NLMixed Procedure to Derive the Standard Errors for Partial Credit Tests

Cheow Cher, WONG

Singapore Examinations and Assessment Board

Paper for the Pacific Rim Objective Measurement Symposium 2015

Abstract

Lord (1982) and Ogasawara (2001a) derived formula for deriving the asymptotic standard errors for true score equating using 2PL and 3PL dichotomous items for equivalent and non-equivalent groups. Recently, Wong (2015) proposes extensions to cater to polytomous items, using the Generalised Partial Credit Model (GPCM) and Graded Response Model. In SAS NLMixed, the proposed formula works well for in estimating standard errors for true score equating involving equivalent and non-equivalent groups of examinees.

The Partial Credit Model (PCM) is different from the GPCM, as it assumes a constant slope for the item parameters. The method of equating is also different, with the use of a single shift value, compared to two equating coefficients for the GPCM. Hence, studies in PCM are needed to determine if NLMixed support the use of the formulas. This paper proposes the use of NLMixed to estimate true score equating standard errors for the PCM using formulas, for common item equating and concurrent calibration equating. A simple example with step-by-step explanation will be used to illustrate use of the formulas. Studies were also conducted based on Monte-Carlo simulations and results are presented in this paper.

Introduction

Equating or linking could be conducted for scores in a number of metric, namely the ability scale, true score scale or observed score scale. The true score scale is used in equating, due to familiarity with the scale, as it resembles the "number correct" score for dictomous items. True scores are used as though they are observed score, to form tables of equivalent scores between tests forms. For instance, with equating, we could claim that a score of 35 marks in form U last year, is equivalent to a score of 32 marks in form V this year.

The purpose of asymptotic standard error (SE) is to quantify the amount of error in the equated score, due to sampling. This inform on the accuracy of the critical equated scores (e.g. pass-fail score), and could influence decisions based on these cut-scores. APA, AERA, and NCME (see *Standards for Educational and Psychological Testing*, 1999, p.57) also recommend the reporting of SE in equating.

There are generally two approaches to obtain estimates for the SE. The first approach involves the use of a large number of bootstrap samples that are generated from the estimated item parameters. It could be rather time consuming to perform such studies, which may involve performing 100-1000 equatings using the bootstrap samples, to obtain estimates of SE. The second approach makes use of asymptotic standard error formulas derived using the delta method, to approximate the SE. There is no need to simulate any sample, as the formulas make use of the estimated parameters, as well as the variance-covariance matrix obtained during calibration. Formulas for IRT dictomous models were derived by Ogasawara (2001a) for dicotomous 2PL and 3PL models. Wong (2015) also proposed formulas for polytomous models, namely the Graded Response Model (GRM) and Generalised Partial Credit (GPCM) IRT models. His studies using SAS NLMixed tended to support the use of these formulas.

The purpose of this paper is to propose the procedure to estimate asymptotic SE using NLMixed, for the Partial Credit model (Masters, 1982). The model is different from the GPCM, as it assumes a constant slope across items. In 2004, the use of NLMixed to calibrate IRT models, including the Rasch family of models was proposed by Wilson et al. (Wilson & De Boeck, 2004; Sheu, Chen & Wang, 2005; Tuerlinckx & Wang, 2004). It generated studies on the use of this procedure in various contexts, as it offers an alternative to specialised IRT calibration softwares. NLMixed estimates the parameters using a marginal maximum likelihood estimation method, which treat ability as random effects. In the estimation of item parameters, an ability distribution is used, and the procedure does not automatically lead to estimates of ability for examinees. Estimation programs often present different challenges to implement solutions based on IRT models. Hence, this paper contributes to studies on the use of the procedure, for PCM true score equating and estimation of asymptotic standard errors.

Two popular equating designs in Rasch-related equating will be considered in this paper - the common-item equating and the concurrent equating. Other equating designs (e.g. common-person equating) are not attempted here. These could be areas for future research. In the common-item equating design, two forms (say Tests U and V) to be equated share

common items, and the estimated parameters of the common items are used to derive a scaling constant, to place Test V on the same scale as Test U (see Table 1). Two separate calibrations are needed; one for data from Test U and another for the data from Test V, giving rise to two sets of item parameters. In the concurrent equating design, there is only one estimation run that make use of the combined data from the two tests, and assigning missing values to those cells labelled as 'Not Applicable' in Table 1.

