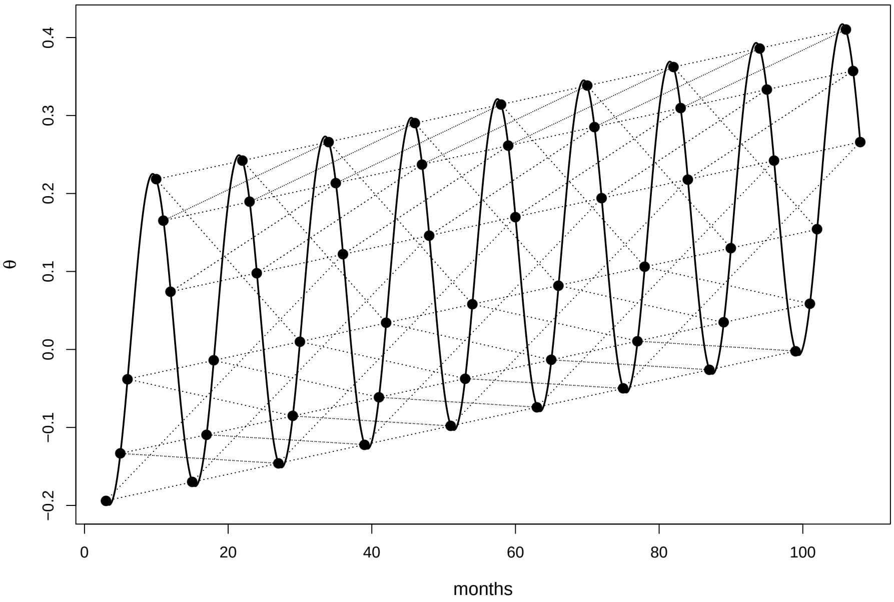
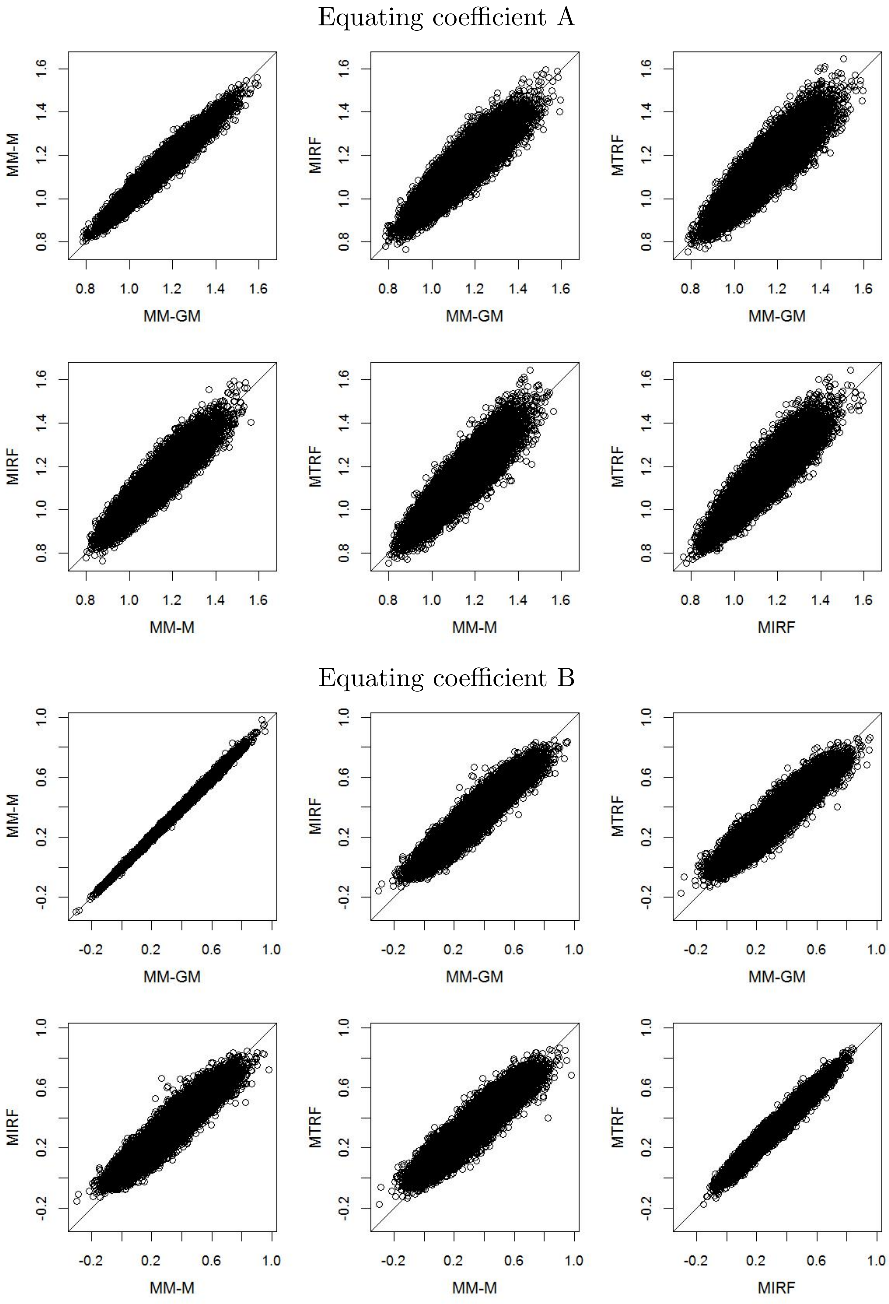


