UPPSALA
UNIVERSITET

Working Paper 2014:2

Department of Statistics

# IRT Observed-score Kernel Equating

Björn Andersson, Uppsala University

*and* Marie Wiberg, Umeå University

Working Paper 2014:2
May  2014
Department of Statistics
Uppsala University
Box 513
SE-751 20 UPPSALA
SWEDEN

Title: IRT Observed-score Kernel Equating

Author(s): Björn Andersson and Marie Wiberg
E-mail: bjorn.andersson@statistics.uu.se

# IRT Observed-Score Kernel Equating

Björn Andersson

Uppsala University

Marie Wiberg

Umeå University

## Abstract

Item response theory (IRT) observed-score kernel equating is described for the non-equivalent groups with anchor test equating design using either chain equating or post-stratification equating. The equating function is treated in a multivariate setting and the asymptotic covariance matrices of IRT observed-score kernel equating functions are derived. Equating is conducted using the two-parameter and three-parameter logistic models with simulated data and data from a standardized achievement test. The results show that IRT observed-score kernel equating offers small standard errors and low equating bias under most settings considered.

IRT Observed-Score Kernel Equating

## Introduction

Equating is a statistical method which is often employed in standardized testing programs to ensure that the results from different administrations of standardized tests are comparable. In observed-score equipercentile equating, the scores on two different tests from separate administrations are made comparable by matching the percentiles of the score distributions of the two tests. One type of observed-score equating is the kernel method of test equating (von Davier, Holland, & Thayer, 2004), an observed-score equating framework which enables the equating of standardized tests in all common equating designs. The framework allows for the usage of score probabilities which are either observed or estimated with a statistical model to conduct the equating. One alternative is to estimate an item response theory (IRT) model from the observed data and calculate the score probabilities derived from the IRT model. These score probabilities are then used to estimate continuous approximations to the discrete test score distributions and the resulting equating is called an IRT observed-score equating (Kolen & Brennan, 2004; Lord & Wingersky, 1984). In kernel equating the score probabilities have typically been estimated using log-linear models (Holland & Thayer, 1989; von Davier et al., 2004). Score probabilities derived from IRT models have also been considered (von Davier, 2010; Wiberg, van der Linden, & von Davier, 2014) but the asymptotic results have not been described. The purpose here is to describe IRT observed-score equating in the kernel equating framework. The results from IRT observed-score equating using linear interpolation (Ogasawara, 2003) are integrated in the framework and, similar to the treatment in Rijmen, Qu, and Davier (2011) for kernel equating using log-linear models, the asymptotic covariance matrix of the IRT observed-score kernel equating function is derived, extending the results from the linear interpolation case.

The structure is as follows. First, a description of different IRT models and their properties is given and the results of kernel equating are summarized. The score probabilities for the NEAT CE and NEAT PSE equating methods are defined and the asymptotic covariance matrices of these score probabilities are described. With these results it is shown how score probabilities from IRT models can be utilized in observed-score kernel equating. Next, the provided derivations are exemplified using simulated data and data from a standardized achievement test. Lastly, the results are discussed and concluding remarks are given.

## IRT Observed-Score Kernel Equating

Consider a test design where test $X$ with $k_X$ items is administered to a sample from population $P$ and test $Y$ with $k_Y$ items is administered to a sample from a separate population $Q$. In addition, the test $A$ with $k_A$ items is administered to both the sample from population $P$ and the sample from population $Q$. Let $N$ and $M$ denote the sample sizes for each test group. Such a test design is called a non-equivalent groups with anchor test (NEAT) design.

### IRT Models

Let $P_{Xl_X}(\theta)$, $P_{Yl_Y}(\theta)$, $P_{A_Pl_A}(\theta)$ and $P_{A_Ql_A}(\theta)$ be the probabilities to answer items $l_X \in \{1, 2, \ldots, k_X\}$, $l_Y \in \{1, 2, \ldots, k_Y\}$ , $l_A \in \{1, 2, \ldots, k_A\}$ of tests $X$, $Y$ and $A$ on populations $P$ and $Q$, respectively, correctly, viewed as functions of the ability level $\theta$. The assumptions of the IRT models considered here are that the ability is unidimensional, that the responses to each item are independent conditional on the ability, that the probabilities to answer each item correctly are invariant to the distribution of the ability and that there is no speededness to the test (Hambleton & Swaminathan, 1984). With the three parameter logistic (3-PL) model the probability is

$$P_{Xl_X}(\theta) = c_{Xl_X} + \frac{1 - c_{Xl_X}}{1 + \exp[-a_{Xl_X}(\theta - b_{Xl_X})]}, \tag{1}$$

where $a_{Xl_X}$ is the discrimination parameter for item $l_X$, $b_{Xl_X}$ is the difficulty parameter for item $l_X$ and $c_{Xl_X}$ is the guessing parameter for item $l_X$ (Hambleton & Swaminathan, 1984).

Equation 1 implies the two-parameter logistic (2-PL) model if $c_{Xl_X} = 0$ and the one-parameter logistic (1-PL) model if $c_{Xl_X} = 0$ and $a_{Xl_X} = 1$. The corresponding probabilities for test $Y$ and test A on populations $P$ and Q, $P_{Yl_Y}(\theta)$, $P_{A_Pl_A}(\theta)$ and $P_{A_Ql_A}(\theta)$, are defined analogously.

In the rest of the paper it is assumed that the 3-PL model is used. Let $\boldsymbol{\alpha}_P = (\boldsymbol{\alpha}_X, \boldsymbol{\alpha}_{A_P})$ and $\boldsymbol{\alpha}_Q = (\boldsymbol{\alpha}_Y, \boldsymbol{\alpha}_{A_Q})$ denote the $3(k_X + k_A) \times 1$ and $3(k_Y + k_A) \times 1$ vectors of all item parameters for the samples from populations $P$ and $Q$. If $\boldsymbol{\alpha}_P$ and $\boldsymbol{\alpha}_Q$ are estimated using maximum likelihood (Bock & Aitkin, 1981; Lord, 1980), the estimators $\hat{\boldsymbol{\alpha}}_P$ and $\hat{\boldsymbol{\alpha}}_Q$ are, under suitable regularity conditions, asymptotically normal distributed (Lord, 1983) with the respective covariance matrices $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_P}$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_Q}$. This is assumed to hold in what follows. In practice the covariance matrices can be estimated with, for example, the observed information matrix or the related, robust, sandwich estimator (Louis, 1982; Yuan, Cheng, & Patton, 2013).