Table 1: Common-item equating for simplified example

		items		
Test U Examinee Group 1	persons	Unique items Subtest *X* (2 items)	Common items Anchor test R_1 (3 items)	Not Applicable
Test V Examinee Group 2		Not Applicable	Common items Anchor test R_2 (3 items)	Unique items Subtest *Y* (2 items)

Equating of Tests Modelled Using the Partial Credit Model (PCM) for Non Equivalent Group of Examinees

The proposed formulas for the PCM model, adapted from those proposed for the GPCM by Wong (2015), are illustrated here using a simplified example. Suppose two groups of examinees, Groups 1 and 2, take Tests *U* and *V* as shown in Table 1. Suppose Test U has two unique items in subtest X and three common items in anchor test R_1. Similarly Test V has two unique items in subtest Y and three common items in anchor test R_2. Let's assume that all the subtests involve only polytomous items with three categories of number correct scores (t=0,1 or 2), with a score of 0 for category 1, a score of 1 for category 2 and so on. This means that each item has two threshold parameters (denoted by b_{kgh}). If an examinee with ability θ attempted the gth item of the kth test, then according to Master's (1982) PCM, the probability function of a getting a score of t is given as follows:

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

Note that the value of 2 in the summation symbol corresponds to our simplified assumption of having only three categories for each item. Suppose we wish to put the scale of Test V onto the scale for Test U. The probability function for items in Test U is in the form given by Equation (1). For items in Test V, the function takes on a slightly different form to cater to the "shift" value. It is:

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

where *B* is the "shift" value (i.e. also known as the equating coefficient), to put the two tests on the same scale.

Next, we define the true scores of tests *V* as:

$$\eta = \sum_{g=1}^{n_{R_2}}\sum_{t=1}^{2} tP_{R_2gt}(\theta) + \sum_{g=1}^{n_Y}\sum_{t=1}^{2} tP_{Ygt}(\theta) \quad (3)$$

Here n_Y=2 and $n_{R_2} = 3$. This is the familiar expression making use of the expected score formula, i.e. $\sum xP(X = x \mid \theta)$. The true score for Test U denoted by ξ is defined in a similar manner. To equate the two tests using the common items method, we first obtain the estimates of both the item parameters (i.e. $\hat{\underset{\sim}{\alpha}}$ which is a column vector made up of the estimated item thresholds b_{kgh}) and the shift value/equating coefficient (i.e. $\hat{B}$). The former is obtained from the NLMixed calibration, whilst the equating coefficient is computed from the item estimates of the common items. Let $\hat{\underset{\sim}{\alpha}}$ and $\hat{B}$ be collectively denoted by the vector $\hat{\beta} = (\hat{\underset{\sim}{\alpha}}', \hat{B})'$, then the asymptotic variance of $\hat{\eta}$ is obtained using the delta method as follows:

$$a\,\mathrm{var}(\hat{\eta}) = \frac{\partial \eta}{\partial \underset{\sim}{\beta}'} a\,\mathrm{cov}(\hat{\beta}) \frac{\partial \eta}{\partial \underset{\sim}{\beta}} \quad (4)$$

The steps to derive the terms on the RHS of (4) are given in Table 2. The formulas for the PCM model shown in this are adapted from those for the GPCM presented in the paper by Wong (2015). Using our simplified example, the item parameters obtained from NLMixed and the values obtained using the formulas corresponding to an ability value of -0.55 are shown in Table 3. This ability value is arbitrarily selected for illustration purpose, as it correspond to a true score of 4 for Test U. In practice, a number of ability values could be used, corresponding to the true scores of interest in Test U (e.g. those corresponding to the cut-scores in Test U).

Concurrent Equating

To work out the SE for concurrent equating, the steps are largely the same but simpler. Only the formulas related to Test V in Tables 2 or 3 are needed and there is also no need for partial derivatives related to B, which is set to zero. In addition, the aCov matrix comprises of only one variance-covariance matrix.

