Asymptotic Standard Errors for Item Response Theory True Score Equating of Polytomous Items

Cheow Cher, WONG

Singapore Examinations and Assessment Board

Paper for the National Council on Measurement in Education Annual Conference

April, 2014. Philadelphia, Pennsylvania, April 2 - April 6

Abstract

Building on previous works by Lord and Ogasawara for dichotomous items, this paper proposes a derivation for the asymptotic standard errors of true score equating involving polytomous items for non-equivalent groups of examinees.   The proposed formula were validated using concurrent calibration equating and *mean-mean* equating of simulated bootstrap samples.

Introduction

Formula for asymptotic standard errors of Item Response Theory (IRT) true score equating for common item non-equivalent groups were first derived by Lord (1982) for the three parameter logistic (3PL) model intended for dichotomous items. Using the same 3PL model, Ogasawara (2000, 2001a) presented and derived asymptotic standard errors formulae for some of the common equating methods. These formula could be used in placed of other methods of estimating standard of errors like bootstrapping methods. Building on their approaches, this paper proposes a derivation for standard errors of IRT true score equating using Master's (1982) Partial Credit Model (PCM) and Muraki's (1992) Generalised Partial Credit Model (GPCM), for non-equivalent groups of examinees. The approach is intended for item parameters estimated using the marginal maximum likelihood (MMLE) method, where ability distributions are assumed to be known.

This study is motivated by the fact that few studies of asymptotic standard errors for polytomous exists. Many non-analytic equating standard errors studies have been conducted using the common-item non-equivalent group design, mostly involving dichotomous items. A few studies involving polytomous items make use of random group design (Harris, Welch & Wang, 1994) or bootstrapping methods. In the book by Kolen and Brennan (Kolen and Brennan, 2004, page 250), it was noted that:

> "Computer subroutines for calculating standard errors of some IRT equating methods are available from Ogasawara (2003b). Also, standard errors of equating have not been derived for polytomous IRT models. Additional empirical works is needed to assess the accuracy of the IRT standard errors that have been derived."

As polytomous items are often used in testing situations, it is important to understand how sampling errors affects the equated scores. Whilst parametric and non-parametric bootstrap methods (Efron and Tibshirani, 1993) exists to estimate these standard errors, it may be time consuming to conduct such studies, which may involve 100-1000 simulated datasets, and performing equating of these datasets over 100-1000 times. Using asymptotic standard error formula, one is able to estimate the standard error from the data used for equating, without the need to do any simulations. Lord (1982) first introduced the concept of asymptotic standard error for true score equating of dichotomous items using the 3PL model to estimate sampling error in the equated score, given a particular ability estimate. The study involved external anchor tests for non-equivalent groups with equating performed via a chain. Ogasawara (2001a), also using the 3PL model, further derived a set of formulae for three types of equating methods. The first type of equating methods involves chained equating where true scores of two tests are equated without using any equating coefficient. The second type of equating methods involves the use of IRT equating coefficients, derived from moments or characteristic curves of common items. These methods include the 'mean-mean' method, the 'mean-standard deviation' method, and the characteristic curve method. Finally, the third type of equating makes use of concurrent calibration of the two tests to be equated.

Instead of the 3PL model used for dichotomous items, this study adapts the derivations of Ogasawara (2000, 2001a) to formulate the corresponding formula for polytomous items, using Master's (1982) Partial Credit Model (PCM) and Muraki's (1992) Generalised Partial Credit Model (GPCM). It should be noted in practice, model's assumptions may be violated to some extent and the outcomes of using a model should be examined for its robustness. The derived formula are then verified using randomly generated data, involving the mean-mean and concurrent calibration equating methods. These two equating methods were selected for a start, as they were relatively easier to program. Also, it makes use of IRT true score in the equating, which is sometimes used in place of number correct score in recommending cuts scores. This is different from another class of equating involving observed scores (Kolen & Brennan, 2004).

Asymptotic Standard Errors for True Score Equating of Polytomous Items Involving Equating Coefficients

Mean-Mean Equating of Tests Modelled Using the Partial Credit Model (PCM)

This section proposes the formula needed to compute the asymptotic standard error for the mean-mean equating of non-equivalent groups of examinees. To simplify the presentation, the asymptotic standard error formulae using the PCM model is derived first, before presenting extensions to the GPCM. To facilitate comparison with Ogasawara's (2001a) paper, similar notations are used in this paper. To reiterate, suppose two groups of examinees, Groups 1 and 2, take tests *U* and *V* as follows:

Examinee Group 1:	Test *U*: (subtest *X*	subtest *R*	N.A.	)
Examinee Group 2:	Test *V*: (N.A.	subtest *R*	subtest *Y*	)

Subtest R comprises the items common to both groups of examinees, and the estimated item parameters of these items are used to equate the two tests. For the mean-mean equating method, Test *U* and Test *V* are calibrated separately in two calibration runs, giving rise to two sets of item parameter estimates.

To simplify the presentation, let us consider that all the subtests involve only polytomous items with three categories of number correct scores (*t*=0,1 or 2), assuming that category 1 is assigned a score of 0, category 2 is assigned a score of 1 and so on. This means that each item has 3 categories and two item threshold parameters (denoted by b_{kgh}). Suppose an examinee with ability θ attempted the *g*th item of the *k*th subtest. For Master's (1982) PCM, the probability function of a getting a score of *t* is given as follows:

$$P_{kgt}(\theta) = \frac{\exp[\sum_{h=0}^{t}(\theta - b_{kgh})]}{\sum_{t=0}^{2}\exp[\sum_{h=0}^{t}(\theta - b_{kgh})]}. \quad (1)$$

For subtests *X* and *R₁*, the probability function is in the form in (1) but for subtests *R₂* and *Y* the function takes on a slightly different form to cater to the equating coefficient:

$$P_{kgt}(\theta) = \frac{\exp[\sum_{h=0}^{t}(\theta - b_{kgh} - B)]}{\sum_{t=0}^{2}\exp[\sum_{h=0}^{t}(\theta - b_{kgh} - B)]}. \tag{2}$$

Here, *B* is the equating coefficient for the mean-mean method, to put the two tests on the same scale. Using the thresholds for the common items (i.e. subtest R), we have

$$b_{R_1gh} = b_{R_2gh} + B. \tag{3}$$

To perform true score equating, we need the formula for the true scores of both tests. Using usual statistical formula for expected scores, the following equations give the true scores of tests *U* and *V* respectively:

$$\xi = \sum_{g=1}^{n_X}\sum_{t=1}^{2} tP_{Xgt}(\theta) + \sum_{g=1}^{n_{R_1}}\sum_{t=1}^{2} tP_{R_1gt}(\theta) \tag{4}$$

$$\eta = \sum_{g=1}^{n_{R_2}}\sum_{t=1}^{2} tP_{R_2gt}(\theta) + \sum_{g=1}^{n_Y}\sum_{t=1}^{2} tP_{Ygt}(\theta) \tag{5}$$

The terms involving the summation of *t* from 1 to 2 in these formula stem from using the elementary way of computing expected true scores for polytomous items, by summing over the terms for the three possible scores.