## Kernel Equating

In kernel equating, continuous approximations to the discrete score distributions of the tests to be equated are estimated by utilizing a kernel. The Gaussian kernel has often been used, since it offers the properties of continuity and differentiability which are necessary for the derivations of the asymptotic distributions.

For an equipercentile equating, the true equating function from $X$ to $Y$ is defined as

$$\mathrm{eq_Y}(x) = G^{-1}\left[F(x)\right], \tag{2}$$

where $F(\cdot)$ and $G(\cdot)$ are the distribution functions of tests $X$ and $Y$, respectively. The distribution functions $F(\cdot)$ and $G(\cdot)$ are typically not continuous so in equating continuous approximations are used. In kernel equating, the continuous approximations are denoted $F_{h_X}(\cdot)$ and $G_{h_Y}(\cdot)$ and are functions of the observed or estimated score probabilities for tests $X$ and $Y$, constructed using a Gaussian kernel. Following the notation in von Davier et al. (2004), for score value $x$ of test $X$, the continuous approximation to $F(x)$ is

$$F_{h_X}(x) = \sum_{j=0}^{k_X} r_j \Phi\left(\frac{x - a_X x_j - (1 - a_X)\mu_X}{a_X h_X}\right), \tag{3}$$

where $r_j$ is the score probability for the $j$-th score value, $\Phi(\cdot)$ denotes the standard normal distribution function, $x_j$ is the $j$-th score value, $\mu_X$ is the mean of the test scores, $h_X$ is the bandwidth and

$$a_X = \sqrt{\frac{\sigma_X^2}{\sigma_X^2 + h_X^2}}, \tag{4}$$

where $\sigma_X^2$ is the variance of the test scores. The continuous approximation $G_{h_Y}(\cdot)$ is defined correspondingly. The bandwidth $h_X$ is selected by minimizing the function $\text{PEN}(h_X) = \sum_{j=0}^{k_X}[r_j - \frac{d}{dx}F_{h_X}(x)|^{x=x_j}]^2$. Note that $F_{h_X}(\cdot)$ is a continuous and differentiable function with respect to the score probabilities $r_j$.

With a NEAT design either a chain equating (CE) or a post-stratification equating (PSE) method can be employed. The kernel NEAT CE function for score value $x$ is

$$e_{\text{Y(CE)}}(x) = G_{h_Y}^{-1}\left(H_{Qh_{AQ}}\left(H_{Ph_{AP}}^{-1}\left(F_{h_X}(x)\right)\right)\right), \tag{5}$$

where $F_{h_X}(\cdot)$, $H_{Ph_{AP}}(\cdot)$, $G_{h_Y}(\cdot)$ and $H_{Qh_{AQ}}(\cdot)$ are the continuous approximations to the distribution functions of the score values of tests $X$ and $A$ on $P$ and $Y$ and $A$ on $Q$, respectively. The kernel NEAT PSE equating function for score value $x$ is defined as

$$e_{\text{Y(PSE)}}(x) = G_{Sh_Y}^{-1}\left[F_{Sh_X}(x)\right], \tag{6}$$

where $G_{Sh_Y}(\cdot)$ and $F_{Sh_X}(\cdot)$ are the continuous approximations to the distribution functions of tests $X$ and $Y$ on the synthetic population $S = w \times P + (1 - w) \times Q$, $w \in [0, 1]$, a mixture of the populations $P$ and $Q$.

## Generating the Score Probabilities From IRT Models

The necessary components for conducting an equating are the vectors of score probabilities implied by the IRT models. The derivation and asymptotic covariance matrices of these vectors for the NEAT CE and PSE designs are given here.

**NEAT CE.** Let $\mathbf{x} = \begin{pmatrix} 0 & 1 & \dots & k_X \end{pmatrix}'$, $\mathbf{y} = \begin{pmatrix} 0 & 1 & \dots & k_Y \end{pmatrix}'$ and $\mathbf{a} = \begin{pmatrix} 0 & 1 & \dots & k_A \end{pmatrix}'$ be the vectors of possible score values of tests $X$, $Y$ and $A$. Now, let $r_i$, $s_j$, $t_{P,k}$ and $t_{Q,k}$ be the probabilities to obtain score values $x_i, y_j$ and $a_k$ for populations $P$ and $Q$. Let $\boldsymbol{r}$ and $\boldsymbol{s}$ be the $(k_X+1)\times 1$ and $(k_Y+1)\times 1$ vectors of probabilities $r_i$ and $s_j$ to obtain each of the score values $x_i \in \{0, 1, \dots, k_X\}$ and $y_j \in \{0, 1, \dots, k_Y\}$ on the tests $X$ and $Y$, respectively, and let $\boldsymbol{t}_P$ and $\boldsymbol{t}_Q$ be the $(k_A+1)\times 1$ vectors of probabilities $t_{P,k}$ and $t_{Q,k}$ to obtain each of the score values $a_k \in \{0, 1, \dots k_A\}$ on test A.

Let $r_i(\theta)$ be the $i$-th score probability for test $X$, viewed as a function of the ability level $\theta$. The score probabilities $r_i(\theta)$ are obtained from the probabilities defined in Equation 1 by the algorithm in Lord and Wingersky (1984). The $i$-th score probability across all ability levels is approximated by

$$r_i \approx \sum_{r=1}^{R} r_i(t_r) W(t_r), \tag{7}$$

where $t_r$ denotes the ability level for the $r$-th quadrature point, $r \in \{1, 2, \dots, R\}$, and where $W(\cdot)$ is a weight function such that each quadrature point is weighted in accordance with the assumptions made about the distribution of the ability level. Corresponding expressions apply for $s_j$, $t_{A_P,k}$ and $t_{A_Q,k}$.