Table 2: Steps to derive the asymptotic standard errors for the PCM

Step	Test		Cells in Table 3	Relevant Formulas
1.	V, U	Obtain the partial derivatives of P_{kgt} wrt ability	[1A],[1B], [4A],[4B]	$\frac{\partial P_{kgt}(\theta)}{\partial \theta} = P_{kgt}(\theta)\left[(t+1) - \sum_{s=0}^{2}\{P_{kgs}(\theta)(s+1)\}\right]$
2.	V, U	Obtain the partial derivatives of P_{kgt} wrt the each item threshold	[2A],[2B], [3A],[3B], [5A],[5B], [6A],[6B]	$\frac{\partial P_{kgt}(\theta)}{\partial b_{kgh}} = \begin{cases} -P_{kgt}(\theta)\left[1 - \sum_{s=h}^{2}\{P_{kgs}(\theta)\}\right] & \text{if } h \le t \\ -P_{kgt}(\theta)\left[0 - \sum_{s=h}^{2}\{P_{kgs}(\theta)\}\right] & \text{otherwise} \end{cases}$
3.	V, U	Obtain the partial derivatives of η wrt abilities	[8]=sum[1]= sum(*1**[1A]+*2**[1B]) [9]=sum[4]= sum(*1**[4A]+*2**[4B])	$\frac{\partial \eta}{\partial \theta} = \sum_{k\in(R_2,Y)} \sum_{g=1}^{n_k}\sum_{t=1}^{2} t \frac{\partial P_{Ygt}(\theta)}{\partial \theta}$ Similar equation for $k \in (R_1, \mathrm{X})$
4.	V	Obtain the partial derivatives of η wrt the each item threshold in test V	[2] = (*1**[2A]+*2**[2B]), [3] = (*1**[3A]+*2**[3B])	$\frac{\partial \eta}{\partial \underset{\sim}{\alpha}_{Ygh}} = \sum_{g=1}^{n_Y}\sum_{t=1}^{2} t \frac{\partial P_{Ygt}(\theta)}{\partial \underset{\sim}{\alpha}_{Ygh}}$; Similar equation for R_2
5.	U	Obtain the partial derivatives of η wrt each item threshold in test U	[5C]= -(*1**[5A]+*2**[5B])/[9] [6C]= -(*1**[6A]+*2**[6B])/[9]	First, obtain $\frac{\partial \theta}{\partial \underset{\sim}{\alpha}_{Xgh}} = \frac{-\sum_{t=1}^{2} t\left[\partial P_{Xgt}(\theta)/\partial \underset{\sim}{\alpha}_{Xgh}\right]}{\sum_{k\in(X,R_1)}\sum_{g=1}^{n_k}\sum_{t=1}^{2} t[\partial P_{kgt}(\theta)/\partial\theta]}$ Similar equation for R_1
			[5]=[8]*[5C] and [6]=[8]*[6C]	$\frac{\partial \eta}{\partial \underset{\sim}{\alpha}_{Xgh}} = \frac{\partial \eta}{\partial \theta}\frac{\partial \theta}{\partial \underset{\sim}{\alpha}_{Xgh}}$ where $\frac{\partial \eta}{\partial \theta}$ is from step 3 above.
6.	V	Obtain the partial derivatives of η wrt equating coefficient *B*	- [8]	$\frac{\partial \eta}{\partial B} = \sum_{k\in(R_2,Y)} \sum_{g=1}^{n_k}\sum_{t=1}^{2} t \frac{\partial P_{kgt}(\theta)}{\partial B}$ = -[8] since $\frac{\partial P_{kgt}(\theta)}{\partial B} = -\frac{\partial P_{kgt}(\theta)}{\partial \theta}$ where [8] is from step 3 above

Step	Test		Relevant Formulas
7.		Form row vector1: {[5],[6],[2],[3],-[8]} which is a 21x1 row vector since 10+10+1=21	$\frac{\partial \eta}{\partial \beta'_{o\!/}}$ where $\hat{\beta} = (\hat{\alpha}'_{o\!/}, \hat{B})'$
8.		Form row vector2. Since 1/(mp)= 1/(2*3)=1/6=0.167, row vector2= {0.167, 0.167, 0.167, 0, 0, 0.167, 0.167, 0.167, 0, 0, -0.167, -0.167, -0.167, 0, 0, -0.167, -0.167, -0.167, 0, 0}	$\frac{\partial B}{\partial \alpha'_{o\!/}}$ Non zero partial derivatives are: $\frac{\partial B}{\partial b_{R_1 j1}} = \frac{\partial B}{\partial b_{R_1 j2}} = \frac{1}{2p}$; $\frac{\partial B}{\partial b_{R_2 j1}} = \frac{\partial B}{\partial b_{R_2 j2}} = -\frac{1}{2p}$ $(j = 1,.., p)$ Note: '2' in the denominator reflects 2 thresholds here
9.		Extract variance-covariance matrices from NLmixed and form aCov, which is a 20x20 matrix	$a\,\mathrm{cov}(\hat{\alpha}_{o\!/})$ This is a block diagonal matrix comprising of the two covariant matrices – The matrix from Test U is on the top left and the matrix from Test V is on the bottom right.
10.		Form vector3 by multiplying TRANSPOSE (vector2) by aCov, which yield {1x20}x{20x20}=1x20 row vector	$a\,\mathrm{cov}(\hat{B}; \hat{\alpha}'_{\%}) = \frac{\partial B}{\partial \alpha'_{o\!/}} a\,\mathrm{cov}(\hat{\alpha}_{\%})$
11.		Form scalar1 by multiply vector2 x aCov x TRANSPOSE (vector2), which yield a scalar, since {1x20}x{20x20}x{1x20}=1x1	$a\,\mathrm{cov}(\hat{B}) = \frac{\partial B}{\partial \alpha'_{o\!/}} a\,\mathrm{cov}(\hat{\alpha}_{\%}) \frac{\partial B}{\partial \alpha_{o\!/}}$
12.		Form partitioned matrix1 as follows: aCov | TRANSPOSE (vector 3) Vector3 | scalar1 This is a 21x21 matrix	Since $\hat{\beta} = (\hat{\alpha}'_{o\!/}, \hat{B})'$ Partitioned matrix: $a\,\mathrm{cov}(\hat{\beta}) = \begin{pmatrix} a\,\mathrm{cov}(\hat{\alpha}_{\%}) & a\,\mathrm{cov}(\hat{\alpha}_{\%}; \hat{B}) \\ a\,\mathrm{cov}(\hat{B}; \hat{\alpha}'_{o\!/}) & a\,\mathrm{cov}(\hat{B}) \end{pmatrix}$
13.		Finally, the variance for a particular theta (here theta=-0.55) is computed as follows: vector1 x matrix1 x TRANSPOSE (vector1). The square root of the variance is the required SE for the specified theta	$a\,\mathrm{var}(\hat{\eta}) = \frac{\partial \eta}{\partial \beta'_{o\!/}} a\,\mathrm{cov}(\hat{\beta}) \frac{\partial \eta}{\partial \beta_{o\!/}}$