To work out the asymptotic standard error of $\hat{\eta}$, the delta method is used. For true score equating involving coefficient, we need the estimates of both the item threshold parameters (i.e. $\underset{\sim}{\hat{\alpha}}$) and the equating coefficients (i.e. $\hat{B}$). The item threshold parameter estimates may be obtained from the calibration program, whilst the equating coefficient is computed from these item parameter estimates using only the common items. Suppose $\underset{\sim}{\hat{\alpha}}$ and $\hat{B}$ are collectively denoted by the vector $\underset{\sim}{\hat{\beta}} = (\underset{\sim}{\hat{\alpha}}', \hat{B})'$, then the asymptotic variance of $\hat{\eta}$ is obtained using the delta method as follows (see Ogasawara, 2001a):

$$a\,\mathrm{var}(\hat{\eta}) = \frac{\partial\eta}{\partial\underset{\sim}{\beta}'}\, a\,\mathrm{cov}(\underset{\sim}{\hat{\beta}})\frac{\partial\eta}{\partial\underset{\sim}{\beta}} \tag{6}$$

The derivations of the terms on the right-hand side (RHS) of (6) are needed. The first task is to derive $\dfrac{\partial\eta}{\partial\underset{\sim}{\alpha}}$, the derivatives of the true score η with respect to the item parameters, across all the tests (i.e. *Y*, *R₂*, *R₁* and *X*). The following derivatives are required which are adapted from the corresponding equations in Ogasawara's (2001a) paper, with additional summations to cater to the categories of the polytomous items. The partial derivatives of η with respect to parameters in subtests *Y* and *R₂* are:

$$\frac{\partial \eta}{\partial \underset{\sim}{\alpha}_{Ygh}} = \sum_{t=1}^{2} t \frac{\partial P_{Ygt}(\theta)}{\partial \underset{\sim}{\alpha}_{Ygh}} \tag{7}$$

$$\frac{\partial \eta}{\partial \underset{\sim}{\alpha}_{R_2gh}} = \sum_{t=1}^{2} t \frac{\partial P_{R_2gt}(\theta)}{\partial \underset{\sim}{\alpha}_{R_2gh}} \tag{8}$$

For partial derivatives of η with respect to parameters in subtest *X*, we have:

$$\frac{\partial \eta}{\partial \underset{\sim}{\alpha}_{Xgh}} = \frac{\partial \eta}{\partial \theta} \frac{\partial \theta}{\partial \underset{\sim}{\alpha}_{Xgh}} \tag{9}$$

with

$$\frac{\partial \eta}{\partial \theta} = \sum_{k \in (R_2, Y)} \sum_{g=1}^{n_k} \sum_{t=1}^{2} t \frac{\partial P_{Ygt}(\theta)}{\partial \theta} \tag{10}$$

and

$$\frac{\partial \theta}{\partial \underset{\sim}{\alpha}_{Xgh}} = \frac{-\sum_{t=1}^{2} t \left[\partial P_{Xgt}(\theta) / \partial \underset{\sim}{\alpha}_{Xgh} \right]}{\sum_{k \in (X, R_1)} \sum_{g=1}^{n_k} \sum_{t=1}^{2} t [\partial P_{kgt}(\theta) / \partial \theta]} \, , \tag{11}$$

where (11) makes use of implicit functions. Similarly for subtest *R₁*,

$$\frac{\partial \eta}{\partial \underset{\sim}{\alpha}_{R_1gh}} = \frac{\partial \eta}{\partial \theta} \frac{\partial \theta}{\partial \underset{\sim}{\alpha}_{R_1gh}} \, , \tag{12}$$

where similar expression as (11) can be derived for $\frac{\partial \theta}{\partial \underset{\sim}{\alpha}_{R_1gh}}$ on the RHS of (12)

Aside from the partial derivatives of η with respect to the item parameters, we also need the partial derivative with respect to the equating coefficient *B* used in mean-mean equating. The derivative needed is:

$$\frac{\partial \eta}{\partial B} = \sum_{k \in (R_2, Y)} \sum_{g=1}^{n_k} \sum_{t=1}^{2} t \frac{\partial P_{kgt}(\theta)}{\partial B} \tag{13}$$

To work out equations (7)-(12), we could inspect the terms on the right hand side of these equations. It can be seen two groups of partial derivatives are needed - the derivative of P_{kgt} with respect to θ (e.g. $\frac{\partial P_{Ygt}(\theta)}{\partial \theta}$) and the derivative of P_{kgt} with respect to $\underset{\sim}{\alpha}_{kgh}$ (e.g. $\frac{\partial P_{Ygt}(\theta)}{\partial \underset{\sim}{\alpha}_{Ygh}}$). These could be derived using the quotient rule for differentiation, as the probability functions in (1) and (2) are expressed in quotient form. This will be illustrated in the next few paragraphs.

Let us consider subtests *X* and *R₁*. To simply the presentation, we introduce shorthand notations $e1 = \exp(\theta - b_{kg1})$ and $e2 = \exp(\theta - b_{kg2})$. Then,

$$\text{for t=0, } P_{kg0}(\theta) = \frac{1}{1+e1+e1e2}, \tag{14a}$$

$$\text{for } t\text{=1, } P_{kg1}(\theta) = \frac{e1}{1+e1+e1e2}, \tag{14b}$$

$$\text{for } t\text{=2, } P_{kg2}(\theta) = \frac{e1e2}{1+e1+e1e2}. \tag{14c}$$

We need not be concerned with derivatives for *t*=0 as they do not contribute to the computation of true scores. Using quotient rule for differentiation, the following derivatives are derived:

$$\text{For } t\text{=1, } \frac{\partial P_{kg1}(\theta)}{\partial \theta} = \frac{e1(1-e1e2)}{(1+e1+e1e2)^2}, \ \frac{\partial P_{kg1}(\theta)}{\partial b_{kg1}} = \frac{-e1}{(1+e1+e1e2)^2}, \frac{\partial P_{kg1}(\theta)}{\partial b_{kg2}} = \frac{e1^2 e2}{(1+e1+e1e2)^2}. \tag{15a-c}$$

$$\text{For } t\text{=2, } \frac{\partial P_{kg2}(\theta)}{\partial \theta} = \frac{e1e2(2+e1)}{(1+e1+e1e2)^2}, \frac{\partial P_{kg2}(\theta)}{\partial b_{kg1}} = \frac{-e1e2}{(1+e1+e1e2)^2}, \frac{\partial P_{kg2}(\theta)}{\partial b_{kg2}} = \frac{-e1e2(1+e1)}{(1+e1+e1e2)^2}. \tag{16a-c}$$

It is reiterated that these partial derivative are intended for our set-up of using three categories for the polytomous items, hence there are two partial derivatives required for the two threshold parameters (i.e. b_{kg1} and b_{kg2}). The partial derivatives for adjacent scores are related in a sequential manner. For instance, equations (16) can be derived by making use of results derived in (15), by using the product rule for differentiation and the relationship between (14b) and (14c), where $P_{kg2}(\theta) = (e2)P_{kg1}(\theta)$.