Since the estimators $\hat{\boldsymbol{\alpha}}_P$ and $\hat{\boldsymbol{\alpha}}_Q$ are independent and $(\hat{\boldsymbol{r}}, \hat{\boldsymbol{t}}_P)$ is a function of $\hat{\boldsymbol{\alpha}}_P$ only and $(\hat{\boldsymbol{s}}, \hat{\boldsymbol{t}}_Q)$ is a function of $\hat{\boldsymbol{\alpha}}_Q$ only, the asymptotic covariance matrix of the estimator of the score probabilities, $(\hat{\boldsymbol{r}}, \hat{\boldsymbol{t}}_P, \hat{\boldsymbol{s}}, \hat{\boldsymbol{t}}_Q)$, is

$$\boldsymbol{\Sigma}_{(\hat{\boldsymbol{r}}, \hat{\boldsymbol{t}}_P, \hat{\boldsymbol{s}}, \hat{\boldsymbol{t}}_Q)} = \begin{pmatrix} \boldsymbol{\Sigma}_{(\hat{\boldsymbol{r}}, \hat{\boldsymbol{t}}_P)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{(\hat{\boldsymbol{s}}, \hat{\boldsymbol{t}}_Q)} \end{pmatrix}, \tag{8}$$

where

$$\boldsymbol{\Sigma}_{(\hat{\boldsymbol{r}}, \hat{\boldsymbol{t}}_P)} = \frac{\partial(\boldsymbol{r}, \boldsymbol{t}_P)}{\partial \boldsymbol{\alpha}_P} \boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_P} \left[ \frac{\partial(\boldsymbol{r}, \boldsymbol{t}_P)}{\partial \boldsymbol{\alpha}_P} \right]' \tag{9}$$

and

$$\boldsymbol{\Sigma}_{(\hat{\boldsymbol{s}}, \hat{\boldsymbol{t}}_Q)} = \frac{\partial(\boldsymbol{s}, \boldsymbol{t}_Q)}{\partial \boldsymbol{\alpha}_Q} \boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_Q} \left[ \frac{\partial(\boldsymbol{s}, \boldsymbol{t}_Q)}{\partial \boldsymbol{\alpha}_Q} \right]', \tag{10}$$

where $\frac{\partial(\boldsymbol{r},\boldsymbol{t}_P)}{\partial\boldsymbol{\alpha}_P}$ and $\frac{\partial(\boldsymbol{s},\boldsymbol{t}_Q)}{\partial\boldsymbol{\alpha}_Q}$ are matrices of partial derivatives with entries given in Ogasawara (2003).

**NEAT PSE.** In PSE, the two populations $P$ and $Q$ are weighted to create a synthetic population $S$ for which the equating is conducted. This is accomplished by defining two new distributions of the score probabilities for tests $X$ and $Y$ on the synthetic population $S$. The score probabilities derived from IRT models in the NEAT PSE design are defined differently compared to the case of using score probabilities derived from log-linear models, so the results in von Davier et al. (2004) cannot be used directly. Instead, the method described for equipercentile equating using linear interpolation in Ogasawara (2003) will be modified and applied to the kernel equating framework.

The NEAT PSE design requires the usage of equating coefficients $\beta_A$ and $\beta_B$ which place the abilities on the same scale in the two populations (Kolen & Brennan, 2004). To achieve this, the probabilities to answer each item $l_Y$ correctly for the test $Y$ are defined as, (Ogasawara, 2003),

$$P^{\text{PSE}}_{Yl_Y}(\theta) = c_{Yl_Y} + \frac{1 - c_{Yl_Y}}{1 + \exp[-(a_{Yl_Y}/\beta_A)(\theta - \beta_A b_{Yl_Y} - \beta_B)]}. \tag{11}$$

Let $r_{Pi}$ and $s_{Pj}$ be the $i$-th and $j$-th score probabilities for test $X$ and $Y$ on population $P$ and let $r_{Qi}$ and $s_{Qj}$ be the $i$-th and $j$-th score probabilities for tests $X$ and $Y$ on population $Q$. These score probabilities are defined as

$$r_{Pi} \approx \sum_{r=1}^{R} r_{Pi}(t_r)W(t_r), r_{Qi} \approx \sum_{r=1}^{R} r_{Pi}(\beta_A t_r + \beta_B)W(t_r) \tag{12}$$

and

$$s_{Pj} \approx \sum_{r=1}^{R} s_{Qj}(t_r)W(t_r), s_{Qj} \approx \sum_{r=1}^{R} s_{Qj}(\beta_A t_r + \beta_B)W(t_r), \tag{13}$$

where $r_{Pi}(\cdot)$ and $s_{Qj}(\cdot)$ are calculated from the probabilities in Equation 1 and Equation 11, respectively, again with the recursive algorithm in Lord and Wingersky (1984). Let the corresponding vectors of probabilities for each score value be $\boldsymbol{r}_P$, $\boldsymbol{r}_Q$, $\boldsymbol{s}_P$ and $\boldsymbol{s}_Q$. Note that $\boldsymbol{r}_P$ and $\boldsymbol{s}_Q$ are identical to the probabilities $\boldsymbol{r}$ and $\boldsymbol{s}$ used in NEAT CE.

The NEAT PSE design requires the score probabilities $r_{Si}$ and $s_{Sj}$ for each score value on tests $X$ and $Y$ on the synthetic population $S$. Thus, in the NEAT PSE design the score probabilities used in the equating are (von Davier et al., 2004)

$$r_{Si} = w_S r_{Pi} + (1 - w_S) r_{Qi}, \tag{14}$$

and

$$s_{Sj} = w_S s_{Pj} + (1 - w_S) s_{Qj}, \tag{15}$$

where $w_S \in [0, 1]$ is the weight for population $P$ and $1 - w_S$ is the weight for population $Q$. Let the vectors of probabilities to achieve each score value on tests $X$ and $Y$ on population $S$ be denoted $\boldsymbol{r}_S$ and $\boldsymbol{s}_S$. In the above exposition, the parameters from the anchor items which constitute test $A$ are not directly used. The parameters of the anchor items are however used in the estimation of the equating coefficients $\beta_A$ and $\beta_B$. There are two main ways of estimating the coefficients: methods based on moments such as the Mean-Sigma (Marco, 1977), Mean-Mean (Loyd & Hoover, 1980) and Mean-Geometric mean (Mislevy & Bock, 1990) methods and methods based on the test characteristic curves such as the Haebara method (Haebara, 1980) and the Stocking-Lord method (Stocking & Lord, 1983). See Kolen and Brennan (2004) for an overview of the relative merits of each method and Ogasawara (2000, 2001) for the asymptotic standard errors of the equating coefficients. Any of these methods of estimating the equating coefficients may be used in IRT observed-score kernel equating.