Figure 2.
Comparison of the estimates of the equating coefficients obtained with the various methods.

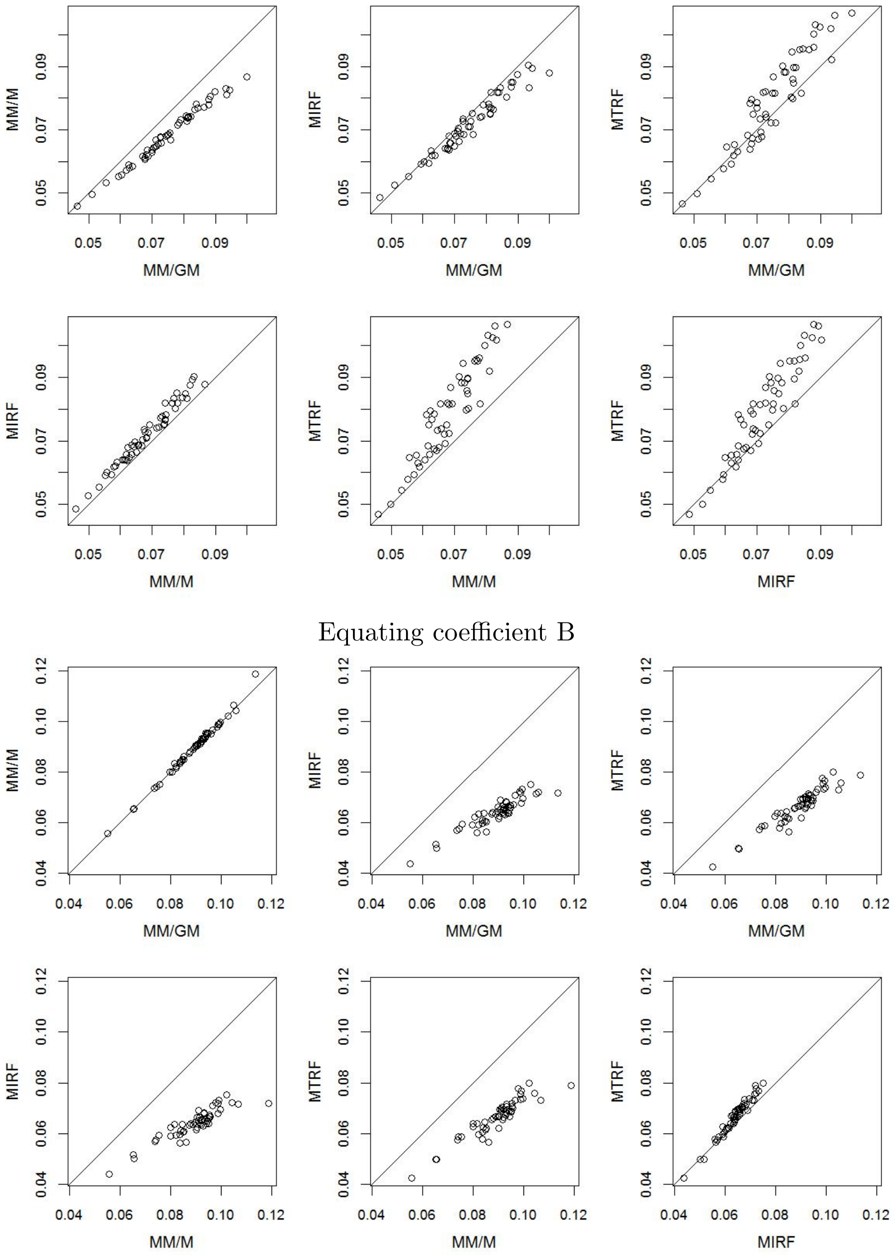

Equating coefficient A

Equating coefficient B

Figure 3.