Using elementary calculus, we can see that these equations (i.e. (15a-c) and (16a-c)) hold for subtests *Y* and *R₂* too, where the shorthand notations become $e1 = \exp(\theta - b_{kg1} - B)$ and $e2 = \exp(\theta - b_{kg2} - B)$.

We now turn to elaborating the partial derivatives for the equating coefficient *B*, by inspecting the terms on the RHS of equation (13). No extra derivation work is needed, as we note the following from studying equation (2):

$$\frac{\partial P_{kg1}(\theta)}{\partial B} = -\frac{\partial P_{kg1}(\theta)}{\partial \theta} \text{ and } \frac{\partial P_{kg2}(\theta)}{\partial B} = -\frac{\partial P_{kg2}(\theta)}{\partial \theta}, \tag{17}$$

where the terms $\frac{P_{kg1}(\theta)}{\partial \theta}$ and $\frac{P_{kg2}(\theta)}{\partial \theta}$ are already derived in (15a) and (16a).

We have already dealt with the term $\dfrac{\partial \eta}{\partial \underset{\sim}{\beta}}$ on the RHS of (6), with the derivation of the remaining $a\,\mathrm{cov}(\hat{\beta})$ elaborated here. The next few equations are adapted from Ogasawara's (2001a) derivations and are presented here for completeness. Now since $\hat{\beta} = (\hat{\underset{\sim}{\alpha}}', \hat{B})'$, the covariance of $\hat{\beta}$ takes the form of a partitioned matrix:

$$a\,\mathrm{cov}(\hat{\beta}) = \begin{pmatrix} a\,\mathrm{cov}(\hat{\underset{\sim}{\alpha}}) & a\,\mathrm{cov}(\hat{\underset{\sim}{\alpha}}; \hat{B}) \\ a\,\mathrm{cov}(\hat{B}; \hat{\underset{\sim}{\alpha}}') & a\,\mathrm{cov}(\hat{B}) \end{pmatrix} \tag{18}$$

where
$$a\,\mathrm{cov}(\hat{B}; \hat{\underset{\sim}{\alpha}}') = \frac{\partial B}{\partial \underset{\sim}{\alpha}'} a\,\mathrm{cov}(\hat{\underset{\sim}{\alpha}}) \tag{19}$$

and
$$a\,\mathrm{cov}(\hat{B}) = \frac{\partial B}{\partial \underset{\sim}{\alpha}'} a\,\mathrm{cov}(\hat{\underset{\sim}{\alpha}}) \frac{\partial B}{\partial \underset{\sim}{\alpha}}. \tag{20}$$

Note that equation (20) also makes use of the delta method, as the coefficient *B* is a function of $\hat{\underset{\sim}{\alpha}}$ in the mean-mean equating method.

Inspecting the RHS of (19) and (20), we note that the term $\dfrac{\partial B}{\partial \underset{\sim}{\alpha}}$ is required. For the mean-mean method, the non-zero derivatives for the *p* common items are:

$$\frac{\partial B}{\partial b_{R_1 j1}} = \frac{\partial B}{\partial b_{R_1 j2}} = \frac{1}{2p} \text{ and } \frac{\partial B}{\partial b_{R_2 j1}} = \frac{\partial B}{\partial b_{R_2 j2}} = -\frac{1}{2p} \quad (j = 1,.., p). \tag{21}$$

Compared to Ogasawara's (2000) equations, an additional number '2' in the denominator is present, which corresponds to our assumption of two thresholds in each polytomous item. With that derivation, we have all the necessary equations to work out the asymptotic standard error for the true score equating of items, using PCM and mean-mean equating. Note that whilst the above derivations assume three categories in each polytomous item, it can easily be generalised to items with more categories, or a test with a mix of items with different number of categories. The derivatives have to be adjusted using elementary calculus, depending on the number of categories in the items.

Extension to the Concurrent Calibration Equating

In the case of concurrent calibration equating, the formula are simpler, as there is no need to have the equating coefficient B. In concurrent calibration equating, all item parameter estimates are on the same scale. Unlike mean-mean equating which requires two separate calibrations, both tests *U* and *V* are calibrated together in a single run in concurrent calibration equating, after aligning the common items in the data matrix. Hence all the derivations in the previous section involving B could be set to zero.

Extension to the Generalised Partial Credit Model (GPCM)

The derivations in the preceding sections can be extended to equating involving Muraki's (1992) GPCM. Compared to the PCM model, it includes the additional item discrimination parameters a_{kg}. The probability functions are:

$$P_{kgt}(\theta) = \frac{\exp[\sum_{h=0}^{t} a_{kg}(\theta - b_{kgh})]}{\sum_{t=0}^{2}\exp[\sum_{h=0}^{t} a_{kg}(\theta - b_{kgh})]} \qquad \text{for subtests } X \text{ and } R_1, \qquad (22)$$

and

$$P_{kgt}(\theta) = \frac{\exp[\sum_{h=0}^{t} (a_{kg}/A)(\theta - Ab_{kgh} - B)]}{\sum_{t=0}^{2}\exp[\sum_{h=0}^{t} (a_{kg}/A)(\theta - Ab_{kgh} - B)]} \quad \text{for subtests } R_2 \text{ and } Y. \qquad (23)$$

Compared to equation (3), there are now two equating coefficients for the mean-mean method, namely A and B where:

$$(a_{R_1gh}, b_{R_1gh}) = (a_{R_2gh}/A, Ab_{R_2gh} + B) \qquad (24)$$

Also, the shorthand notations in equations (14) are updated to $e1 = \exp(a_{kg}(\theta - b_{kg1}))$ for subtests *X* and *R₁*, and $e1 = \exp(\frac{a_{kg}}{A}(\theta - Ab_{kg1} - B))$ for subtests *Y* and *R₂* and so on. For partial derivatives $\frac{\partial P_{kg1}(\theta)}{\partial \theta}$ and $\frac{\partial P_{kg2}(\theta)}{\partial \theta}$ corresponding to those in equations (15a) and (16a), an additional multiplier of a_{kg} (for subtests *X* and *R₁*) or $\frac{a_{kg}}{A}$ (for subtests *Y* and *R₂*) is needed on the RHS of these equations, following rule for differentiation of exponential terms. Similarly for partial derivatives $\frac{\partial P_{kg1}(\theta)}{\partial b_{kg1}}$, $\frac{\partial P_{kg1}(\theta)}{\partial b_{kg2}}$, $\frac{\partial P_{kg2}(\theta)}{\partial b_{kg1}}$ and $\frac{\partial P_{kg2}(\theta)}{\partial b_{kg2}}$ corresponding to those in equations (15b-c) and (16b-c), the additional multiplier of a_{kg} is needed for all subtests (i.e. *X*, *R₁*, *Y* and *R₂*).

