Asymptotic Variance of Linking Coefficient Estimators for Polytomous IRT Models

Applied Psychological Measurement
2018, Vol. 42(3) 192–205
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621617721249
journals.sagepub.com/home/apm

Björn Andersson¹

Abstract

In item response theory (IRT), when two groups from different populations take two separate tests, there is a need to link the two ability scales so that the item parameters of the tests are comparable across the groups. To link the two scales, information from common items are utilized to estimate linking coefficients which place the item parameters on the same scale. For polytomous IRT models, the Haebara and Stocking–Lord methods for estimating the linking coefficients have commonly been recommended. However, estimates of the variance for these methods are not available in the literature. In this article, the asymptotic variance of linking coefficients for polytomous IRT models with the Haebara and Stocking–Lord methods are derived. The results are presented in a general form and specific results are given for the generalized partial credit model. Simulations which investigate the accuracy of the derivations under various settings of model complexity and sample size are provided, showing that the derivations are accurate under the conditions considered and that the Haebara and Stocking–Lord methods have superior performance to several moment methods with performance close to that of concurrent calibration.

Keywords

linking coefficients, equating coefficients, item response theory, standard errors, nonequivalent groups design

Item response theory (IRT) is a powerful framework for analyzing categorical data in educational and psychological testing. The results from IRT analyses are often used to report scores and for inferring population characteristics. When two groups from different populations take two tests, the item parameters for the respective groups are on different scales and cannot be directly compared. To enable the comparison between the parameters of the two tests, the parameters must be placed on a common scale (Kolen & Brennan, 2014). This is often accomplished by having some items which are common between the two tests.

One way to estimate the item parameters such that all parameters are expressed on the same scale is to use a multiple-group estimation procedure, also called concurrent calibration. With such a method, the differences between the ability distributions of the groups are taken into account and the item parameters are estimated with one of the groups as the reference group. A

¹Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, China

Corresponding Author:

Björn Andersson, Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, No. 19 Xijiekou Wai Street, Haidian District, Beijing 100875, China.
Email: bjoern.andersson@bnu.edu.cn

second way to obtain item parameters expressed on the same scale is to use separate estimation for the groups. Because the parameters of the common items should be identical for both groups if the parameters were on the same scale, the information contained in the estimated common items for each group can be used to estimate linking parameters A and B which place all the item parameters of the two tests on a common scale.

When using separate estimation for the item parameters from two groups, two types of estimators of A and B have mainly been used in the literature: methods based on moments and methods based on response functions. The moment methods use the means and variances of the estimated common item parameters to estimate A and B (Loyd & Hoover, 1980; Marco, 1977) while the response function methods minimize a distance measure between the item or test response functions for estimated common item parameters from two different groups (Haebara, 1980; Stocking & Lord, 1983). Several studies using dichotomous and polytomous IRT models have indicated that the response function methods have superior properties to the moment methods (Hanson & Béguin, 2002; S. H. Kim & Cohen, 1998; Kim & Kolen, 2007). It is therefore important to have estimates of the variance of the response function methods. However, in the literature the variance of linking coefficient estimators using response function methods has only been derived for dichotomous IRT models (Ogasawara, 2001).

The purpose of this article is to derive the asymptotic variance of response function estimators of linking coefficients when using polytomous IRT models. The results of the article are important because linking coefficients are utilized in several areas in educational and psychological measurement such as equating and linking of scales and tests. Having estimates of the variance of the response function methods for estimating linking coefficients for polytomous IRT models will add to the utility of the equating and linking methodology.

The article is structured as follows. First, polytomous IRT models are introduced briefly. Then linking coefficients are described and response function estimators of these are defined. The asymptotic covariance matrices of the Haebara and Stocking–Lord methods are then derived. The derivations are illustrated using a simulation study and lastly the results are discussed.

Polytomous IRT Models

In educational and psychological testing, items on a test are often scored in two or more categories. To model data from such a test, a polytomous IRT model is often considered suitable. A general model for such data is the generalized partial credit model (GPCM; Muraki, 1992), where the probability to obtain category k_j on item j , conditional on the ability level θ , is defined as

$$P_{j,k_j}(\theta) = \begin{cases} \frac{1}{1 + \sum_{c=2}^{m_j} \exp \left[\sum_{v=2}^c a_j(\theta - b_{j,v}) \right]}, & k_j = 1 \\ \frac{\exp \left[\sum_{v=2}^{k_j} a_j(\theta - b_{j,v}) \right]}{1 + \sum_{c=2}^{m_j} \exp \left[\sum_{v=2}^c a_j(\theta - b_{j,v}) \right]}, & k_j > 1 \end{cases}, \quad (1)$$

where m_j denotes the number of categories for item j . The parameter a_j is called the discrimination parameter and the parameters $b_{j,v}$ are called the item category parameters. The probability $P_{j,k_j}(\theta)$ is called the item category response function. The item category response functions are used when estimating the linking coefficients with the Haebara method. Let J be the number of

common items on two tests and let $k_j - 1$ denote the score given for category k_j on item j . The test response function, conditional on θ , is then defined as

$$T(\theta) = \sum_{j=1}^J \sum_{k_j=1}^{m_j} (k_j - 1) P_{j,k_j}(\theta). \quad (2)$$

The test response function indicates the expected score for a given ability level and is used when estimating the linking coefficients with the Stocking–Lord method.