Comparison of the standard deviations of the equating coefficients obtained with the various methods.

Equating coefficient A



Equating coefficient B

Figure 4.
Comparison of estimates of the equating coefficient A with the weighted bisector method.

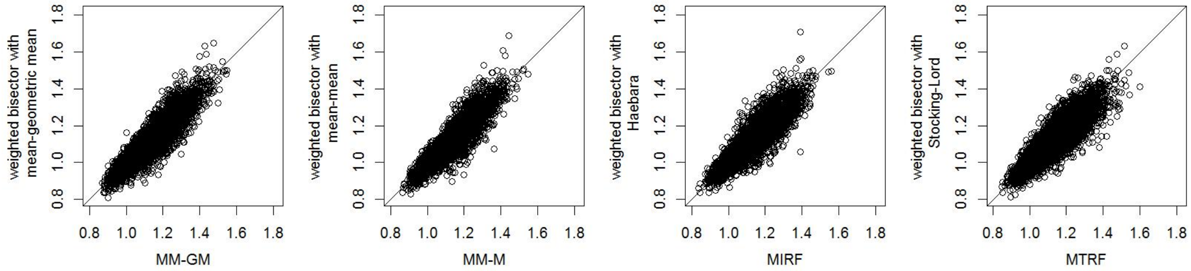


Figure 5.
Comparison of standard deviations of the equating coefficient A with the weighted bisector method.

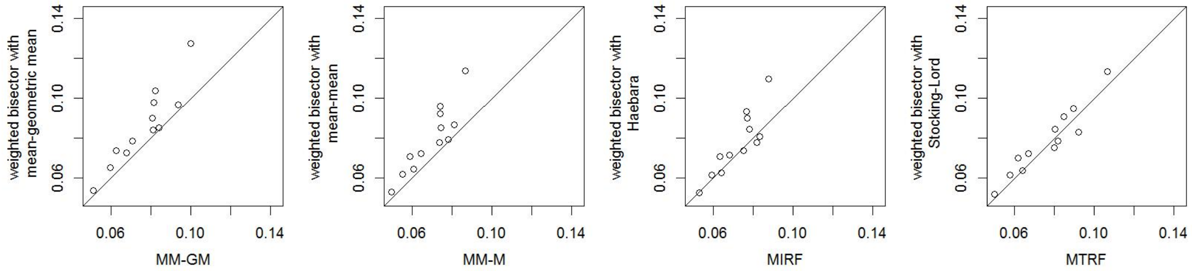
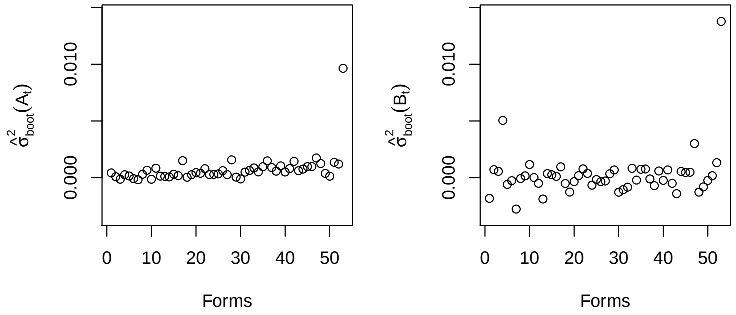


Figure 6.
Variability of equating coefficients due to the selection of common items.



List of Tables

Table 1.
Absolute difference between mean estimates and true values of the equating coefficients for the multiple equating methods.

coefficient	value	MM-GM	MM-M	MIRF	MTRF
A	mean	0.0045	0.0030	0.0018	0.0020
	max	0.0094	0.0068	0.0055	0.0072
B	mean	0.0026	0.0029	0.0016	0.0085
	max	0.0088	0.0095	0.0044	0.0196

Table 2.
Absolute difference between mean standard errors and standard deviations of the equating coefficients for the multiple equating methods.

coefficient	value	MM-GM	MM-M	MIRF	MTRF
A	mean	0.0024	0.0023	0.0014	0.0016
	max	0.0068	0.0056	0.0039	0.0043
B	mean	0.0020	0.0023	0.0013	0.0029
	max	0.0063	0.0070	0.0054	0.0089

Table 3.
Standard deviations of standard errors for the multiple equating methods.

coefficient	value	MM-GM	MM-M	MIRF	MTRF
A	min	0.0023	0.0021	0.0023	0.0022
	mean	0.0053	0.0045	0.0048	0.0060
	max	0.0080	0.0064	0.0067	0.0094
B	min	0.0022	0.0023	0.0012	0.0012
	mean	0.0046	0.0045	0.0026	0.0032
	max	0.0122	0.0151	0.0037	0.0059