Since the estimators of the IRT item parameters $\boldsymbol{\alpha}_P$ and $\boldsymbol{\alpha}_Q$ are independent and have asymptotic covariance matrices $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_P}$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_Q}$, the estimator $(\hat{\boldsymbol{\alpha}}_P, \hat{\boldsymbol{\alpha}}_Q)$ has the asymptotic covariance matrix

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_P, \hat{\boldsymbol{\alpha}}_Q} = \begin{pmatrix} \boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_P} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_Q} \end{pmatrix}. \tag{16}$$

The vector $(\boldsymbol{r}_S, \boldsymbol{s}_S)$ is a function of the vector $(\boldsymbol{r}_P, \boldsymbol{r}_Q, \boldsymbol{s}_P, \boldsymbol{s}_Q)$ which in turn is a function of the vector $(\boldsymbol{\alpha}_X, \boldsymbol{\alpha}_Y, \beta_A, \beta_B)$ which is a function of $(\boldsymbol{\alpha}_P, \boldsymbol{\alpha}_Q)$. Thus, provided that the equating coefficients $\beta_A$ and $\beta_B$ are functions of the item parameters which are continuous and differentiable, the vector $(\boldsymbol{r}_S, \boldsymbol{s}_S)$ is a differentiable, continuous function of the item parameters $(\boldsymbol{\alpha}_P, \boldsymbol{\alpha}_Q)$ and thus its estimator $(\hat{\boldsymbol{r}}_S, \hat{\boldsymbol{s}}_S)$ has asymptotic covariance matrix

$$\boldsymbol{\Sigma}_{(\hat{\boldsymbol{r}}_S, \hat{\boldsymbol{s}}_S)} = \frac{\partial(\boldsymbol{r}_S, \boldsymbol{s}_S)}{\partial(\boldsymbol{\alpha}_P, \boldsymbol{\alpha}_Q)} \boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_P, \hat{\boldsymbol{\alpha}}_Q} \left[ \frac{\partial(\boldsymbol{r}_S, \boldsymbol{s}_S)}{\partial(\boldsymbol{\alpha}_P, \boldsymbol{\alpha}_Q)} \right]', \tag{17}$$

where, by the chain rule of differentiation,

$$\frac{\partial(\boldsymbol{r}_S, \boldsymbol{s}_S)}{\partial(\boldsymbol{\alpha}_P, \boldsymbol{\alpha}_Q)} = \frac{\partial(\boldsymbol{r}_S, \boldsymbol{s}_S)}{\partial(\boldsymbol{r}_P, \boldsymbol{r}_Q, \boldsymbol{s}_P, \boldsymbol{s}_Q)} \frac{\partial(\boldsymbol{r}_P, \boldsymbol{r}_Q, \boldsymbol{s}_P, \boldsymbol{s}_Q)}{(\partial\boldsymbol{\alpha}_X, \boldsymbol{\alpha}_Y, \beta_A, \beta_B)} \frac{\partial(\boldsymbol{\alpha}_X, \boldsymbol{\alpha}_Y, \beta_A, \beta_B)}{\partial(\boldsymbol{\alpha}_P, \boldsymbol{\alpha}_Q)}. \tag{18}$$

The partial derivative matrices in Equation 18 are given in Appendix A.

**Local IRT Observed-Score Kernel Equating.** A special case of IRT observed-score equating is the so called IRT local equating, where an equating is conducted conditional on a given ability level (van der Linden, 2011; Wiberg et al., 2014). Hence the score probabilities for the ability level $\theta_0$ used in local equating are given by $r_i(\theta_0)$, $s_j(\theta_0)$, $t_{A_P,k}(\theta_0)$ and $t_{A_Q,k}(\theta_0)$ for the NEAT CE design and by

$$r_{Si}(\theta_0) = w_S r_{Pi}(\theta_0) + (1 - w_S) r_{Pi}(\beta_A \theta_0 + \beta_B)$$

and

$$s_{Sj}(\theta_0) = w_S s_{Qj}(\theta_0) + (1 - w_S) s_{Qj}(\beta_A \theta_0 + \beta_B)$$

for the NEAT PSE design. The asymptotic results previously stated apply for this special case as well.

## IRT Observed-Score Equating in the Kernel Equating Framework

The score probabilities defined for the various equating designs are used in the kernel equating framework by calculating the continuous distributions needed for the specific design and then calculating the equating function. Hence the kernel equating function from $X$ to $Y$ for all score values is the vector-valued function

$$\mathbf{e}_{\mathrm{Y(D)}}(\mathbf{x}) = \Big( \mathrm{e}_{\mathrm{Y(D)}}(0) \quad \mathrm{e}_{\mathrm{Y(D)}}(1) \quad \ldots \quad \mathrm{e}_{\mathrm{Y(D)}}(k_X) \Big)', \tag{19}$$

where $\mathrm{e}_{\mathrm{Y(D)}}(\cdot)$ is the equating function for a specific design D which is replaced with the various designs available. Herein we focus on the two NEAT designs given in Equation 5 and Equation 6.

The function in Equation 19 is continuous and differentiable with respect to the score value probabilities $r_i$ and $s_j$. Hence, if the estimator of the vector of score probabilities $(\mathbf{r}, \mathbf{s})$ is asymptotically normal distributed with covariance matrix $\mathbf{\Sigma}_{\hat{\mathbf{r}},\hat{\mathbf{s}}}$ then the estimator of $\mathbf{e}_{\mathrm{Y(D)}}(\mathbf{x})$ is also asymptotically normal distributed with covariance matrix (Ferguson, 1996)

$$\mathbf{\Sigma}_{\hat{\mathbf{e}}_{\mathrm{Y(D)}}(\mathbf{x})} = \frac{\partial \mathbf{e}_{\mathrm{Y(D)}}(\mathbf{x})}{\partial(\mathbf{r}, \mathbf{s})} \mathbf{\Sigma}_{\hat{\mathbf{r}},\hat{\mathbf{s}}} \left( \frac{\partial \mathbf{e}_{\mathrm{Y(D)}}(\mathbf{x})}{\partial(\mathbf{r}, \mathbf{s})} \right)', \tag{20}$$

where $\frac{\partial \mathbf{e}_{\mathrm{Y(D)}}(\mathbf{x})}{\partial(\mathbf{r},\mathbf{s})}$ is a matrix of partial derivatives with rows which in the notation of von Davier et al. (2004) are denoted $\mathbf{J}_{e_Y}$. Hence, combining the result in either Equation 8 or Equation 17 with Equation 20 the asymptotic covariance matrix of the equating function for the NEAT CE or NEAT PSE design is given.