Table 3: Item parameters of simplified example and derived values using the asymptotic SE formulas for PCM (ability assumed to be -0.55)

											t=1, h=1	t=2, h=1	t=1, h=2	t=2, h=2				
ITEM	TEST	B	theta	b1	b2	$P_{kg1}(\theta)$	$P_{kg2}(\theta)$	$\frac{\partial P_{kg1}(\theta)}{\partial \theta}$	$\frac{\partial P_{kg2}(\theta)}{\partial \theta}$	$\frac{\partial \eta}{\partial \theta}$	$\frac{\partial P_{kg1}(\theta)}{\partial b_{kg1}}$	$\frac{\partial P_{kg2}(\theta)}{\partial b_{kg1}}$	$\frac{\partial P_{kg1}(\theta)}{\partial b_{kg2}}$	$\frac{\partial P_{kg2}(\theta)}{\partial b_{kg2}}$			$\frac{\partial \eta}{\partial b_{kg1}}$	$\frac{\partial \eta}{\partial b_{kg2}}$
								[1A]	[1B]	[1]	[2A]	[2B]	[3A]	[3B]			[2]	[3]
1	**V**	0.4667227	-0.55	0.5046	1.5567	0.1768	0.0135	0.1408	0.0242	0.1893	-0.1432	-0.0109	0.0024	-0.0133			-0.1650	-0.0242
2	**V**	0.4667227	-0.55	-1.5660	0.4374	0.5522	0.1290	0.1048	0.1535	0.4118	-0.1760	-0.0411	0.0712	-0.1124			-0.2583	-0.1535
3	**V**	0.4667227	-0.55	-2.6520	-1.3430	0.3875	0.5370	-0.1788	0.2892	0.3996	-0.0293	-0.0406	0.2081	-0.2486			-0.1104	-0.2892
4	**V**	0.4667227	-0.55	-0.6364	0.5322	0.3738	0.0794	0.1747	0.1165	0.4078	-0.2044	-0.0434	0.0297	-0.0731			-0.2912	-0.1165
5	**V**	0.4667227	-0.55	-1.0679	1.0782	0.4823	0.0594	0.1924	0.0830	0.3585	-0.2211	-0.0272	0.0286	-0.0558			-0.2755	-0.0830
									Sum=	1.7669[8]								
											t=1, h=1	t=2, h=1	t=1, h=2	t=2, h=2				
ITEM	TEST	B	theta	b1	b2	$P_{kg1}(\theta)$	$P_{kg2}(\theta)$	$\frac{\partial P_{kg1}(\theta)}{\partial \theta}$	$\frac{\partial P_{kg2}(\theta)}{\partial \theta}$	$\frac{\partial \eta}{\partial \theta}$	$\frac{\partial P_{kg1}(\theta)}{\partial b_{kg1}}$	$\frac{\partial P_{kg2}(\theta)}{\partial b_{kg1}}$	$\frac{\partial P_{kg1}(\theta)}{\partial b_{kg2}}$	$\frac{\partial P_{kg2}(\theta)}{\partial b_{kg2}}$	$\frac{\partial \theta}{\partial b_{kg1}}$	$\frac{\partial \theta}{\partial b_{kg2}}$	$\frac{\partial \eta}{\partial b_{kg1}}$	$\frac{\partial \eta}{\partial b_{kg2}}$
								[4A]	[4B]	[4]	[5A]	[5B]	[6A]	[6B]	[5C]	[6C]	[5]	[6]
1	**U**	0	-0.55	1.0309	1.8461	0.1681	0.0153	0.1347	0.0276	0.1898	-0.1372	-0.0125	0.0026	-0.0151	0.0851	0.0145	0.1503	0.0255
2	**U**	0	-0.55	-1.0416	0.9654	0.5460	0.1200	0.1169	0.1456	0.4082	-0.1824	-0.0401	0.0655	-0.1056	0.1377	0.0764	0.2433	0.1350
3	**U**	0	-0.55	-1.9685	-1.0942	0.3372	0.5811	-0.1684	0.2909	0.4133	-0.0275	-0.0474	0.1960	-0.2434	0.0642	0.1526	0.1134	0.2695
4	**U**	0	-0.55	-0.5931	0.5532	0.4367	0.1449	0.1194	0.1845	0.4885	-0.1827	-0.0606	0.0633	-0.1239	0.1594	0.0968	0.2817	0.1710
5	**U**	0	-0.55	-1.0376	0.9774	0.5461	0.1185	0.1184	0.1443	0.4069	-0.1831	-0.0398	0.0647	-0.1045	0.1378	0.0757	0.2434	0.1337
									Sum=	1.907[9]								

Studies Using Simulated Samples

Method

Monte-Carlo simulations were conducted to determine if NLMixed could support the use of the formulas to obtain asymptotic SE (denoted by ASE) for the PCM. 300 pairs of datasets to represent Tests U and V were simulated from random item parameters based on the PCM, and SE of equated results (denoted by SIM) was compared with those suggested by ASE. A total of 6 studies were carried out, four involving three-category items and two involving 4-category items, using common-item or concurrent equating method. The steps to carry out the first study (*PCM & Common-item Equating 1*) involving 26 three-category items and common-item equating are:

(M1) Obtain population threshold parameters for 26 items drawn from the N(0,1) and N(1,1) distributions, and reorder the two thresholds if they are reversed. Form two artificial Test U and Test V, each with 16 items such that the first 6 items represents the common item.

