The impact of differential discrimination on vertical equating

Stephen Humphry
Murdoch University, Western Australia

Mailing address
Stephen Humphry
Murdoch University
Murdoch 6150
Western Australia

Acknowledgements

Permission by the Department of Education & Training, Western Australia is acknowledged for the use of the empirical data featured within the paper. Support within the Department of Education & Training has greatly benefited the research.
The work for the Report was supported in part by an Australian Research Council grant with the Australian National Ministerial Council on Employment, Education, Training and Youth Affairs (MCEETYA) Performance Measurement and Reporting Task Force, UNESCO's International Institute for Educational Planning (IEP), and the Australian Council for Educational Research (ACER) as Industry Partners*.

*Report No. 5 ARC Linkage Grant LP0454080: Maintaining Invariant Scales in State, National and International Level Assessments. D Andrich and G Luo Chief Investigators, Murdoch University

The impact of differential discrimination on vertical equating

Abstract

This paper shows the impact of differential discrimination on the results of vertical equating when the Rasch model is applied. It is shown that where there is differential discrimination, relevant items contribute information, for the purpose of vertical equating, that contains systematic error. The impact of differential item discrimination on vertical equating is examined from a theoretical perspective, and subsequently illustrated in terms of a simulation study and an empirical example. The implication of the results shown in the paper is that data should be interrogated for evidence of differential discrimination, and relevant items should be omitted in order to minimize the impact of systematic error on the results of vertical equating. In order to minimize systematic error, evidence of differential discrimination should be given precedence over evidence of most other forms of misfit. Indices such as Outfit and Infit are commonly used to diagnose misfit associated with differential discrimination. The research described in this paper also highlights a specific limitation of such indices. Accordingly, the importance of using graphical evidence to complement fit indices is emphasized.

Keywords: Discrimination, vertical equating, equating, Rasch model, systematic error, differential item functioning.

The impact of differential discrimination on vertical equating

Introduction

A requirement of the Rasch model (Rasch, 1960, 1961) is that there is a uniform level of discrimination within a specified frame of reference for measurement. In the context of assessment, a *frame of reference* (Rasch, 1977) generally comprises a class of persons responding to a class of items in a well-defined assessment context. It is shown in this paper that violation of this requirement, when it is not accounted for, results in systematic error in estimates of item locations. This error arises because item estimates depend on the distribution of person locations rather than being independent of the distribution. Consequently, the relevant items also contribute information, for the purpose of vertical equating, which contains systematic error.

Studies have been conducted to examine the effects of various factors such as item distribution, number of items, and sample size on test equating (e.g. Suanthong et al, 2000). Such factors do not, however, result in systematic error. It is important to understand the conditions under which systematic error arises in order to optimally utilize evidence from data analyses to minimize equating error.

Conditions in which differential discrimination can be accounted for are described in Humphry (2005, 2006). Rather than trying to account for differential discrimination in the process of estimation, the purpose of this paper is to demonstrate consequences of systematic error for the process and results of vertical equating and to explain relevant implications. In particular, evidence of differential discrimination should be used to exclude relevant items from analyses. The source of systematic error and its consequences are firstly analysed theoretically. The implication of the systematic error is then explained and illustrated in terms of (i) a simulation study which emulates vertical equating between year groups of students and (ii) an empirical example of equating in the context of mathematics assessment.

Background

The process of *vertical equating* is "employed when the groups of examinees differ in ability level and the tests differ in difficulty level" (Baker, 1984, p. 261). This form of equating is commonly distinguished from *horizontal equating*, in which the tests to be equated have similar difficulties and the groups of examinees similar abilities. Vertical equating is usually achieved by embedding common items within different test forms targeted to groups with different mean abilities. The purpose is to estimate item and person parameters relative to the same origin using information from the common items. This may be achieved by joint scaling based on a data matrix containing structurally missing data; i.e. a matrix containing blocks of data which are missing due to the design of an equating study. Alternatively, it may be achieved by scaling assessments separately then equating the mean item estimate for common items on the two forms.

When item response models such as the two and three parameter models are used, it is generally considered necessary to equate both the unit and origin of different metrics because the choices of both the unit and origin are seen to arise from arbitrary constraints (e.g. Baker, 1983, 1984; Kim & Hanson, 2002; Stocking & Lord, 1983). In the Rasch model, in contrast, the objective is generally regarded as equating only the origin because the requirement of uniform discrimination across items contained in both tests implies a common unit (Humphry, 2005, 2006).