Linking Coefficients

Let a_{1j} , a_{2j} , $b_{1j,v}$ and $b_{2j,v}$ denote the parameters for a common item j on two different scales, pertaining to the different Groups 1 and 2. For many polytomous IRT models, such as the GPCM and the graded response model (GRM; Samejima, 1969), the parameters for Groups 1 and 2 using the linking parameters A and B are related as $a_{1j} = a_{2j}/A$ and $b_{1j,v} = Ab_{2j,v} + B$ and thus also through the inverse relationships $a_{2j} = Aa_{1j}$ and $b_{2j,v} = (b_{1j,v} - B)/A$ (Baker, 1992; Kim & Lee, 2006; Lord, 1980). When the parameters a_{1j} , a_{2j} , $b_{1j,v}$, and $b_{2j,v}$ are estimated, these equalities do not hold exactly. The linking parameters A and B are therefore also estimated, by utilizing the estimated parameters of the common items for Groups 1 and 2. There are mainly two types of methods for estimating the linking parameters: moment methods, which use the first or second moments of the estimated common item parameters in estimating the linking parameters, and response function methods, which estimate the linking coefficients by minimizing a distance measure based on the item or test response functions resulting from the estimated common item parameters. One of the moment methods, the mean–mean method, estimates the parameters A and B by

$$(\hat{A}_{MM}, \hat{B}_{MM})' = \begin{pmatrix} \frac{\sum_{j=1}^J \hat{a}_{2j}}{\sum_{j=1}^J \hat{a}_{1j}} \\ \frac{\sum_{j=1}^J \sum_{v=2}^{m_j} \hat{b}_{1j,v}}{\sum_{j=1}^J m_j - J} - \frac{\hat{A}_{MM} \sum_{j=1}^J \sum_{v=2}^{m_j} \hat{b}_{2j,v}}{\sum_{j=1}^J m_j - J} \end{pmatrix}. \quad (3)$$

In an alternative formulation, the geometric mean can be used in place of the arithmetic mean when estimating the A -parameter, obtaining the mean–geometric mean estimator

$$(\hat{A}_{MGM}, \hat{B}_{MGM})' = \begin{pmatrix} \frac{\exp\left(\frac{1}{J} \sum_{j=1}^J \log \hat{a}_{2j}\right)}{\exp\left(\frac{1}{J} \sum_{j=1}^J \log \hat{a}_{1j}\right)} \\ \frac{\sum_{j=1}^J \sum_{v=2}^{m_j} \hat{b}_{1j,v}}{\sum_{j=1}^J m_j - J} - \frac{\hat{A}_{MGM} \sum_{j=1}^J \sum_{v=2}^{m_j} \hat{b}_{2j,v}}{\sum_{j=1}^J m_j - J} \end{pmatrix}. \quad (4)$$

A third moment method uses the standard deviation of the b_{jk} -parameters in the respective groups to estimate the A -parameter. Hence, the mean–sigma estimator is defined as

$$(\hat{A}_{MS}, \hat{B}_{MS})' = \sqrt{\begin{pmatrix} \frac{\sum_{j=1}^J \sum_{v=2}^{m_j} \left(\frac{\sum_{j=1}^J \sum_{v=2}^{m_j} \hat{b}_{1j,v}}{\sum_{j=1}^J m_{j-J}} \right)^2}{\sum_{j=1}^J \sum_{v=2}^{m_j} \left(\frac{\sum_{j=1}^J \sum_{v=2}^{m_j} \hat{b}_{2j,v}}{\sum_{j=1}^J m_{j-J}} \right)^2} \\ \frac{\sum_{j=1}^J \sum_{v=2}^{m_j} \hat{b}_{1j,v}}{\sum_{j=1}^J m_{j-J}} - \frac{\hat{A}_{MS} \sum_{j=1}^J \sum_{v=2}^{m_j} \hat{b}_{2j,v}}{\sum_{j=1}^J m_{j-J}} \end{pmatrix}}. \quad (5)$$

The variances of $(\hat{A}_{MM}, \hat{B}_{MM})'$ and $(\hat{A}_{MS}, \hat{B}_{MS})'$ can be calculated using the delta method because they are explicit functions of the item parameters and formulas for these were given in Wong (2015). The same applies for the mean-geometric mean method and the expression of the delta method large sample variance of $(\hat{A}_{MGM}, \hat{B}_{MGM})'$ is given in the supplementary material. Next, the estimation of the linking coefficients with the response function methods Haebara and Stocking-Lord is described.

Estimating Linking Coefficients With Response Function Methods

Let $P_{j,k_j}^{(1)}(\theta)$ denote the probability, conditional on θ , to answer in category k_j on item j where the parameters are expressed on the scale of Group 1 and let $P_{j,k_j}^{(2)}(\theta)$ denote the corresponding probability where the parameters are expressed on the scale of Group 2. That is, we have

$$P_{j,k_j}^{(1)}(\theta) = \frac{\exp \left[\sum_{v=2}^{k_j} a_{1j}(\theta - b_{1j,v}) \right]}{1 + \sum_{c=2}^{m_j} \exp \left[\sum_{v=2}^c a_{1j}(\theta - b_{1j,v}) \right]} \quad (6)$$

and

$$P_{j,k_j}^{(2)}(\theta) = \frac{\exp \left[\sum_{v=2}^{k_j} a_{2j}(\theta - b_{2j,v}) \right]}{1 + \sum_{c=2}^{m_j} \exp \left[\sum_{v=2}^c a_{2j}(\theta - b_{2j,v}) \right]}. \quad (7)$$

Because $a_{1j} = a_{2j}/A$ and $b_{1j,v} = Ab_{2j,v} + B$ and, equivalently, $a_{2j} = Aa_{1j}$ and $b_{2j,v} = (b_{1j,v} - B)/A$, the probabilities in Equations 6 and 7 can be expressed as

$$\begin{aligned} P_{j,k_j}^{(1)}(\theta) &= \frac{\exp \left[\sum_{v=2}^{k_j} a_{2j}/A (\theta - Ab_{2j,v} - B) \right]}{1 + \sum_{c=2}^{m_j} \exp \left[\sum_{v=2}^c a_{2j}/A (\theta - Ab_{2j,v} - B) \right]} \\ &= \frac{\exp \left[\sum_{v=2}^{k_j} a_{2j} \left(\frac{\theta - B}{A} - b_{2j,v} \right) \right]}{1 + \sum_{c=2}^{m_j} \exp \left[\sum_{v=2}^c a_{2j} \left(\frac{\theta - B}{A} - b_{2j,v} \right) \right]} \\ &= P_{j,k_j}^{(2)} \left(\frac{\theta - B}{A} \right) \end{aligned} \quad (8)$$

and

$$\begin{aligned}
 P_{j,k_j}^{(2)}(\theta) &= \frac{\exp \left[\sum_{v=2}^{k_j} A a_{1j} (\theta - (b_{1j,v} - B)/A) \right]}{1 + \sum_{c=2}^{m_j} \exp \left[\sum_{v=2}^c A a_{1j} (\theta - (b_{1j,v} - B)/A) \right]} \\
 &= \frac{\exp \left[\sum_{v=2}^{k_j} a_{1j} (A\theta + B - b_{1j,v}) \right]}{1 + \sum_{c=2}^{m_j} \exp \left[\sum_{v=2}^c a_{1j} (A\theta + B - b_{1j,v}) \right]} \\
 &= P_{j,k_j}^{(1)}(A\theta + B).
 \end{aligned} \tag{9}$$

When the parameters are estimated, the equalities in Equations 8 and 9 will not hold exactly for all values of θ and all categories k_j on each item j . The resulting probabilities in Equations 8 and 9 are used to define objective functions which are minimized with respect to the linking coefficients in order to estimate the linking coefficients.