Unlike PCM, the partial derivatives involving a_{kg} for subtests *X* and *R₁* would also be needed:

$$\text{For } t\text{=1,} \quad \frac{\partial P_{kg1}(\theta)}{\partial a_{kg1}} = \frac{e1(\theta - b_{kg1}) - e1^2 e2(\theta - b_{kg2})}{(1 + e1 + e1e2)^2} \qquad (25)$$

$$\text{For } t\text{=2,} \quad \frac{\partial P_{kg2}(\theta)}{\partial a_{kg2}} = \frac{e1e2((1 + e1)(\theta - b_{kg2}) + \theta - b_{kg1})}{(1 + e1 + e1e2)^2} \qquad (26)$$

The corresponding equations for subtests *Y* or *R*$_2$ are:

For t=1, $$\frac{\partial P_{kg1}(\theta)}{\partial a_{kg1}} = \frac{e1(\theta - Ab_{kg1} - B) - e1^2 e2(\theta - Ab_{kg2} - B)}{A(1+e1+e1e2)^2} \tag{27}$$

For t=2, $$\frac{\partial P_{kg2}(\theta)}{\partial a_{kg2}} = \frac{e1e2\left[(1+e1)(\theta - Ab_{kg2} - B) + (\theta - Ab_{kg1} - B)\right]}{A(1+e1+e1e2)^2} \tag{28}$$

Finally, in addition to (17), the additional partial derivatives with respect to *A* would be:

$$\frac{\partial P_{kg1}(\theta)}{\partial A} = -\frac{(\theta - B)}{A}\frac{\partial P_{kg1}(\theta)}{\partial \theta} \text{ and } \frac{\partial P_{kg2}(\theta)}{\partial A} = -\frac{(\theta - B)}{A}\frac{\partial P_{kg2}(\theta)}{\partial \theta} \tag{29}$$

With slight adaptation to cater to polytomous items, Ogasawara’s (2000) derivations for mean-mean equating involving both *A* and *B* coefficients would also be needed.

Verification of Formula Using Simulated Samples

Method

To verify the formula, simulation studies using samples derived from different polytomous models (i.e. PCM and GPCM), different equating methods (i.e. mean-mean and concurrent) and different sets of population item parameters were conducted. As mentioned before, the two equating methods selected were the concurrent calibration equating and the mean-mean equating, as these methods are less complex to program. The use of different sets of population item parameters (see Appendix A) was intended as additional checks to verify the formula. Four possible combinations of studies were conducted – namely *PCM & Mean-Mean Equating*, *PCM & Concurrent Equating*, *GPCM & Mean-Mean Equating* and *GPCM & Concurrent Equating*. A total of eight studies were conducted. The item parameters for the first study in each combination were generated from the Normal(-0.5,1) and Normal(0.5,1) for the two threshold parameters, and the discrimination parameters (for GPCM studies only) were generated by adding 0.3 to random values generated from the Uniform distribution. For the second study in each combination, the two thresholds were generated from the Normal(0,1) and Normal(1,1), and the discrimination parameters were generated in the same manner as the first study in each combination.

The steps involved in each study are described as follows. First, using the 26 generated population item parameters, two artificial tests (Tests *U* and *V*) were created, each with 16 three-category items, with six common items between the two tests. Second, using the population item parameters, a total of 200 parametric bootstrap samples were simulated to represent tests *U* and *V*, each with 1000 examinees’ responses. For the *PCM & Mean-Mean Equating* study, the 1000 examinees’ abilities were simulated using random draws of values from the Normal(0,1) and Normal (0.5,1) for tests *U* and *V* respectively. For *GPCM & Mean-Mean Equating*, the Normal(0,1) and Normal(0.5,1.2) were used to simulate the abilities, as GPCM allows for the modelling of the item discrimination parameters. In the case of studies involving the concurrent calibration method, examinees' abilities were all drawn from the Normal(0,1) for both tests *U* and *V*. The response of an examinee to an item was simulated by comparing the segments defined by cumulative score distribution of the polytomous item, with a random draw from the Uniform distribution. In our example of three-category items, the cumulative score distribution divides the [0,1] probability space for a given ability into three segments – P(0 <= *t* < 1), P(1 <= *t* < 2) and P(*t* = 2). Depending on the value of the random draw from the Uniform distribution and the segment it falls on, a score of 0, 1 or 2 was assigned as the response.

Third, each simulated sample was calibrated using the NLMixed procedure (Sheu, Chen & Wang, 2005; Tuerlinckx & Wang, 2004) found in the SAS statistical package, assuming a 15 point quadrature for ability. An example of the subroutine is shown in Appendix B. The item parameters, as well as their variance-covariance matrix, could be produced in the SAS calibration and stored as outputs. Using these outputs, the analytical asymptotic standard errors were derived using the formula.

The empirical standard errors were also computed, which involved conducting 100 equating for each study. For the mean-mean equating method, the estimated parameters of the two tests were placed on the same scale using the difference in the mean threshold values of the six common items (i.e. 12 thresholds). For the concurrent calibration method, as the simulated responses of both tests were calibrated in a single run of NLMixed, the estimated parameters of both tests were already on the same scale. With parameters on the same scale, the equated scores $\hat{\eta}$ in test *V* given true scores in test *U* could then be obtained, using the PCM or GPCM formula. Finally, the empirical standard errors of equating were computed using the standard deviation of the equated scores $\hat{\eta}$, over the 100 pairs of simulated samples.

For each study, the first 10 asymptotic standard errors (denoted by ASE in the figures, i.e. ASE1-ASE10) were computed and plotted against the empirical standard error (denoted by SIM), to evaluate if the ASEs are close to the SIM by inspecting the graphs. In practice, only one ASE graph is possible as there is only one set of real data.