(M2) Simulate 300 Test U datasets, each with simulated responses of 1000 examinees based on the PCM, assuming that the ability is drawn from the N(0,1) distribution. Similarly, simulate 300 Test V datasets, but assuming that the ability is drawn from the N(0.5,1) distribution where the 0.5 is intended to represent the shift value.

(M3) Calibrate the datasets using SAS NLMixed, and output the item parameters and variance-covariance matrices from the calibrations. A sample SAS program for estimation is given in Annex A.

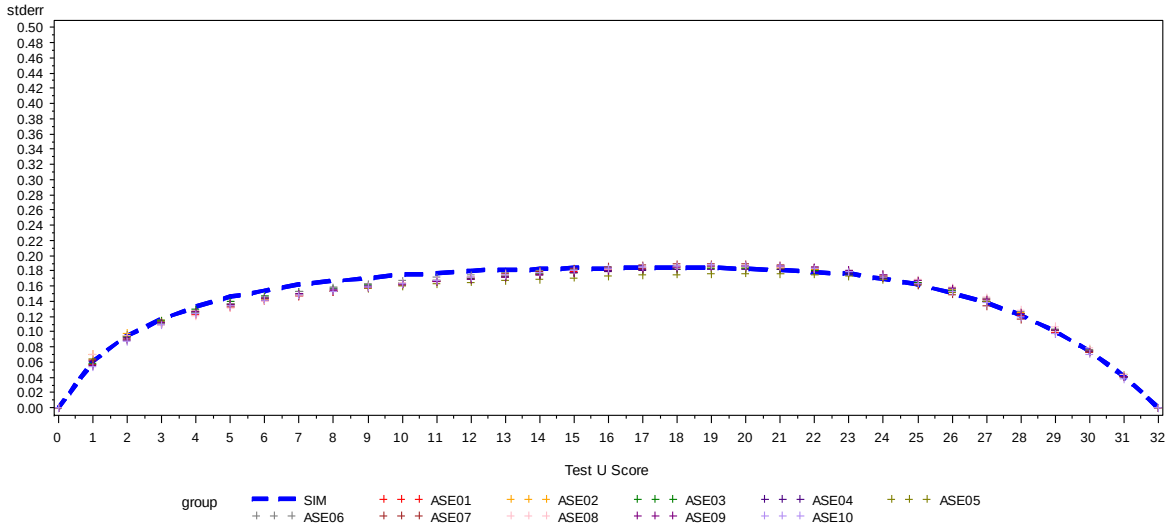
(M4) For each pair of tests, work out the shift value, and perform common-item equating across selected abilities corresponding to integer values of Test U. The SDs of the equated scores over the 300 equatings are the SEs (i.e. SIM).

(M5) To obtain the ASE, use the estimated item parameters and variance-covariance matrices, and the steps in Tables 2-3 to obtain the ASE for the selected abilities. The first 10 pairs of datasets were used to plot 10 ASE curves.

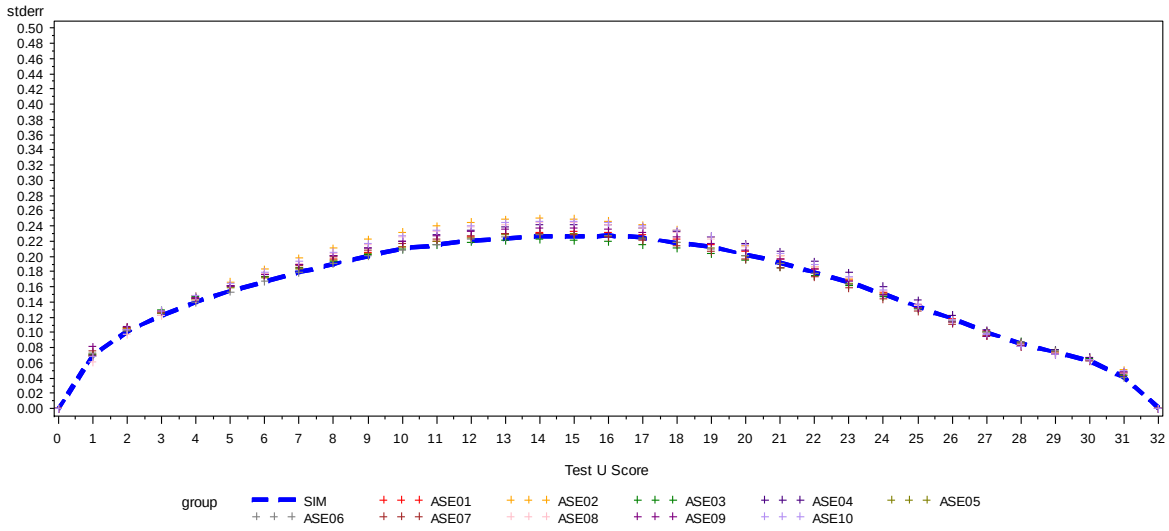
(M6) Plot the 10 ASEs against the SIM. If the ASEs are close to the SIM, then the use of the ASE formulas for PCM is tenable in NLMixed.