The Haebara and Stocking–Lord methods estimate the linking coefficients A and B by minimizing the squared distance between the item category response (IR) functions and test response (TR) functions of the common items for two groups, respectively. The minimization is done over the entire latent distribution, something which necessitates the calculation of integrals with no explicit solution. Here, Gauss–Hermite quadrature will be utilized to approximate the integrals. Let $\theta_1, \dots, \theta_L, \dots, \theta_L$ denote the quadrature points and let $W(\cdot)$ denote the associated weight function. For the Haebara method, consider the objective function

$$f_{IR} = f_{IR1} + f_{IR2}, \tag{10}$$

where

$$f_{IR1} = \sum_{l=1}^L \sum_{j=1}^J \sum_{k_j=1}^{m_j} \left(P_{j,k_j}^{(1)}(\theta_l) - P_{j,k_j}^{(2)}\left(\frac{\theta_l - B}{A}\right) \right)^2 W(\theta_l) \tag{11}$$

and

$$f_{IR2} = \sum_{l=1}^L \sum_{j=1}^J \sum_{k_j=1}^{m_j} \left(P_{j,k_j}^{(2)}(\theta_l) - P_{j,k_j}^{(1)}(A\theta_l + B) \right)^2 W(\theta_l). \tag{12}$$

For the Stocking–Lord method, consider the objective function

$$f_{TR} = f_{TR1} + f_{TR2}, \tag{13}$$

where

$$f_{TR1} = \sum_{l=1}^L \left(\sum_{j=1}^J \sum_{k_j=1}^{m_j} (k_j - 1) \left(P_{j,k_j}^{(1)}(\theta_l) - P_{j,k_j}^{(2)}\left(\frac{\theta_l - B}{A}\right) \right) \right)^2 W(\theta_l) \tag{14}$$

and

$$f_{\text{TR2}} = \sum_{l=1}^L \left(\sum_{j=1}^J \sum_{k_j=1}^{m_j} (k_j - 1) \left(P_{j,k_j}^{(2)}(\theta_l) - P_{j,k_j}^{(1)}(A\theta_l + B) \right) \right)^2 W(\theta_l). \quad (15)$$

When estimating A and B , the resulting probabilities from estimates of a_{1j} , $b_{1j,v}$ and a_{2j} , $b_{2j,v}$ are used in the above expressions. To estimate the parameters A and B , the expressions in Equations 10 and 13 are minimized with respect to $(A, B)'$, something which necessitates numerical optimization. Following the notation in Ogasawara (2001), let $D \in \{\text{IR}, \text{TR}\}$ and define $(\hat{A}_D, \hat{B}_D)'$ as the estimator of the parameter vector $(A, B)'$ with either the Haebara (IR) or Stocking–Lord (TR) method. In the literature, three different objective functions have been considered for either of the two approaches, Haebara and Stocking–Lord. By considering only f_{IR1} or f_{TR1} , the forward transformation is achieved, while by considering only f_{IR2} or f_{TR2} , the backward transformation is achieved. By considering the summation of f_{IR1} and f_{IR2} or f_{TR1} and f_{TR2} , the combination transformation is achieved. The focus in this article will be on the most general formulation of the methods, which is the combination transformation. For further discussion, see S. Kim and Kolen (2007).

Asymptotic Covariance of Response Function Estimators of Linking Coefficients

Let α denote the vector of item parameters for all the common items expressed on two different scales. Because f_D is minimized, we have that $\frac{\partial f_D}{\partial(A, B)} = (0, 0)$. Furthermore, because f_D is a continuous and differentiable function of the item parameters α and provided that the matrix $\frac{\partial^2 f_D}{\partial(A, B)' \partial(A, B)}$ has a nonzero determinant, we may apply the implicit function theorem to ascertain the existence of functions $g_{A(D)}$ and $g_{B(D)}$ such that $A = g_{A(D)}(\alpha)$ and $B = g_{B(D)}(\alpha)$. From the implicit function theorem we then have that the partial derivatives of $(g_{A(D)}(\alpha), g_{B(D)}(\alpha))'$ with respect to α can be calculated by

$$\frac{\partial(g_{A(D)}(\alpha), g_{B(D)}(\alpha))'}{\partial \alpha} = - \left(\frac{\partial^2 f_D}{\partial(A, B)' \partial(A, B)} \right)^{-1} \frac{\partial^2 f_D}{\partial(A, B)' \partial \alpha'}. \quad (16)$$

The matrices $\frac{\partial^2 f_D}{\partial(A, B)' \partial(A, B)}$ and $\frac{\partial^2 f_D}{\partial(A, B)' \partial \alpha'}$ in Equation 16 are calculated with straightforward techniques but the derivations are quite long, so the matrices are provided as supplementary material. Let $\sum_{\hat{\alpha}}$ denote the asymptotic covariance matrix of the estimator of α . By using the delta method for implicit functions (Benichou & Gail, 1989), the asymptotic covariance matrix of $(\hat{A}_D, \hat{B}_D)'$ is given by

$$\sum_{(\hat{A}_D, \hat{B}_D)'} = \frac{\partial(g_{A(D)}(\alpha), g_{B(D)}(\alpha))'}{\partial \alpha} \sum_{\hat{\alpha}} \left[\frac{\partial(g_{A(D)}(\alpha), g_{B(D)}(\alpha))'}{\partial \alpha} \right]'. \quad (17)$$

When estimating $\sum_{(\hat{A}_D, \hat{B}_D)'}$, the parameter vector α in the above equation is replaced by the estimated item parameter vector.