Results

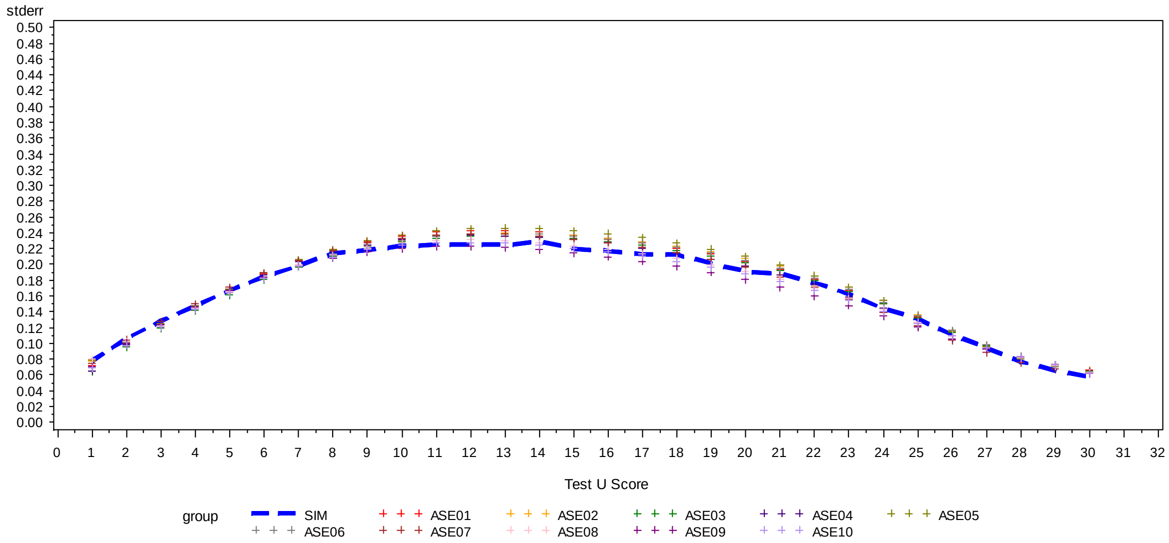
Figures 1a and 1b show the results for the *PCM & Mean-Mean Equating* studies. The means of the equating coefficient *B* were 0.505 and 0.503 respectively, close to the expected value of 0.5. These graphs compare the standard error obtained from the equating of 100 pairs of simulated samples (SIM), and the asymptotic standard error obtained from the first 10 simulated samples (ASE1-ASE10). Figure 1a shows that the curves are close. This is also the case for Figure 1b, where another set of population item parameters was used. This supports the validity of the derived formula.

Figures 2a and 2b show the results for the *PCM & Concurrent Equating* studies, using the same population item parameters as those in Figures 1a and 1b respectively (see Annex A). Figure 2a or 2b shows that the ASE curves are close to those of SIMs, verifying the formula.

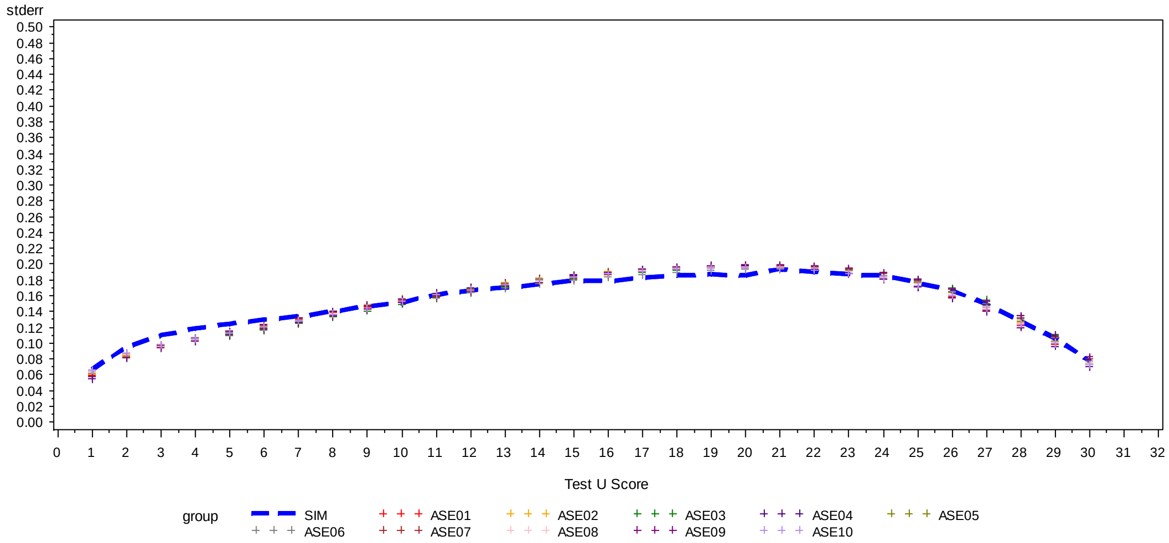
Figures 3a and 3b show the results for the *GPCM & Mean-Mean Equating* studies. The means of the equating coefficients (A, B) were (0.524, 1.216) and (0.508, 1.197) respectively, close to the expected values of (0.5, 1.2). The ASE curves tend to follow the shape of SIMs. However, a greater deviation between ASEs and SIMs is observed here, compared to the corresponding results for PCM (see Figures 1a and 1b). A greater variation amongst the ASEs is also detected. Nonetheless, the similar shapes and closeness of the SIMs and ASE curves lend support to the validity of the formula.

Figures 4a and 4b show the results for the *GPCM & Concurrent Equating* studies. These studies make use of the same item parameters as those in Figures 3a and 3b respectively. The ASE curves are close to the SIM curves. Compared to the *GPCM & Mean-Mean Equating* studies, the *GPCM & Mean-Mean Equating* (see Figures 3a-b) appears to give relatively the worse results.