## Numerical Results and Applications

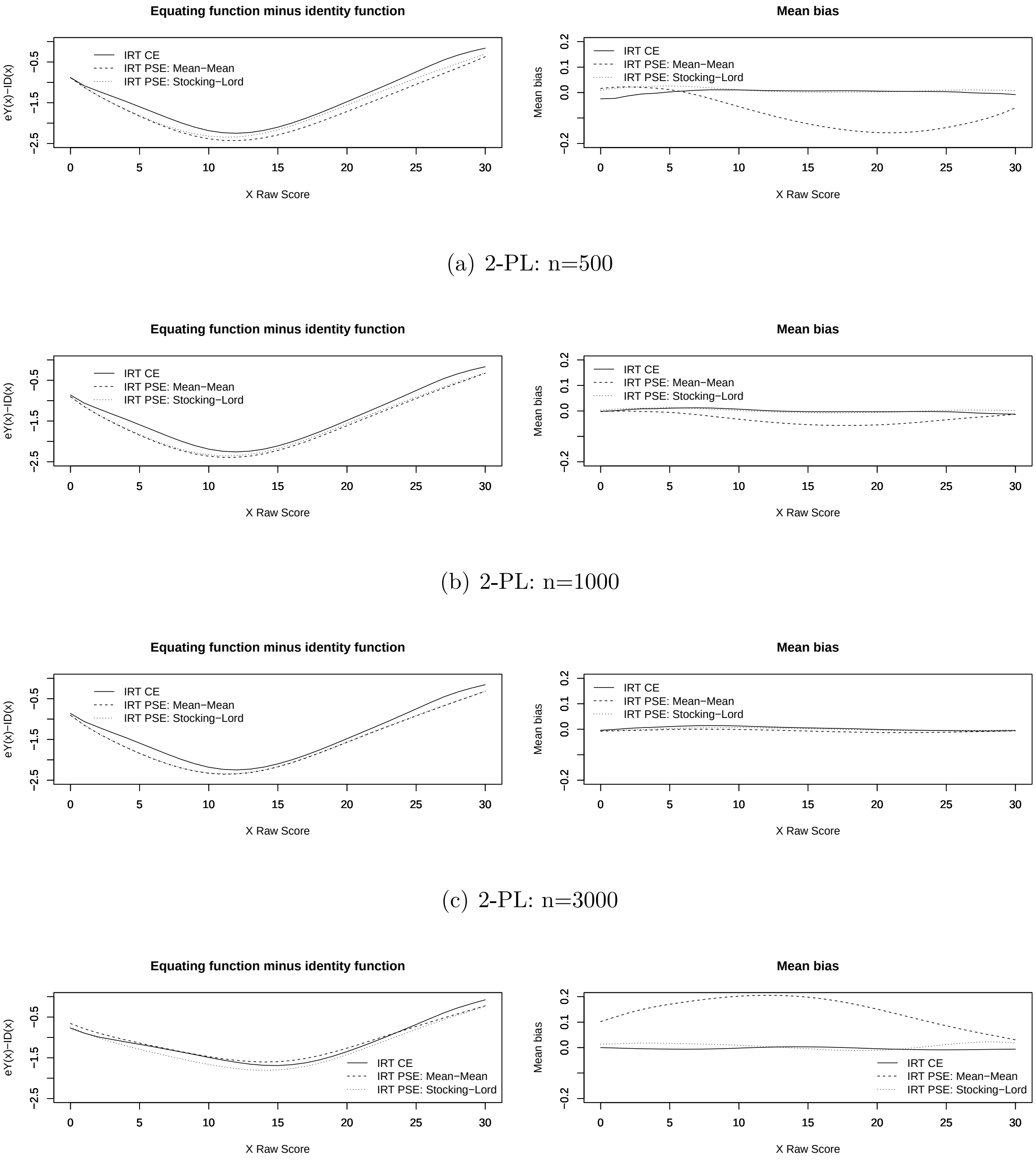
The computations were done using R (R Development Core Team, 2013). The IRT models were estimated using the package **ltm** (Rizopoulos, 2006), the equating coefficients and their partial derivative vectors were calculated with the package **equateIRT** (Battauz, 2013) and IRT observed-score kernel equating was done with the package **kequate** (Andersson, Bränberg, & Wiberg, 2013).

### Simulation Study with the 2-PL and 3-PL Models

To verify the provided derivations and offer a comparison between the 2-PL and 3-PL models simulations were conducted. Data was simulated for a 30 item main test in accordance with the 2-PL and 3-PL model specification for two test groups from populations $P$ and $Q$ respectively. An external anchor test consisting of 20 items was also simulated. For each item the discrimination parameter was randomly drawn from $\sim U(0.5, 2)$, the difficulty parameter from $\sim N(0, 1)$ and (for the 3-PL model) the guessing parameter from $\sim U(0.15, 0.20)$. The same discrimination and difficulty parameters were used for the 2-PL and 3-PL models. The parameters were selected in this manner in order to have a setting which closely resembles that of a real test and to facilitate comparisons with Ogasawara (2003), where the parameters were drawn similarly. The abilities of the test takers were drawn from $\sim N(0, 1)$ for the group from population $P$ and from $\sim N(0.5, 1.2)$ for the group from population $Q$. Such abilities result in the population equating coefficients $\beta_A = 1.2$ and $\beta_B = 0.5$. For the 2-PL model sample sizes $n = 500, 1000, 3000$ were used. Due to the low convergence rate for the 3-PL model with the smaller sample sizes only $n = 3000$ was used with the 3-PL model. The non-convergence rate with the 3-PL model was 31% while it was less than 0.2% for each of the 2-PL models. The results are based on 1000 replications where all the models properly converged.

The equating functions relative to the identity function and the mean biases of the equating functions are plotted in Figure 1. There are only small differences between the equating functions. In the 2-PL equating the PSE equatings are virtually identical. There are some small differences between the 3-PL PSE equatings as displayed in Figure 1(d). None of the equating functions are ever more apart than the difference that matters, defined as ±0.5 raw score points (Dorans & Feigenbaum, 1994). The mean biases are very low except for the PSE Mean-Mean method with sample size 500 using the 2-PL model and when using the 3-PL model, seen in Figures 1(a) and 1(d).