The *PCM &* Common-item *Equating 2* is similar to the first study, with population thresholds drawn from N(-0.5,1) and N(0.5,1) distributions in step (M1). *PCM & Common-item Equating 3* involves 26 four-category items, with population threshold parameters simulated from the N(-1,0.5), N(0,0.5) and N(1,0.5) distributions. The concurrent equating study *PCM & Concurrent Equating 1* makes use of the same simulated datasets as *PCM & Common-item Equating* 1 study and so on.

PCM & COMMON ITEM EQUATING 1

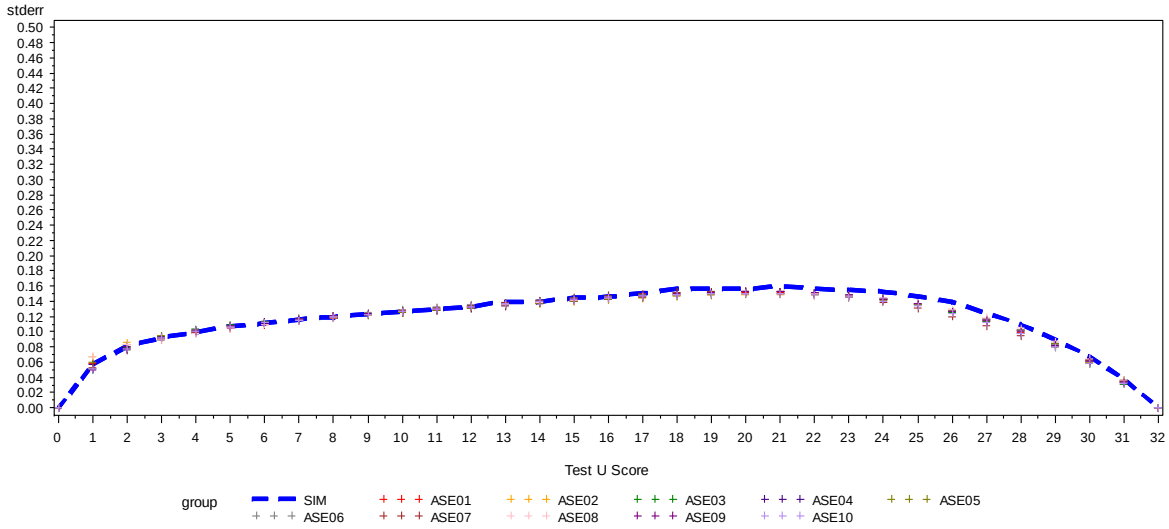


PCM & COMMON ITEM EQUATING 2

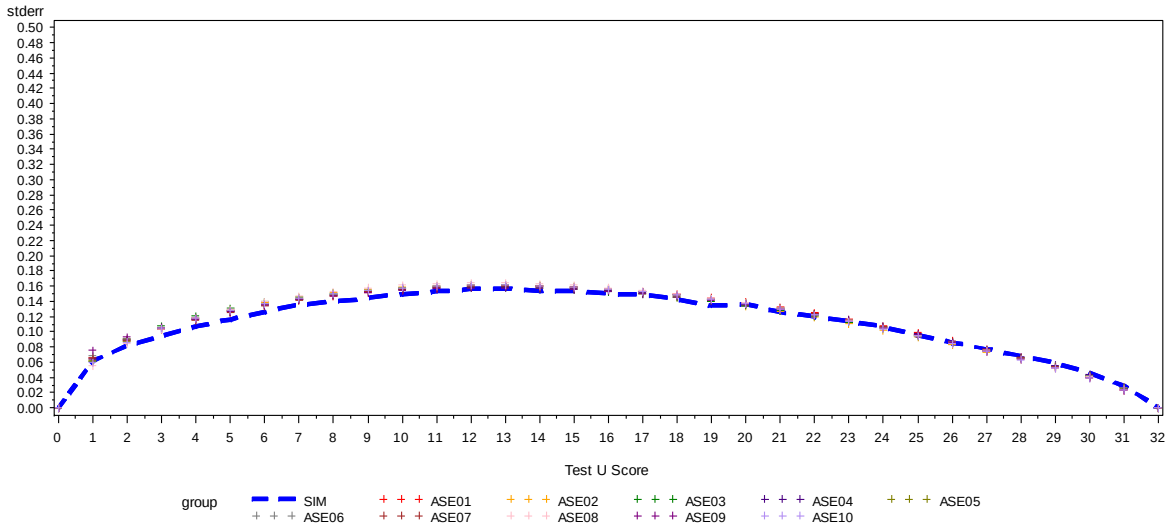


Figures 1a-b: Studies using PCM and Common Item Equating involving items with three categories. Asymptotic Standard Errors derived from the first 10 sets of samples (ASE1-ASE10), compared with standard error derived from 300 bootstrap samples (SIM).