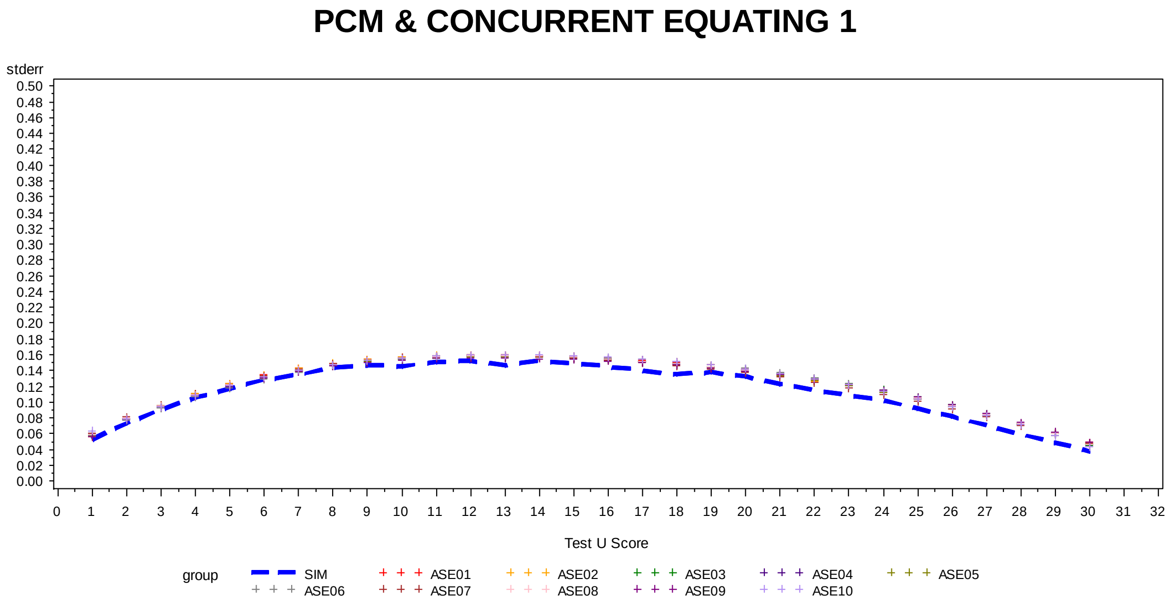
PCM & M-M EQUATING 1

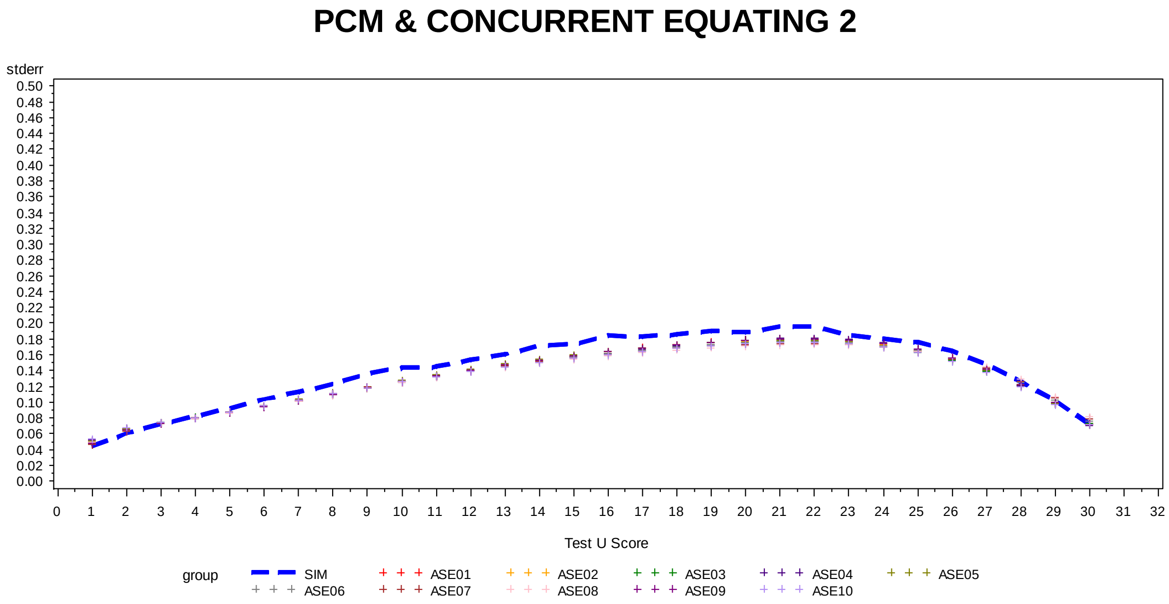


PCM & M-M EQUATING 2

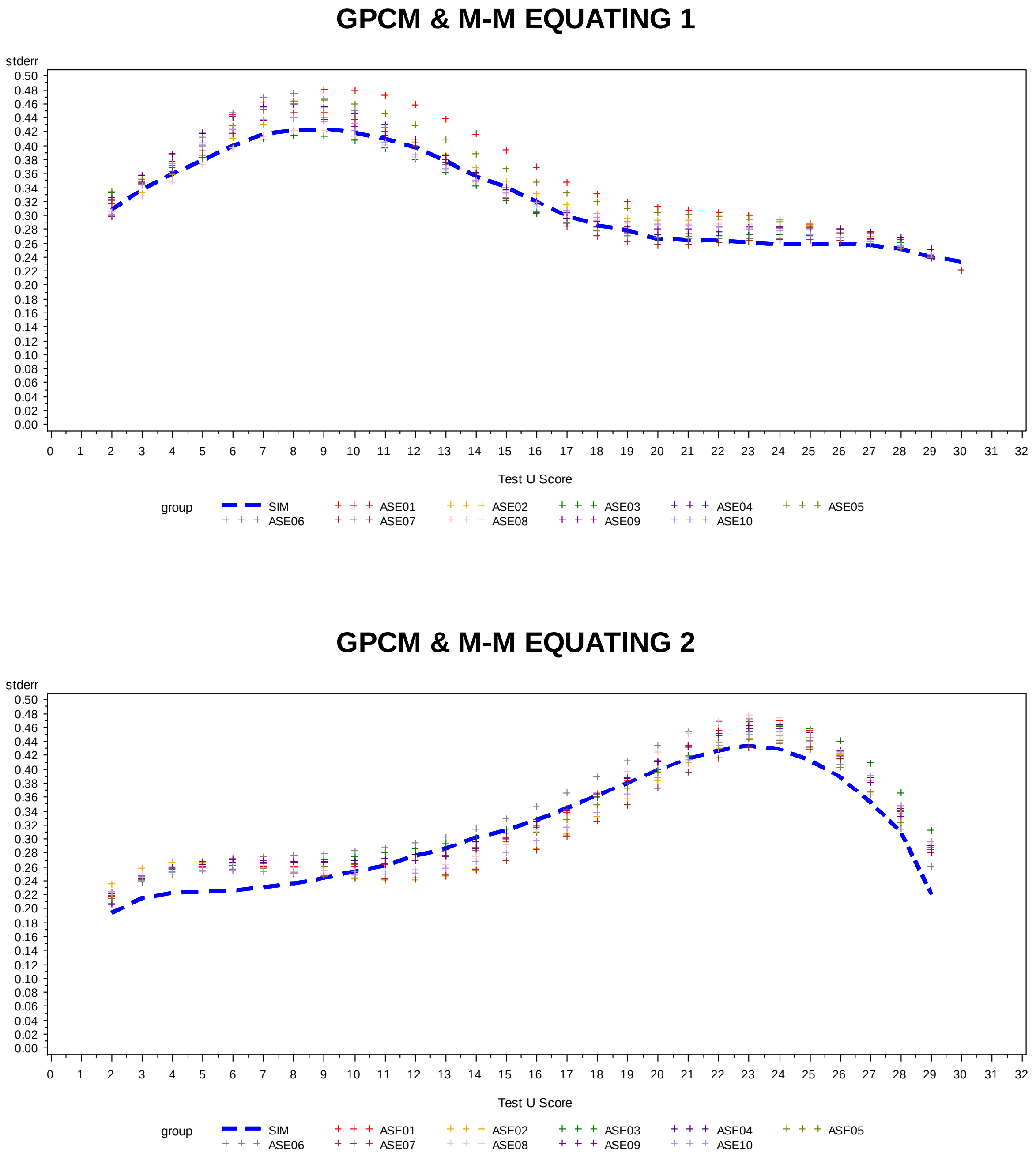


Figures 1a-b: Studies using PCM and Mean-Mean Equating. Asymptotic Standard Errors derived from the first 10 sets of samples (ASE1-ASE10), compared with standard error derived from 200 bootstrap samples (SIM).