The simulated and analytical standard errors of equating (SEEs) are plotted in Figure 2. Note the difference in scale of Figure 2(a) compared to Figures 2(b)-2(d). Overall, the SEEs using CE are the smallest and the SEEs using PSE Mean-Mean are the largest. There is not a substantial difference in the SEEs between CE and PSE Stocking-Lord. The differences between the PSE Mean-Mean and PSE Stocking-Lord methods are in line with the results in Ogasawara (2001). The simulated and analytical SEEs are very closely matched for all equatings except for the PSE Mean-Mean method with n=500 for the 2-PL model and the 3-PL model, where the analytical SEEs are larger



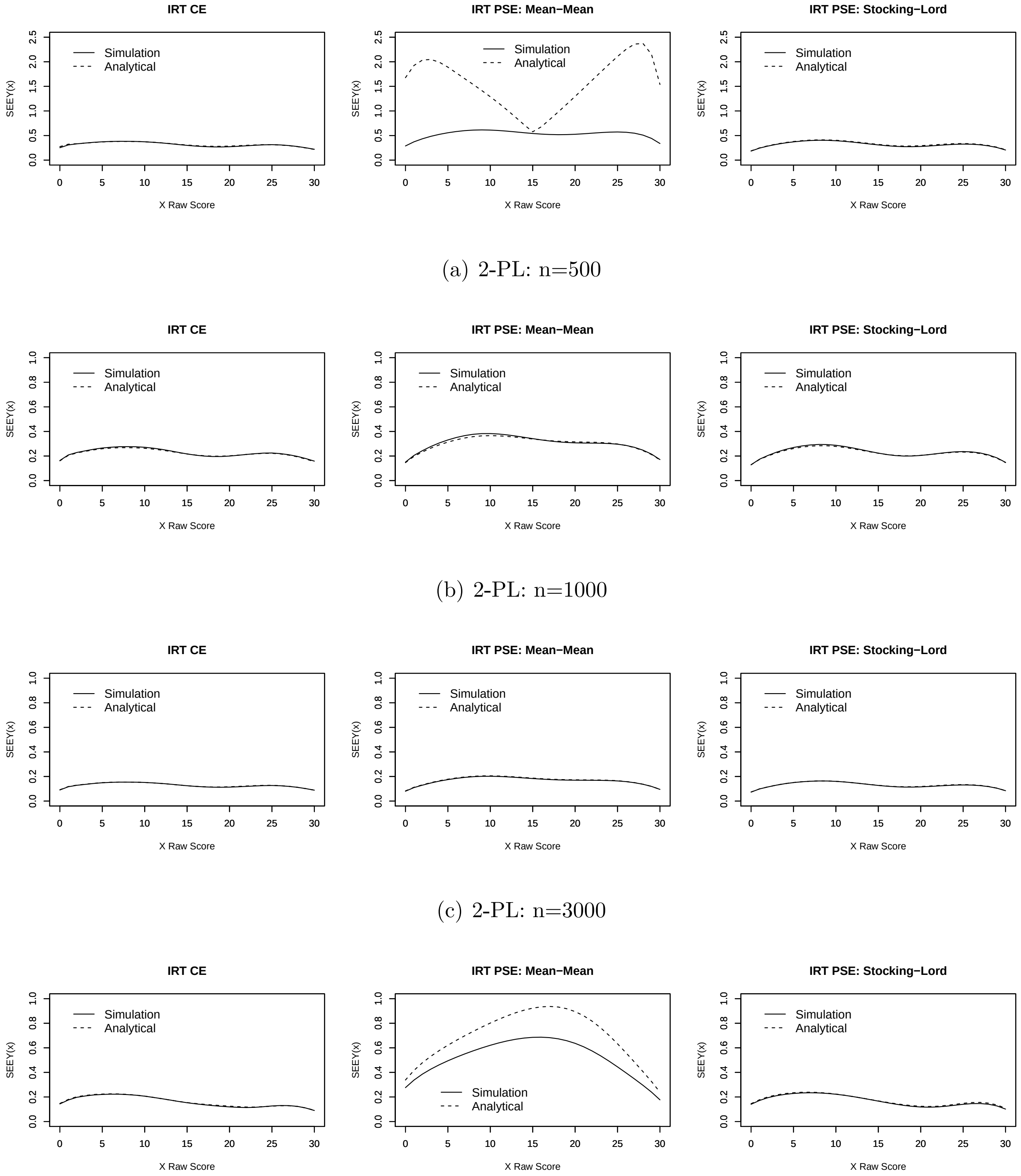
(a) 2-PL: n=500

(b) 2-PL: n=1000

(c) 2-PL: n=3000

(d) 3-PL: n=3000

*Figure 1*. The 2-PL and 3-PL equating functions minus the identity function and equating functions mean biases.



(a) 2-PL: n=500

(b) 2-PL: n=1000

(c) 2-PL: n=3000

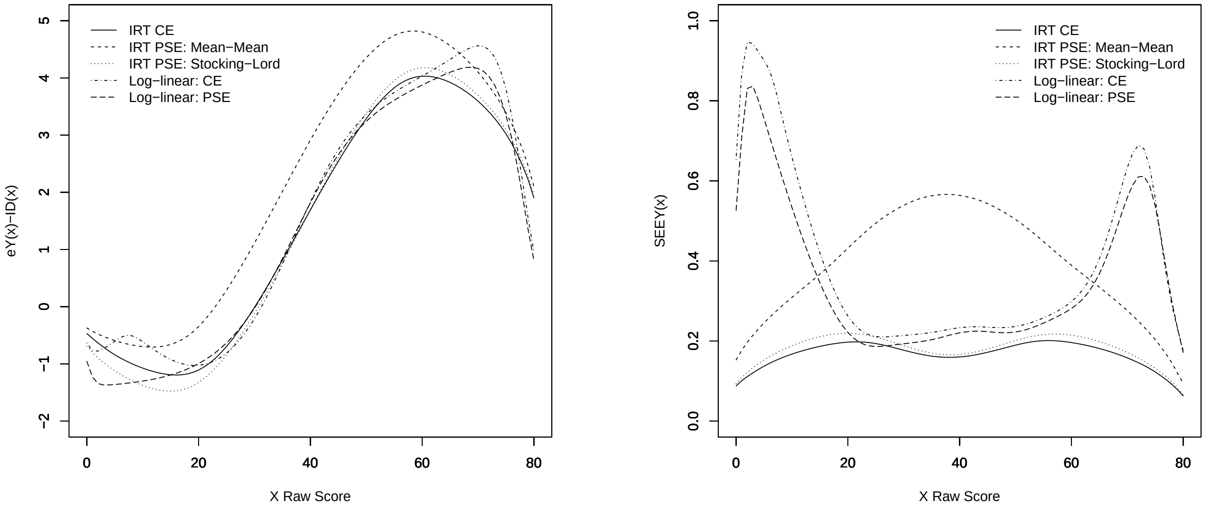
(d) 3-PL: n=3000

*Figure 2*. Simulated and analytical SEEs for the 2-PL and 3-PL equating functions.

than the simulation-based SEEs (Figure 2(d)). As expected, the SEEs for the 2-PL equating are lower than those for the 3-PL equating. With the CE and PSE Stocking-Lord methods the SEEs using the 3-PL model are on average 25% and 34% higher, respectively.