Simulation Study

Design

Data for two tests X and Y consisting of 20 or 25 items with either three or five categories (scored 0–2 or 0–4) were simulated using the GPCM. Two settings were considered for each item type, one where the tests X and Y had five common items and one where they had 10

common items. For the three-category items, the discrimination parameters for tests X and Y were selected from $U(0.3, 1.3)$ and the item category parameters were selected from $N(-.5, 1)$ and $N(.5, 1)$. This parameter selection procedure is identical to that used for the three-category item tests in Wong (2015). With the five-category items, for the unique items on tests X and Y , the discrimination parameters were selected by drawing random numbers from the log-normal distribution with parameters $\mu = 0$ and $\sigma = 0.2$ while the item category parameters for each item were selected from the $N(-1.5, .2)$, $N(-.5, .2)$, $N(.5, .2)$, and $N(1.5, .2)$ distributions. This procedure is identical to the parameter selection in Kim and Lee (2006). The parameters for the common items were selected to be identical to the common item parameters for the GPCM given in Kim and Lee (2006). The ability distributions for the two groups were selected to be $N(0, 1)$ and $N(0.5, 1.2)$. Thus, the true linking coefficients from Group 1 to Group 2 were $A = 1.2$ and $B = 0.5$. Sample sizes 250, 500, 1,000, and 2,000 were considered and 2,000 replications were used for each setting.

The item parameters were estimated using marginal maximum likelihood with the statistical programming language R (R Development Core Team, 2016), either separately for each group or simultaneously in a multigroup setting (concurrent calibration). Version 1.20.1 of the R (R Development Core Team, 2016) package mirt (Chalmers, 2012) was used for the item parameter estimation. With separate estimation, both groups were assumed to have an underlying $N(0, 1)$ distribution while for concurrent calibration the first group was assumed to have an underlying $N(0, 1)$ distribution and the second group was assumed to have an underlying $N(\mu_2, \sigma_2)$ distribution, where μ_2 and σ_2 were free parameters to be estimated. To be able to estimate the parameters μ_2 and σ_2 , the common item parameters were restricted to be equal between the two groups. For the concurrent calibration method, the estimators of the linking coefficients A and B are with the setup considered here the estimators of the square root of the variance, $\hat{\sigma}_2 = \sqrt{\hat{\sigma}_2^2}$, and the mean, $\hat{\mu}_2$, of the latent distribution for Group 2. The sandwich estimator was used to estimate the asymptotic covariance matrix of the item parameters in the separate estimation. The asymptotic covariance matrix was not calculated for concurrent calibration because Version 1.20.1 of mirt does not provide estimates of the variance of $\hat{\mu}_2$ and $\hat{\sigma}_2^2$. After estimation, the Haebara, Stocking-Lord, mean-mean, mean-geometric mean, mean-sigma, and concurrent calibration estimates of the linking coefficients were calculated using newly written R code. The Monte Carlo standard error (MCSE), average estimated asymptotic standard error (ASE) and bias were calculated for each condition, except that ASE was not calculated for the concurrent calibration method because the asymptotic covariance matrix was not calculated. The relative efficiency (RE) between two estimators $\hat{T}_1$ and $\hat{T}_2$ defined as

$$RE = \frac{MSE(\hat{T}_1)}{MSE(\hat{T}_2)}$$

was also calculated, where MSE denotes the Mean Squared Error. The nonparametric bootstrap was used to calculate the standard errors of the MCSE and ASE and the confidence intervals for the RE in the simulation study.

Results

The results from the simulation with the three-category items are given in Table 1 for the case of five common items and in Table 2 for the case of 10 common items. The ASEs are accurate for all sample sizes and methods considered and hence there is no difference in the accuracy of the ASE between the five different linking coefficient estimators. The standard errors for the Haebara and Stocking-Lord methods are lower than those for the moment methods and the

Table 1. ASE and MCSE ($\times 10$) and Bias ($\times 10$) for Estimators of Linking Coefficients A and B With the Three-Category Items and Five Common Items.

Sample size	Estimator	A-parameter			B-parameter		
		ASE	MCSE	Bias	ASE	MCSE	Bias
250	H	1.33	1.37 (0.02)	0.08 (0.03)	1.35	1.33 (0.02)	-0.05 (0.03)
	SL	1.39	1.45 (0.03)	0.11 (0.03)	1.38	1.37 (0.02)	0.01 (0.03)
	MGM	1.65	1.71 (0.03)	0.17 (0.04)	1.59	1.62 (0.03)	0.13 (0.04)
	MM	1.57	1.61 (0.03)	0.14 (0.04)	1.61	1.63 (0.03)	0.12 (0.04)
	MS	2.05	2.06 (0.05)	0.23 (0.04)	1.70	1.69 (0.03)	0.14 (0.04)
	CC	—	1.36 (0.02)	0.08 (0.03)	—	1.32 (0.02)	0.01 (0.03)
500	H	0.94	0.96 (0.02)	0.04 (0.02)	0.95	0.94 (0.02)	-0.00 (0.02)
	SL	0.98	1.00 (0.02)	0.05 (0.02)	0.97	0.96 (0.02)	0.03 (0.02)
	MGM	1.14	1.15 (0.02)	0.08 (0.03)	1.10	1.09 (0.02)	0.09 (0.02)
	MM	1.10	1.12 (0.02)	0.07 (0.03)	1.11	1.10 (0.02)	0.09 (0.02)
	MS	1.33	1.32 (0.02)	0.09 (0.03)	1.15	1.14 (0.02)	0.10 (0.03)
	CC	—	0.96 (0.02)	0.03 (0.02)	—	0.93 (0.01)	0.02 (0.02)
1,000	H	0.66	0.65 (0.01)	0.01 (0.01)	0.67	0.65 (0.01)	-0.01 (0.01)
	SL	0.69	0.69 (0.01)	0.02 (0.02)	0.69	0.67 (0.01)	0.01 (0.01)
	MGM	0.80	0.82 (0.01)	0.04 (0.02)	0.77	0.75 (0.01)	0.04 (0.02)
	MM	0.77	0.79 (0.01)	0.03 (0.02)	0.78	0.76 (0.01)	0.04 (0.02)
	MS	0.91	0.90 (0.01)	0.04 (0.02)	0.80	0.78 (0.01)	0.04 (0.02)
	CC	—	0.65 (0.01)	0.00 (0.01)	—	0.65 (0.01)	-0.00 (0.01)
2,000	H	0.47	0.48 (0.01)	0.01 (0.01)	0.47	0.48 (0.01)	-0.02 (0.01)
	SL	0.48	0.49 (0.01)	0.01 (0.01)	0.48	0.49 (0.01)	-0.01 (0.01)
	MGM	0.56	0.54 (0.01)	0.02 (0.01)	0.54	0.54 (0.01)	-0.00 (0.01)
	MM	0.54	0.53 (0.01)	0.01 (0.01)	0.54	0.55 (0.01)	-0.00 (0.01)
	MS	0.64	0.63 (0.01)	0.03 (0.01)	0.56	0.56 (0.01)	0.01 (0.01)
	CC	—	0.47 (0.01)	0.00 (0.01)	—	0.48 (0.01)	-0.02 (0.01)