PPCM & CONCURRENT EQUATING 1



PCM & CONCURRENT EQUATING 2



Figures 2a-b: Studies using PCM and Concurrent Equating involving items with three categories. Asymptotic Standard Errors derived from the first 10 sets of samples (ASE1-ASE10), compared with standard error derived from 300 bootstrap samples (SIM).

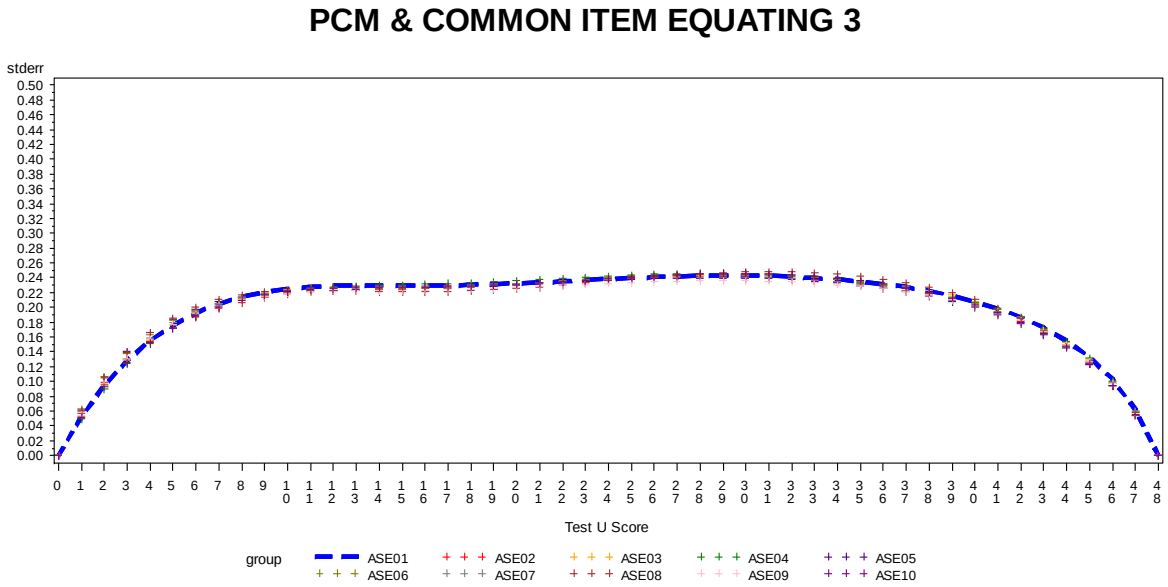


Figure 3a: Studies using PCM and Common Item Equating involving items with four categories. Asymptotic Standard Errors derived from the first 10 sets of samples (ASE1-ASE10), compared with standard error derived from 300 bootstrap samples (SIM).

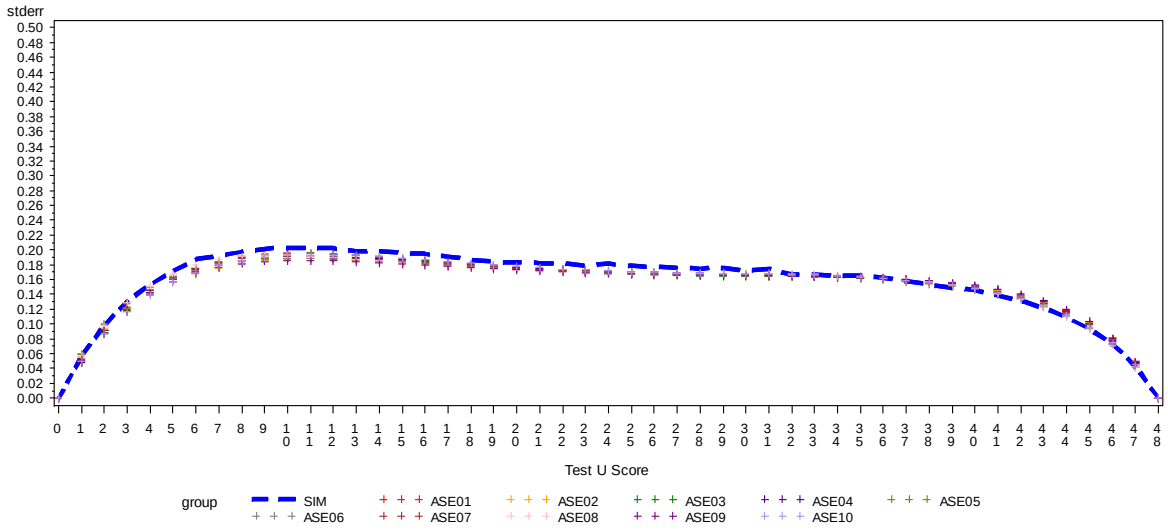


Figure 4a: Studies using PCM and Concurrent Equating involving items with four categories. Asymptotic Standard Errors derived from the first 10 sets of samples (ASE1-ASE10), compared with standard error derived from 300 bootstrap samples (SIM).