### An Application to a Standardized Achievement Test

To provide an example of the presented method as applied to a real test setting, data from a standardized achievement test was used. The main tests $X$ and $Y$ had 80 items and a 40-item anchor test $A$ was also administered. All items were binary scored multiple choice items. The sample sizes for the two groups were 6465 and 5263, respectively. Only the 2-PL model was fitted since the 3-PL model did not properly converge. Five different equating functions were considered. IRT observed score kernel equating was conducted using both CE and PSE with the Mean-Mean and Stocking-Lord equating coefficients. As a comparison, bivariate log-linear models were also fitted using five univariate moments and two bivariate moments for each test. Then kernel equating using log-linear models in the NEAT CE and PSE designs was applied. The equating functions and their respective SEEs were calculated. As displayed in Figure 3, the equating function and SEE are quite different for many of the designs. The IRT equatings share the same shape of the equating function but the IRT PSE using the Mean-Mean equating coefficients is consistently above the other two IRT equatings as seen in Figure 3(a). The equating functions for IRT CE and IRT PSE with Stocking-Lord are virtually identical, except for score values below 20 where IRT CE is consistently above. The two log-linear equatings have almost identical equating functions to IRT CE and IRT PSE with the Stocking-Lord method in the middle range of the equated scores but differ substantially for lower and higher scores. The SEEs shown in Figure 3(b) indicate that the log-linear equatings are associated with a high sampling variability for scores below 20 and above 60. The SEEs for IRT observed-score equating using CE and PSE with the Stocking-Lord method are almost identical across all score values. IRT observed-score equating using PSE with the Mean-Mean equating coefficients



(a) Equating function.

(b) SEE.

*Figure 3*. The equating function minus the identity function and SEEs for the five designs used for the large scale assessment test.

have higher SEEs than the other IRT methods. Overall, the SEEs are the lowest for the equating functions using IRT CE and IRT PSE with Stocking-Lord equating coefficients.

## Concluding remarks

The utility of IRT observed-score equating as applied to the kernel equating framework has been demonstrated. The advantages of the kernel equating framework are the generality provided by the properties of the kernel used and the flexibility given by the usage of different kernels. Particularly, when using a suitable kernel, the equating function for the kernel method does not have points of non-differentiability which is an issue for an equating function utilizing linear interpolation. Here, only the Gaussian kernel was used but any kernel which is continuous and differentiable can be utilized with the asymptotic results intact. The multivariate treatment of the equating function offered in this paper extends the results of the linear interpolation case given in Ogasawara (2003) and also extends the hypothesis testing methods provided in Rijmen et al. (2011), since hypotheses about the equating function across all score points simultaneously can be tested.

Observed-score kernel equating using IRT models works well for sample sizes as low as 500, provided that the 2-PL model is used. For even smaller sample sizes, equating is generally not recommended but log-linear models can instead be considered. When using maximum likelihood estimation, the 3-PL model has poor properties in the small sample case (Kim, 2006; Li & Lissitz, 2004) and should only be used for large samples. However, for large sample sizes, 3000 or higher, it has been shown that the 3-PL model is a viable option. A downside to using the 3-PL models is that the resulting SEEs are on average approximately 25-35% larger than with the comparable 2-PL models. Because of its properties the 3-PL model will typically provide differences in the equating function compared to the 2-PL model in the lower score range. Hence, if the lower score range is of particular interest the 3-PL model should be pursued. However, in practice some of the guessing parameters might need to be set equal to zero in order to achieve successful convergence.

Common to all IRT models considered in this paper is the reliance on rather strict assumptions which may not hold perfectly in practice. These assumptions are stronger than those for equating using log-linear models or equating using the observed frequencies directly. An interesting investigation worthy of future work would be to compare the performance of these less restrictive methods to the IRT observed-score equating case. Preliminary results in this paper indicate that utilizing IRT models offers less sampling variability than the usage of log-linear models does.

## Appendix

### The Partial Derivatives For the Score Probabilities From IRT Models

The $[(k_X+1)+(k_Y+1)] \times [2(k_X+1)+2(k_Y+1)]$ matrix $\frac{\partial(\boldsymbol{r}_S,\boldsymbol{s}_S)}{\partial(\boldsymbol{r}_P,\boldsymbol{r}_Q,\boldsymbol{s}_P,\boldsymbol{s}_Q)}$ is

$$\frac{\partial(\boldsymbol{r}_S,\boldsymbol{s}_S)}{\partial(\boldsymbol{r}_P,\boldsymbol{r}_Q,\boldsymbol{s}_P,\boldsymbol{s}_Q)} = \begin{pmatrix} w_S\mathrm{diag}(\mathbf{1}_{k_X+1}) & (1-w_S)\mathrm{diag}(\mathbf{1}_{k_X+1}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & w_S\mathrm{diag}(\mathbf{1}_{k_Y+1}) & (1-w_S)\mathrm{diag}(\mathbf{1}_{k_Y+1}) \end{pmatrix}. \tag{21}$$