Note. ASE = asymptotic standard error; MCSE = Monte Carlo standard error; H = Haebara; SL = Stocking-Lord; MGM = mean-geometric mean; MM = mean-mean; MS = mean-sigma; CC = concurrent calibration.

standard errors are smaller with 10 common items compared with the case of five common items. The Haebara method has smaller MCSE than the Stocking-Lord method for all conditions. The largest difference between the two response function methods is for the *A*-parameter with five common items, where the Haebara method has around 5% lower MCSE for sample sizes 250, 500, and 1,000. Overall, the concurrent calibration method and the Haebara method have MCSEs which are the lowest and there is no clear difference between these two methods. The moment methods have larger standard errors with the mean-sigma method having the highest standard errors. The biases are positive but quite small for most conditions and estimators, and become lower with an increased sample size. The largest biases exist for the moment methods with sample size 250 but even these are not substantial.

The results from the simulation with the five-category items are given in Table 3 for five common items and in Table 4 for 10 common items. The ASEs are accurate for all settings and methods. Overall, the results show that there are virtually no differences between the Haebara and Stocking-Lord methods. For all settings, the moment methods perform the worst with higher standard errors and higher bias. In contrast to the results with three-category items, the mean-sigma method performs the best among the moment methods with respect to the standard errors. The differences between the moment methods and the other estimators are largest when estimating the *A*-parameter. The concurrent calibration method has the lowest MCSEs for all conditions except when estimating the *B*-parameter with 10 common items. The differences to the Haebara and Stocking-Lord methods are very small.

Table 2. ASE and MCSE ($\times 10$) and Bias ($\times 10$) for Estimators of Linking Coefficients A and B With the Three-Category Items and 10 Common Items.

Sample size	Estimator	A-parameter			B-parameter		
		ASE	MCSE	Bias	ASE	MCSE	Bias
250	H	1.11	1.13 (0.02)	0.07 (0.03)	1.20	1.18 (0.02)	-0.03 (0.03)
	SL	1.14	1.19 (0.02)	0.10 (0.03)	1.23	1.21 (0.02)	0.04 (0.03)
	MGM	1.30	1.36 (0.03)	0.15 (0.03)	1.40	1.40 (0.03)	0.18 (0.03)
	MM	1.22	1.26 (0.02)	0.11 (0.03)	1.38	1.37 (0.02)	0.17 (0.03)
	MS	1.90	1.86 (0.05)	0.26 (0.04)	1.55	1.49 (0.04)	0.22 (0.03)
	CC	—	1.13 (0.03)	0.07 (0.03)	—	1.18 (0.02)	0.04 (0.03)
500	H	0.78	0.80 (0.01)	0.02 (0.02)	0.85	0.84 (0.01)	-0.02 (0.02)
	SL	0.80	0.83 (0.01)	0.03 (0.02)	0.86	0.85 (0.01)	0.01 (0.02)
	MGM	0.89	0.92 (0.02)	0.06 (0.02)	0.95	0.93 (0.02)	0.07 (0.02)
	MM	0.85	0.87 (0.01)	0.04 (0.02)	0.94	0.92 (0.01)	0.06 (0.02)
	MS	1.17	1.16 (0.02)	0.09 (0.02)	1.00	0.99 (0.02)	0.08 (0.02)
	CC	—	0.79 (0.01)	0.02 (0.02)	—	0.83 (0.01)	0.00 (0.02)
1,000	H	0.55	0.54 (0.01)	0.01 (0.01)	0.60	0.59 (0.01)	-0.03 (0.01)
	SL	0.57	0.57 (0.01)	0.02 (0.01)	0.61	0.60 (0.01)	-0.01 (0.01)
	MGM	0.63	0.64 (0.01)	0.04 (0.01)	0.66	0.66 (0.01)	0.02 (0.01)
	MM	0.60	0.61 (0.01)	0.02 (0.01)	0.66	0.65 (0.01)	0.01 (0.01)
	MS	0.80	0.79 (0.01)	0.04 (0.02)	0.69	0.69 (0.01)	0.02 (0.02)
	CC	—	0.54 (0.01)	0.01 (0.01)	—	0.59 (0.01)	-0.02 (0.01)
2,000	H	0.39	0.39 (0.01)	0.00 (0.01)	0.42	0.43 (0.01)	-0.02 (0.01)
	SL	0.40	0.40 (0.01)	0.00 (0.01)	0.43	0.43 (0.01)	-0.01 (0.01)
	MGM	0.44	0.43 (0.01)	0.00 (0.01)	0.47	0.47 (0.01)	0.00 (0.01)
	MM	0.42	0.42 (0.01)	0.00 (0.01)	0.46	0.46 (0.01)	0.00 (0.01)
	MS	0.55	0.55 (0.01)	0.03 (0.01)	0.49	0.49 (0.01)	0.01 (0.01)
	CC	—	0.39 (0.01)	-0.00 (0.01)	—	0.42 (0.01)	-0.02 (0.01)

Note. ASE = asymptotic standard error; MCSE = Monte Carlo standard error; H = Haebara; SL = Stocking-Lord; MGM = mean-geometric mean; MM = mean-mean; MS = mean-sigma; CC = concurrent calibration.