Results

Figures 1a-b and 3a show the results for the *PCM & Common Item Equating 1, 2 and 3* studies respectively. The means of the equating coefficient *B* were 0.496, 0.509 and 0.492 respectively, close to the expected value of 0.5. The 10 ASEs and SIM curves were close, lending support to the use of the formulas in NLMixed. This is also the case for the concurrent equating studies (see Figures 2a-b and 4a for the *PCM & Concurrent Equating* 1, 2 and 3 studies)

Discussion

The proposed formula to compute the asymptotic standard errors seems to work well in NLMixed, for the PCM. Results are generally comparable between the empirically computed (SIM) and analytically derived (ASE) standard errors. This is true for studies using the different equating methods (i.e. concurrent or common-item), different number of categories and different population item parameters. It demonstrated the possibility of using outputs from commercial software like SAS NLMixed 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. The use of NLMixed for the Rasch family may require additional checks to determine if the scales are invariant, as the slope is fixed at a constant. This could be done by the usual quality control plots suggested by Wright & Stone (1979), by plotting the common item parameters in Test U and Test V, after putting them on the same scale. If these points fall on the identity line, then the two scales can be considered to be on the same scale. To illustrate, Figure 5 shows one such plot from the *PCM & Common-Item Equating 1* study.

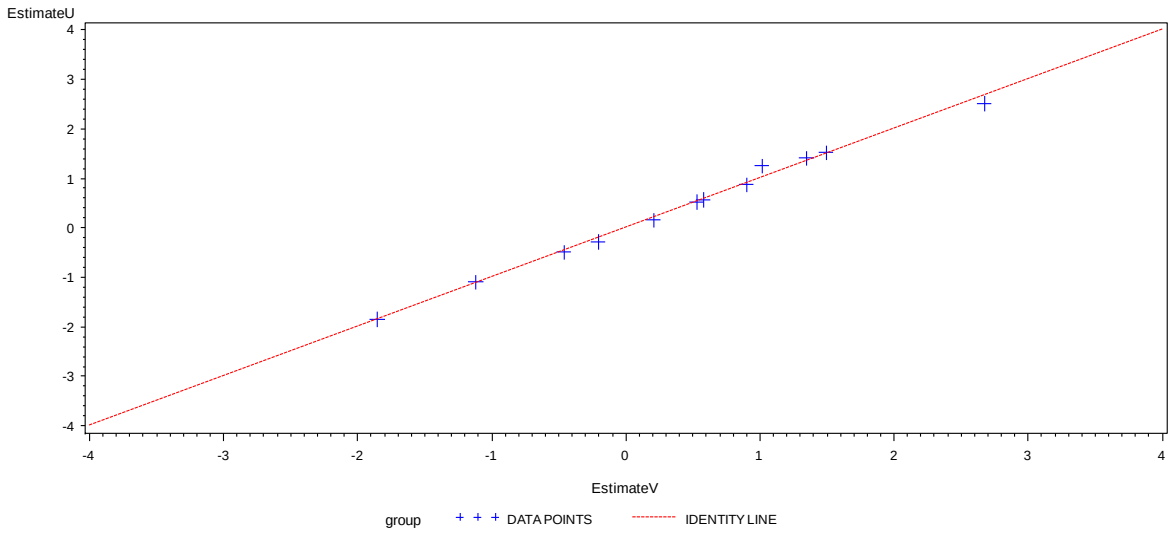


Figure 5 Quality control plot to check for the invariance of the Rasch scales after equating

There is scope for more studies related to the Rasch family of models. For instance, could NLMixed support equating and the estimation of SE using common-person equating?

How could the formulas be adapted for use with the Rating Scale model? The studies using NLMixed in this paper also surface some possible research questions. To cater to the PCM in NLMixed, in addition to the item parameters, the SD of the random effect ability distribution was also estimated. Additional studies on the possible effects of this variation during estimation could be conducted. The proposed formulas could also be studied using different commercial softwares or estimation methods, to determine if these observations are replicable.

Appendix A

Using SAS NLMixed for the 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=20 COV NOAD;

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

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;

eta1 = exp(theta-d1);
eta2 = exp((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);
if (p>1E-8) then ll=log(p);
else ll=-1E100;
MODEL Q ~ general(ll);
RANDOM theta ~ normal(0,sd**2) 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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Assocation.

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.

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

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

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

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometrika Monographs*, 34 (Suppl. 4).

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.

Wilson, M., & De Boeck, P (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43-74). New York: Springer.

Wright, B. D., and Stone, M. H. (1979). Best test design. Chicago: MESA Press.

Wong, C. C. (2015), Asymptotic Standard Errors for Item Response Theory True Score Equating of Polytomous Items. Journal of Educational Measurement, 52: 106–120.

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.

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-corrects cores. *Applied Psychological Measurement*,19, 231-241.