Figures 2a-b: Studies using PCM and Concurrent Equating. Asymptotic Standard Errors derived from the first 10 sets of samples (ASE1-ASE10), compared with standard error derived from 200 bootstrap samples (SIM).

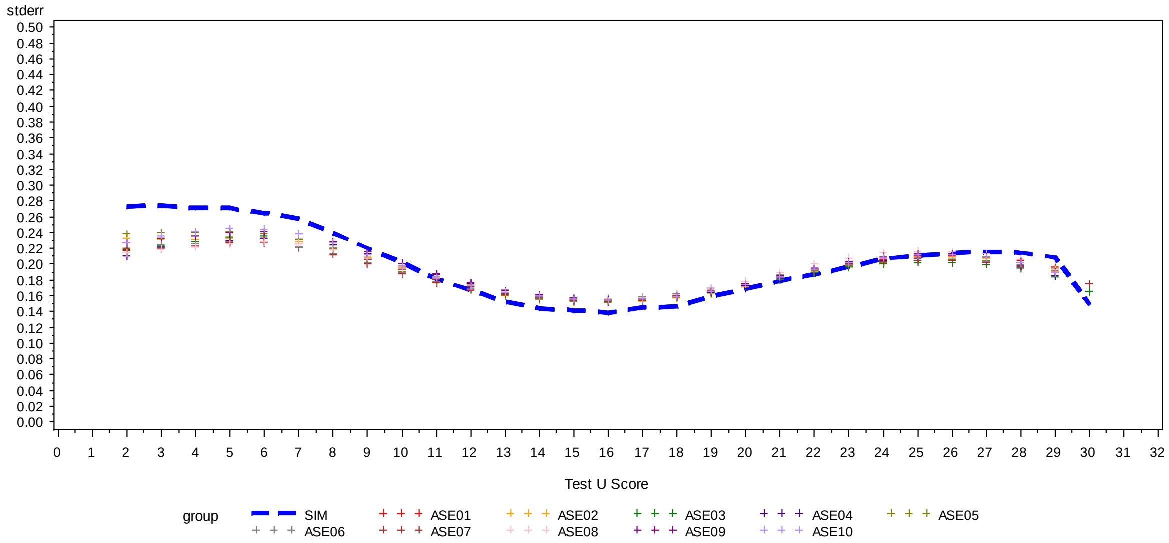


GPCM & M-M EQUATING 1

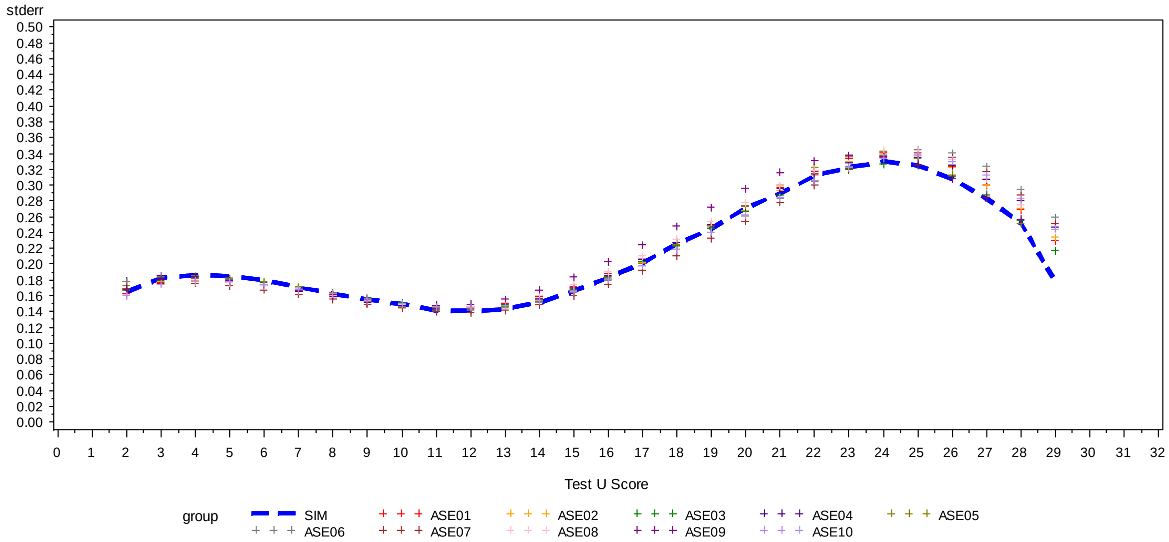
GPCM & M-M EQUATING 2

Figures 3a-b: Studies using GPCM and Mean-Mean Equating. Asymptotic Standard Errors derived from the first 10 sets of samples (ASE1-ASE10), compared with standard error derived from 200 bootstrap samples (SIM).

GPCM & CONCURRENT EQUATING 1



GPCM & CONCURRENT EQUATING 2



Figures 4a-b: Studies using GPCM and Concurrent Equating. Asymptotic Standard Errors derived from the first 10 sets of samples (ASE1-ASE10), compared with standard error derived from 200 bootstrap samples (SIM).

Discussion

The proposed formula to compute the asymptotic standard errors for the GPCM and PCM were derived and verified in this paper. Results are generally comparable between the empirically computed and analytically derived standard errors. This is true for studies using the different models (i.e. GPCM or PCM), different equating methods (i.e. concurrent or mean-mean), and different population item parameters. Amongst the studies in this paper, the *GPCM & Mean-Mean Equating* studies seem to produce relatively the worst results. This could be due to the need to estimate more parameters (i.e. both item parameters and equating coefficients), compared to the other studies.

The studies also demonstrated the possibility of using outputs from commercial software like SAS to compute asymptotic standard error for equating, which may be more accessible for some researchers, as the variance-covariance matrix is produced during the calibration.

There is scope for more studies, to lend more support to the accuracy of the formula. More studies could be conducted using other equating methods (e.g. characteristic curves methods), different population item parameters, or different number of quadrature points, as well as related studies of bias in equating. The proposed solution and formula using different commercial software could also be attempted, to determine if these observations are replicable. Finally, the approach to derive the formula for asymptotic standard errors presented in this paper could be extended to other polytomous item response theory models like Samejima's Graded Response model or Andrich's Rating Scale model.