In Tables 5 and 6, the confidence intervals for the relative efficiencies of concurrent calibration compared with each of the other linking coefficient estimators are displayed. For the RE, a value larger than 1 means that the estimator has lower MSE than concurrent calibration and a value smaller than 1 means that the estimator has higher MSE than concurrent calibration. Overall, the differences between concurrent calibration and the response function methods are small. The best-performing estimator relative to concurrent calibration is the Haebara method, which has an efficiency which is comparable to concurrent calibration overall. The Stocking-Lord method is almost as good, except that for the tests with three-category items the RE is lower than that for the Haebara method. The moment methods have lower relative efficiencies for all settings compared with the response function methods, with the mean-sigma method performing worse than the mean-mean and mean-geometric mean methods with the three-category items but performing better than them with the five-category items.

Discussion

In this article, the asymptotic variance of linking coefficients using the Haebara and Stocking-Lord methods were derived for polytomous IRT models. While the specific results were given only for the GPCM, it is straightforward to apply the results to other polytomous IRT models such as the GRM. For score reporting with IRT, the results of this article can be applied to

Table 3. ASE and MCSE ($\times 10$) and Bias ($\times 10$) for Estimators of Linking Coefficients A and B With the Five-Category Items and Five Common Items.

Sample size	Estimator	A-parameter			B-parameter		
		ASE	MCSE	Bias	ASE	MCSE	Bias
250	H	1.01	1.01 (0.02)	0.06 (0.02)	1.14	1.14 (0.02)	-0.01 (0.03)
	SL	1.00	1.01 (0.02)	0.05 (0.02)	1.14	1.14 (0.02)	0.01 (0.03)
	MGM	1.21	1.22 (0.02)	0.10 (0.03)	1.20	1.21 (0.02)	0.06 (0.03)
	MM	1.22	1.22 (0.02)	0.10 (0.03)	1.20	1.21 (0.02)	0.06 (0.03)
	MS	1.11	1.13 (0.02)	0.10 (0.03)	1.17	1.18 (0.02)	0.03 (0.03)
	CC	—	1.00 (0.02)	0.05 (0.02)	—	1.14 (0.02)	0.02 (0.03)
500	H	0.71	0.72 (0.01)	0.01 (0.02)	0.80	0.80 (0.01)	-0.02 (0.02)
	SL	0.71	0.72 (0.01)	0.01 (0.02)	0.80	0.80 (0.01)	-0.01 (0.02)
	MGM	0.85	0.87 (0.01)	0.03 (0.02)	0.84	0.85 (0.01)	0.01 (0.02)
	MM	0.85	0.87 (0.01)	0.02 (0.02)	0.84	0.85 (0.01)	0.01 (0.02)
	MS	0.78	0.79 (0.01)	0.00 (0.02)	0.82	0.82 (0.01)	-0.00 (0.02)
	CC	—	0.71 (0.01)	0.01 (0.02)	—	0.80 (0.01)	-0.01 (0.02)
1,000	H	0.50	0.51 (0.01)	0.02 (0.01)	0.57	0.56 (0.01)	-0.03 (0.01)
	SL	0.50	0.50 (0.01)	0.02 (0.01)	0.57	0.56 (0.01)	-0.02 (0.01)
	MGM	0.60	0.61 (0.01)	0.03 (0.01)	0.59	0.58 (0.01)	-0.01 (0.01)
	MM	0.60	0.61 (0.01)	0.03 (0.01)	0.59	0.59 (0.01)	-0.01 (0.01)
	MS	0.55	0.56 (0.01)	0.02 (0.01)	0.58	0.57 (0.01)	-0.01 (0.01)
	CC	—	0.50 (0.01)	0.01 (0.01)	—	0.56 (0.01)	-0.02 (0.01)
2,000	H	0.35	0.35 (0.01)	0.00 (0.01)	0.40	0.40 (0.01)	-0.01 (0.01)
	SL	0.35	0.35 (0.01)	-0.00 (0.01)	0.40	0.40 (0.01)	-0.01 (0.01)
	MGM	0.42	0.42 (0.01)	-0.01 (0.01)	0.42	0.42 (0.01)	-0.01 (0.01)
	MM	0.42	0.42 (0.01)	-0.01 (0.01)	0.42	0.42 (0.01)	-0.01 (0.01)
	MS	0.39	0.38 (0.01)	0.01 (0.01)	0.41	0.41 (0.01)	-0.00 (0.01)
	CC	—	0.35 (0.01)	-0.00 (0.01)	—	0.40 (0.01)	-0.01 (0.01)

Note. ASE = asymptotic standard error; MCSE = Monte Carlo standard error; H = Haebara; SL = Stocking–Lord; MGM = mean–geometric mean; MM = mean–mean; MS = mean–sigma; CC = concurrent calibration.

observed-score equating, as described in Andersson (2016), and to true-score equating, described in Wong (2015).

There are several versions of the Haebara and Stocking–Lord methods used in the literature. The most general forms of these two methods were used in this article, meaning that the results for the other versions follow directly from the forms considered here. Note that the derivations in the article also apply for the dichotomous IRT models which are special cases of the GPCM, such as the two-parameter logistic (2-PL) model. In this sense, the results of the article generalize the results of Ogasawara (2001) to all the versions of the Haebara and Stocking–Lord methods in the literature.

The results of the numerical study indicate that the ASEs are accurate for sample sizes as low as 250, suggesting that the derivations are appropriate to use in practice. Adding to several studies which have shown the superiority of the response function methods compared with the moment methods, this study indicates that the Haebara and Stocking–Lord methods outperform the moment methods mean–mean, mean–geometric mean, and mean–sigma with respect to both the sampling variance and the bias. Nevertheless, the ASEs for the moment methods were as accurate as those for the response function methods. The biases for the Haebara and Stocking–Lord methods were negligible and approximately the same as for concurrent calibration. Even so, a useful extension to this line of work is to derive the asymptotic bias of estimators of linking coefficients under correctly and incorrectly specified models, as has been done for correctly

Table 4. ASE and MCSE ($\times 10$) and Bias ($\times 10$) for Estimators of Linking Coefficients A and B With the Five-Category Items and 10 Common Items.