The $[2(k_X+1)+2(k_Y+1)] \times (3k_X+3k_Y+2)$ matrix $\frac{\partial(\boldsymbol{r}_P,\boldsymbol{r}_Q,\boldsymbol{s}_P,\boldsymbol{s}_Q)}{(\partial\boldsymbol{\alpha}_X,\boldsymbol{\alpha}_Y,\beta_A,\beta_B)}$ is

$$\frac{\partial(\boldsymbol{r}_P,\boldsymbol{r}_Q,\boldsymbol{s}_P,\boldsymbol{s}_Q)}{(\partial\boldsymbol{\alpha}_X,\boldsymbol{\alpha}_Y,\beta_A,\beta_B)} = \begin{pmatrix} \frac{\partial\boldsymbol{r}_P}{\partial\boldsymbol{\alpha}_X} & \mathbf{0} & \mathbf{0} \\ \frac{\partial\boldsymbol{r}_Q}{\partial\boldsymbol{\alpha}_X} & \mathbf{0} & \frac{\partial\boldsymbol{r}_Q}{\partial(\beta_A,\beta_B)} \\ \mathbf{0} & \frac{\partial\boldsymbol{s}_P}{\partial\boldsymbol{\alpha}_Y} & \frac{\partial\boldsymbol{s}_P}{\partial(\beta_A,\beta_B)} \\ \mathbf{0} & \frac{\partial\boldsymbol{s}_Q}{\partial\boldsymbol{\alpha}_Y} & \mathbf{0} \end{pmatrix}, \tag{22}$$

where $\frac{\partial\boldsymbol{r}_P}{\partial\boldsymbol{\alpha}_X}$, $\frac{\partial\boldsymbol{s}_Q}{\partial\boldsymbol{\alpha}_Y}$, $\frac{\partial\boldsymbol{r}_Q}{\partial\boldsymbol{\alpha}_X}$, $\frac{\partial\boldsymbol{s}_P}{\partial\boldsymbol{\alpha}_Y}$, $\frac{\partial\boldsymbol{s}_P}{\partial(\beta_A,\beta_B)}$ and $\frac{\partial\boldsymbol{r}_Q}{\partial(\beta_A,\beta_B)}$ are partial derivative matrices with entries given in Ogasawara (2003).

Lastly, the $(3k_X+3k_Y+2) \times [3(k_X+k_A)+3(k_Y+k_A)]$ matrix $\frac{\partial(\boldsymbol{\alpha}_X,\boldsymbol{\alpha}_Y,\beta_A,\beta_B)}{\partial(\boldsymbol{\alpha}_P,\boldsymbol{\alpha}_Q)}$ is

$$\frac{\partial(\boldsymbol{\alpha}_X,\boldsymbol{\alpha}_Y,\beta_A,\beta_B)}{\partial(\boldsymbol{\alpha}_P,\boldsymbol{\alpha}_Q)} = \begin{pmatrix} \mathrm{diag}(\mathbf{1}_{3k_X}) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathrm{diag}(\mathbf{1}_{3k_Y}) & \mathbf{0} \\ \frac{\partial\beta_A}{\partial\boldsymbol{\alpha}_X} & \frac{\partial\beta_A}{\partial\boldsymbol{\alpha}_{A_P}} & \frac{\partial\beta_A}{\partial\boldsymbol{\alpha}_Y} & \frac{\partial\beta_A}{\partial\boldsymbol{\alpha}_{A_Q}} \\ \frac{\partial\beta_B}{\partial\boldsymbol{\alpha}_X} & \frac{\partial\beta_B}{\partial\boldsymbol{\alpha}_{A_P}} & \frac{\partial\beta_B}{\partial\boldsymbol{\alpha}_Y} & \frac{\partial\beta_B}{\partial\boldsymbol{\alpha}_{A_Q}} \end{pmatrix}, \tag{23}$$

where the partial derivative vectors in the last two rows depend on the method of estimating the equating coefficients. See Ogasawara (2000) for equating coefficients using moments and Ogasawara (2001) for equating coefficients using the Haebara and Stocking-Lord methods.

# References

Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software*, *55*(6), 1–25. Retrieved from `http://www.jstatsoft.org/v55/i06/`

Battauz, M. (2013). equateirt: Direct, chain and average equating coefficients with standard errors using irt methods [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=equateIRT` (R package version 1.0-1)

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*(4), 443–459.

Dorans, N., & Feigenbaum, M. (1994). Equating issues engendered by changes to the sat and psat/nmsqt. *Technical issues related to the introduction of the new SAT and PSAT/NMSQT*, 91–122.

Ferguson, T. (1996). *A course in large sample theory.* Chapman & Hall.

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

Hambleton, R. K., & Swaminathan, H. (1984). *Item response theory: Principles and applications* (Vol. 7). Springer.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (Technical Report No. 89-84). Princeton, NJ: Educational Testing Service.

Kim, S. (2006). A comparative study of irt fixed parameter calibration methods. *Journal of Educational Measurement*, *43*(4), 355–381.

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

Li, Y. H., & Lissitz, R. W. (2004). Applications of the analytically derived asymptotic standard errors of item response theory item parameter estimates. *Journal of Educational Measurement*, *41*(2), 85–117.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.*

Hillsdale, NJ: Erlbaum.

Lord, F. M. (1983). Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika*, *48*(3), 425–435.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of irt true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*, 452–461.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 226–233.

Loyd, B. H., & Hoover, H. (1980). Vertical equating using the rasch model. *Journal of Educational Measurement*, *17*(3), 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*(2), 139–160.

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

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

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement*, *25*(1), 53–67.

Ogasawara, H. (2003). Asymptotic standard errors of irt observed-score equating methods. *Psychometrika*, *68*, 193–211.

R Development Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/ (ISBN 3-900051-07-0)

Rijmen, F., Qu, Y., & Davier, A. A. (2011). Hypothesis testing of equating differences in the kernel equating framework. In A. A. Davier (Ed.), *Statistical models for test equating, scaling, and linking* (p. 317-326). Springer New York.

Rizopoulos, D. (2006, 11 20). ltm: An r package for latent variable modeling and item response analysis. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from `http://www.jstatsoft.org/v17/i05`

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied psychological measurement*, *7*(2), 201–210.

van der Linden, W. J. (2011). Local observed-score equating. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking.* Springer.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York: Springer-Verlag.

von Davier, A. A. (2010). Equating observed-scores: The percentile rank, gaussian kernel, and irt observed-score equating methods. In *International meeting of psychometric society.*

Wiberg, M., van der Linden, W. J., & von Davier, A. A. (2014). Local observed-score kernel equating. *Journal of Educational Measurement*, *51*, 57–74.

Yuan, K.-H., Cheng, Y., & Patton, J. (2013). Information matrices and standard errors for mles of item parameters in irt. *Psychometrika*, 1-23. doi: 10.1007/s11336-013-9334-4