Appendix A

Population Item Parameters Used to Simulate Examinees' Responses for the Various Studies

1st Study of Each Combination					2nd Study of Each Combination		
GPCM & Mean/Mean Equating 1 **GPCM** & Concurrent Equating 1					**GPCM** & Mean/Mean Equating 2 **GPCM** & Concurrent Equating 2		
	Included	Included	Included		Included	Included	Included
PCM & Mean/Mean Equating 1 **PCM** & Concurrent Equating 1					**PCM** & Mean/Mean Equating 2 **PCM** & Concurrent Equating 2		
	Included	Included	Excluded		Included	Included	Excluded
Test R (common items)	**b1**	**b2**	**a**		**b1**	**b2**	**a**
1	1.33806	3.08258	0.79763		0.5789	1.58495	1.01307
2	0.21331	1.02245	1.29233		-1.68365	1.55287	0.76445
3	-0.62052	0.86036	0.48273		0.58717	0.57203	0.95577
4	0.57923	1.05657	0.53858		0.42444	0.92405	0.54243
5	1.4392	0.29652	0.6463		-0.94425	-0.20737	0.33482
6	0.82241	-2.01982	1.19491		-0.42301	1.18193	0.64467
Test X							
7	-0.24341	1.26142	0.98801		1.04386	0.88357	0.8774
8	0.0251	-1.05536	1.29273		1.0179	1.3931	0.49124
9	-0.81588	0.73243	0.72398		1.13689	1.48594	0.38911
10	0.44013	0.50921	1.16191		-2.36292	0.96759	1.24919
11	-1.3015	0.5786	0.32351		0.17206	0.50831	0.85046
12	0.09979	2.97816	1.12001		1.44349	1.20696	0.83222
13	-0.28374	-1.44705	0.85695		-1.0292	3.46612	1.05332
14	-1.90943	2.09595	0.7641		0.90394	0.81524	0.53393
15	-1.24689	0.0966	0.33429		0.34143	2.17371	0.4313
16	-1.14301	1.19819	1.21465		-0.02732	1.49808	0.73662
Test Y							
17	-0.3626	1.1303	0.42531		1.76616	1.9649	1.27168
18	-0.05881	0.67114	0.37231		-1.15887	1.41843	0.60408
19	-1.62947	1.0397	0.48049		0.07449	1.70189	0.46944
20	-2.96021	0.30297	0.98042		0.89841	-0.21586	0.94881
21	-0.01764	0.29818	0.46142		1.90587	1.36124	0.65042
22	-0.1051	0.21991	0.72959		-2.45609	-0.03249	0.5422
23	-0.72166	-0.45564	1.11316		1.44083	1.13126	0.86995
24	0.57653	0.95833	1.29675		-0.31688	1.2674	0.44457
25	-0.22971	-1.49936	1.1765		-1.23284	0.88399	0.69617
26	-1.84183	1.34844	0.49588		-0.30707	1.76409	1.0011

Appendix B

Using SAS NLMixed for the Generalised Partial Credit Model

```
* Call in simulated dataset 1
DATA ff1;
  SET IN.simdata1;
  CASE=_N_;
  KEEP CASE Q1-Q16;
RUN;

* Import categorical item data;
DATA F1; SET ff1;
ARRAY aQ(16) Q1-Q16;
DO i=1 TO 16;
item=i; Q=aQ(i); OUTPUT;
END;
RUN;

* Create dummy variables;
DATA F1; SET F1;
ARRAY dummy (16) i1-i16;
DO d=1 TO 16;
IF item=d THEN dummy(d)=1; ELSE dummy(d)=0;
END;
DROP i d Q1-Q16;
RUN;

PROC NLMIXED DATA=F1 METHOD=GAUSS TECHNIQUE=QUANEW QPOINTS=15 COV NOAD;

* All model parameters must be listed here with start values;
PARMS d101-d116=-1 d201-d216=0 a01-a16=2;

d1 = d101*i1 + d102*i2 + d103*i3 + d104*i4 + d105*i5 + d106*i6 + d107*i7 + d108*i8 + d109*i9 +
d110*i10 + d111*i11 + d112*i12 + d113*i13 + d114*i14 + d115*i15 + d116*i16;

d2 = d201*i1 + d202*i2 + d203*i3 + d204*i4 + d205*i5 + d206*i6 + d207*i7 + d208*i8 + d209*i9 +
d210*i10 + d211*i11 + d212*i12 + d213*i13 + d214*i14 + d215*i15 + d216*i16;

a = a01*i1 + a02*i2 + a03*i3 + a04*i4 + a05*i5 + a06*i6 + a07*i7 + a08*i8 + a09*i9 + a10*i10 +
a11*i11+ a12*i12 + a13*i13 + a14*i14 + a15*i15 + a16*i16 ;

eta1 = exp((a)*((theta-d1)));
eta2 = exp((a)*((theta-d1)+(theta-d2)));
* Probabilities for each category estimated;
IF Q=0 THEN p = 1 / (1 + eta1 + eta2 );
ELSE IF Q=1 THEN p = eta1 / (1 + eta1 + eta2);
ELSE IF Q=2 THEN p = eta2 / (1 + eta1 + eta2);
ll = log(p);
MODEL Q ~ general(ll);

RANDOM theta ~ normal(0,1) SUBJECT = case ;

* All item parameter estimates and the variance-covariance matrix saved to named datasets;
ODS OUTPUT ParameterEstimates=OUT.item_parameter;
ODS OUTPUT CovMatParmEst=OUT.variance_cov;
RUN;
```

References

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

Efron, B. & Tibshirani, R.J. (1993). *An introduction to the bootstrap (Monographs on Statistics and Applied Probability 57)*. New York: Chapman and Hall.

Harris, D.J., Welch, C.J., and Wang, T. (1994). *Issues in equating performance assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

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

Lord, F. M. (1982). Standard errors of an equating by item response theory. *Applied Psychological Measurement*, 6, 463-472.

Masters,G . N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

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

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

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

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

Ogasawara, H. (2003). *EL 1.0.* Unpublished computer subroutines.

Ogasawara, H., Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, 47, 397-412.

van der Linden, W. J., & Luecht, R. M. (1998). Observed-score equating as a test assembly problem. *Psychometrika*, 63, 401-418.

Zeng, L., & Kolen, M. J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement*,19, 231-241.

Tuerlinckx, F., & Wang, W. C. (2004). Models for polytomous data. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 75-109). New York: Springer.

Sheu C., Chen C., Su Y., Wang W.(2005). Using SAS PROC NLMIXED to fit item response theory models. *Behavior Research Methods*, 37, 202-218.