Sample size	Estimator	A-parameter			B-parameter		
		ASE	MCSE	Bias	ASE	MCSE	Bias
250	H	0.90	0.91 (0.01)	0.05 (0.02)	1.08	1.07 (0.02)	-0.01 (0.02)
	SL	0.89	0.91 (0.01)	0.04 (0.02)	1.08	1.07 (0.02)	0.01 (0.02)
	MGM	1.00	1.03 (0.02)	0.07 (0.02)	1.11	1.11 (0.02)	0.06 (0.02)
	MM	1.01	1.02 (0.02)	0.07 (0.02)	1.11	1.11 (0.02)	0.05 (0.02)
	MS	0.95	0.97 (0.02)	0.04 (0.02)	1.09	1.09 (0.02)	0.04 (0.02)
	CC	—	0.90 (0.01)	0.04 (0.02)	—	1.07 (0.02)	0.03 (0.02)
500	H	0.63	0.64 (0.01)	0.01 (0.01)	0.76	0.76 (0.01)	-0.03 (0.02)
	SL	0.63	0.64 (0.01)	0.01 (0.01)	0.76	0.76 (0.01)	-0.02 (0.02)
	MGM	0.71	0.72 (0.01)	0.02 (0.02)	0.78	0.78 (0.01)	0.01 (0.02)
	MM	0.71	0.72 (0.01)	0.02 (0.02)	0.78	0.78 (0.01)	0.01 (0.02)
	MS	0.67	0.67 (0.01)	0.02 (0.01)	0.77	0.77 (0.01)	-0.01 (0.02)
	CC	—	0.63 (0.01)	0.01 (0.01)	—	0.76 (0.01)	-0.01 (0.02)
1,000	H	0.45	0.44 (0.01)	0.01 (0.01)	0.54	0.53 (0.01)	-0.03 (0.01)
	SL	0.44	0.44 (0.01)	0.01 (0.01)	0.54	0.53 (0.01)	-0.03 (0.01)
	MGM	0.50	0.50 (0.01)	0.01 (0.01)	0.55	0.54 (0.01)	-0.02 (0.01)
	MM	0.50	0.50 (0.01)	0.01 (0.01)	0.55	0.54 (0.01)	-0.02 (0.01)
	MS	0.47	0.47 (0.01)	0.01 (0.01)	0.54	0.53 (0.01)	-0.02 (0.01)
	CC	—	0.44 (0.01)	0.01 (0.01)	—	0.53 (0.01)	-0.03 (0.01)
2,000	H	0.31	0.31 (0.00)	0.00 (0.01)	0.38	0.38 (0.01)	-0.01 (0.01)
	SL	0.31	0.31 (0.00)	0.00 (0.01)	0.38	0.38 (0.01)	-0.01 (0.01)
	MGM	0.35	0.36 (0.01)	-0.01 (0.01)	0.39	0.40 (0.01)	-0.01 (0.01)
	MM	0.35	0.35 (0.01)	-0.01 (0.01)	0.39	0.40 (0.01)	-0.01 (0.01)
	MS	0.33	0.33 (0.01)	0.01 (0.01)	0.38	0.39 (0.01)	-0.01 (0.01)
	CC	—	0.31 (0.00)	-0.00 (0.01)	—	0.38 (0.01)	-0.01 (0.01)

Note. ASE = asymptotic standard error; MCSE = Monte Carlo standard error; H = Haebara; SL = Stocking-Lord; MGM = mean-geometric mean; MM = mean-mean; MS = mean-sigma; CC = concurrent calibration.

Table 5. Confidence Intervals for the Relative Efficiency of Linking Coefficient Estimators From Separate Estimation Compared With Concurrent Calibration, Three-Category Items.

		A-parameter relative efficiency 95% CI				
Common items	Sample size	H	SL	MGM	MM	MS
5	250	[0.97, 1.00]	[0.85, 0.91]	[0.59, 0.67]	[0.68, 0.75]	[0.40, 0.47]
	500	[0.98, 1.00]	[0.90, 0.95]	[0.65, 0.72]	[0.69, 0.76]	[0.49, 0.56]
	1,000	[0.99, 1.01]	[0.87, 0.92]	[0.60, 0.66]	[0.65, 0.71]	[0.49, 0.56]
	2,000	[0.98, 1.00]	[0.93, 0.97]	[0.72, 0.80]	[0.75, 0.82]	[0.53, 0.59]
10	250	[0.99, 1.01]	[0.88, 0.93]	[0.66, 0.73]	[0.78, 0.85]	[0.33, 0.41]
	500	[0.98, 1.00]	[0.90, 0.94]	[0.71, 0.78]	[0.81, 0.87]	[0.43, 0.49]
	1,000	[0.98, 1.00]	[0.90, 0.94]	[0.69, 0.75]	[0.77, 0.83]	[0.44, 0.51]
	2,000	[0.98, 1.00]	[0.94, 0.98]	[0.78, 0.84]	[0.82, 0.89]	[0.46, 0.53]
		B-parameter relative efficiency 95% CI				
Common items	Sample size	H	SL	MGM	MM	MS
5	250	[0.99, 1.02]	[0.92, 0.96]	[0.64, 0.71]	[0.62, 0.69]	[0.58, 0.66]
	500	[0.98, 1.00]	[0.92, 0.96]	[0.70, 0.76]	[0.67, 0.74]	[0.63, 0.70]
	1,000	[0.97, 1.00]	[0.92, 0.97]	[0.72, 0.79]	[0.70, 0.77]	[0.66, 0.73]
	2,000	[0.97, 1.00]	[0.93, 0.97]	[0.74, 0.81]	[0.72, 0.78]	[0.68, 0.75]
10	250	[1.01, 1.04]	[0.95, 0.99]	[0.68, 0.75]	[0.71, 0.77]	[0.57, 0.67]
	500	[0.99, 1.01]	[0.95, 0.99]	[0.77, 0.83]	[0.79, 0.85]	[0.68, 0.75]
	1,000	[0.98, 1.00]	[0.93, 0.97]	[0.76, 0.82]	[0.78, 0.84]	[0.69, 0.76]
	2,000	[0.98, 1.00]	[0.96, 0.99]	[0.79, 0.85]	[0.81, 0.87]	[0.73, 0.79]

Note. Bold font indicates that the confidence interval does not cover 1. CI = confidence interval; H = Haebara; SL = Stocking-Lord; MGM = mean-geometric mean; MM = mean-mean; MS = mean-sigma.

Table 6. Confidence Intervals for the Relative Efficiency of Linking Coefficient Estimators From Separate Estimation Compared With Concurrent Calibration, Five-Category Items.

Common items	Sample size	A-parameter relative efficiency 95% CI				
		H	SL	MGM	MM	MS
5	250	[0.96, 0.99]	[0.97, 0.99]	[0.63, 0.70]	[0.63, 0.70]	[0.76, 0.82]
	500	[0.97, 0.99]	[0.98, 0.99]	[0.64, 0.71]	[0.64, 0.71]	[0.78, 0.84]
	1,000	[0.97, 0.99]	[0.98, 0.99]	[0.64, 0.71]	[0.64, 0.71]	[0.77, 0.83]
	2,000	[0.98, 1.00]	[0.99, 1.00]	[0.65, 0.72]	[0.65, 0.72]	[0.80, 0.86]
10	250	[0.97, 0.99]	[0.98, 0.99]	[0.74, 0.81]	[0.75, 0.81]	[0.84, 0.89]
	500	[0.98, 1.00]	[0.99, 1.00]	[0.75, 0.81]	[0.75, 0.81]	[0.86, 0.92]
	1,000	[0.98, 1.00]	[0.99, 1.00]	[0.74, 0.81]	[0.74, 0.81]	[0.85, 0.90]
	2,000	[0.98, 1.00]	[0.99, 1.00]	[0.74, 0.81]	[0.74, 0.81]	[0.87, 0.92]
Common items	Sample size	B-parameter relative efficiency 95% CI				
		H	SL	MGM	MM	MS
5	250	[1.00, 1.01]	[0.99, 1.01]	[0.86, 0.91]	[0.86, 0.91]	[0.91, 0.95]
	500	[1.00, 1.01]	[1.00, 1.01]	[0.87, 0.92]	[0.97, 0.92]	[0.94, 0.98]
	1,000	[1.00, 1.01]	[0.99, 1.00]	[0.90, 0.95]	[0.90, 0.95]	[0.94, 0.98]
	2,000	[0.99, 1.00]	[0.99, 1.01]	[0.88, 0.93]	[0.88, 0.92]	[0.93, 0.97]
10	250	[1.00, 1.02]	[0.99, 1.01]	[0.91, 0.95]	[0.91, 0.95]	[0.95, 0.98]
	500	[1.00, 1.01]	[1.00, 1.01]	[0.93, 0.97]	[0.92, 0.96]	[0.96, 0.99]
	1,000	[0.99, 1.01]	[0.99, 1.00]	[0.95, 0.99]	[0.95, 0.99]	[0.97, 1.00]
	2,000	[0.99, 1.01]	[1.00, 1.01]	[0.93, 0.96]	[0.93, 0.96]	[0.96, 0.99]

Note. Bold font indicates that the confidence interval does not cover 1. CI = confidence interval; H = Haebara; SL = Stocking–Lord; MGM = mean–geometric mean; MM = mean–mean; MS = mean–sigma.

specified dichotomous IRT models with the mean–mean and mean–sigma methods (Ogasawara, 2011).

This study indicates that when using marginal maximum likelihood estimation, the response function methods are almost as good as concurrent calibration with respect to the bias and standard error of the linking coefficients. For sample size 250 with both the three-category and five-category items, the Haebara method was sometimes even better than concurrent calibration, although the improvement was small. The Haebara method had better performance than the Stocking–Lord method for the tests with three-category items but not for the tests with five-category items. Some previous studies have indicated that the concurrent calibration method has overall better performance than using linking coefficients with separate estimation (Hanson & Béguin, 2002; Kim & Kolen, 2007) but examples of studies indicating the opposite are also available (S. H. Kim & Cohen, 1998). However, these studies used estimation methods which differed to the one used in this article, where the standard marginal maximum likelihood method was used. Furthermore, this study utilized simulated data from groups with differences in both the latent mean and the latent variance which the referenced studies did not.

The asymptotic variances of linking coefficient estimators derived in this article and in previous articles only account for the variability in estimating the item parameters. Other sources of variability such as the selection of the common items from a pool of items remain unaccounted for (Haberman, Lee, & Qian, 2009; Michaelides & Haertel, 2014). For large sample sizes, the common item selection could be the main source of variability because the variability of the item parameter estimation reduces with the sample size while the variability of the common item selection does not.

Last, the method of using separate estimation and then calculating the linking coefficients has clear benefits compared with concurrent calibration. For example, with many successive calibrations it may not be possible to achieve convergence for the full data using concurrent calibration even though for each individual group it is possible to achieve convergence. It is also easier to diagnose potential problems when estimating the item parameters separately (Hanson & Béguin, 2002). It should also be noted that in this study the estimation with concurrent calibration took approximately 10 times longer to conduct compared with separate estimation. Hence, the method of concurrent calibration may be computationally infeasible to conduct in practice, especially for large data sets.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplementary material is available for this article online.

References

- Andersson, B. (2016). Asymptotic standard errors of observed-score equating with polytomous IRT models. *Journal of Educational Measurement*, 53, 459-477.
- Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement*, 16, 87-96.
- Benichou, J., & Gail, M. H. (1989). A delta method for implicitly defined random variables. *The American Statistician*, 43, 41-44.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29.
- Haberman, S. J., Lee, Y.-H., & Qian, J. (2009). *Jackknifing techniques for evaluation of equating accuracy* (Research Report No. RR-09-39). Princeton, NJ: Educational Testing Service.
- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22, 144-149.
- 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.
- Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, 22, 131-143.
- 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.
- Kim, S., & Lee, W.-C. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement*, 43, 53-76.
- Kolen, M. J., & Brennan, R. J. (2014). *Test equating: Methods and practices* (3rd ed.). New York, NY: Springer-Verlag.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Loyd, B. H., & Hoover, H. (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.
- 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.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement*, 25, 53-67.
- Ogasawara, H. (2011). Applications of asymptotic expansion in item response theory linking. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 261-280). New York, NY: Springer.
- R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph Supplement No. 17.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- Wong, C. C. (2015). Asymptotic standard errors for item response theory true score equating of polytomous items. *Journal of Educational Measurement*, 52, 106-120.