The Rasch model for dichotomous data is usually stated as

$$\Pr\{X_{ni} = 1\} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}, \qquad (1)$$

where β_n is the location of person n and δ_i is the location of item i on a latent continuum. More generally, the Rasch model can be stated as

$$\Pr\{X_{sni} = 1\} = \frac{\exp(\rho_s(\beta_n - \delta_i))}{1 + \exp(\rho_s(\beta_n - \delta_i))}, \qquad (2)$$

where ρ_s is an arbitrary scaling factor (Humphry, 2005, 2006) which must be of uniform magnitude within a specified frame of reference for measurement. In this paper, the scaling factor ρ_s is not treated in the same way as the item discrimination parameter of Birnbaum's (1968) *two parameter logistic model*. Instead, it is treated as a parameter which pertains to an empirical factor associated with a set of items. Accordingly, the parameter is subscripted for the item set *s* rather than for each individual item. Differences between the magnitudes of ρ_s , $s = 1,...,S$ within a *single* frame of reference are referred to as *differential item discrimination*.

Theoretical analysis of error in item locations associated with discrimination

This paper is concerned with situations in which item locations are estimated using estimation methods based on Equation (1) but the response data for certain items empirically accord with Equation (2). In order to analyse the theoretical impact of differential discrimination on item estimates in these situations, it is instructive to begin by defining

$$\beta_n - \delta_i^{'} \equiv \beta_n - (\beta_n - \rho_s\beta_n + \rho_s\delta_i) \equiv \rho_s(\beta_n - \delta_1)\,. \qquad (3)$$

Substituting the first term of Equation (3) into Equation (1), and noting the last term is contained within Equation (2), it follows that in the presence of differential discrimination as defined above, the theoretical item location is necessarily

$$\delta_i^{'} \equiv \beta_n - \rho_s\beta_n + \rho_s\delta_i \equiv \rho_s\delta_i + (1-\rho_s)\beta_n\ . \qquad (4)$$

In turn, it follows from Equation (4) that the value of $\delta_i^{'}$ depends on the person location when $\rho_s \neq 1$. Accordingly, we let

$$\delta_{i(\beta)}^{'} \equiv \rho_s\delta_i + (1-\rho_s)\beta_n \qquad (5)$$

for the purpose of following analysis and exposition, where the subscript in parentheses denotes that the item location is dependent on the person location.

Clearly, when $\rho_s = 1$, it follows that $\delta_i^{'} = \delta_i$ in Equation (4), and it is therefore possible to estimate item parameters independently of person parameters as shown by Rasch (1960). However, when $\rho_s \neq 1$ there is a violation of the model which affects maximum likelihood estimates because the observed total score for the item, shown on the left hand side of Equation (3), depends on the magnitude of ρ_s, while the solution equation does not take the magnitude of ρ_s into account. When not accounted for, differential discrimination manifests within patterns of scores and the total score for an item. In turn, these manifestations impact on the results of estimations.

To illustrate the nature of the dependence of the item location on the person locations in the presence of differential discrimination defined in Equation (5), it is instructive to consider the values of $\delta_{i(\beta)}^{'}$ corresponding with person locations at specific points along a continuum when $\rho_s \neq 1$. Accordingly, Table 1 shows the locations of $\delta_{i(\beta)}^{'}$ for selected person locations, as listed, when $\delta_i = 1$ and $\rho_s = 2.0$.

Table 1: Theoretical item locations for selected abilities in the presence of violation of the requirement of uniform discrimination

	β	$\delta_{i(\beta)}^{'}$	π	I
	-3.0	5.0	0.018	0.018
	-2.5	4.5	0.029	0.028
	-2.0	4.0	0.047	0.045
	-1.5	3.5	0.076	0.070
	-1.0	3.0	0.119	0.105
	-0.5	2.5	0.182	0.149
	0.0	2.0	0.269	0.197
	0.5	1.5	0.378	0.235
	1.0	1.0	0.500	0.250
	1.5	0.5	0.622	0.235
	2.0	0.0	0.731	0.197
	2.5	-0.5	0.818	0.149
	3.0	-1.0	0.881	0.105
Mean	0.0	2.0		

In Table 1, π is the probability of a correct response given by Equation (2) and $\mathrm{I} = \pi(1-\pi)$ is the Fisher information, which is the negative of the second derivative of the log likelihood function with respect to the item parameter (e.g. Andrich, 1988).

Two key points are noted in relation to Table 1. First, $\overline{\delta}_i^{'} = \rho_s \delta_i$ for the listed person locations, which are symmetric about the mean person location. Second, the Fisher information is symmetric about $\delta_i = 1.0$, and therefore the information is *not* symmetric about $\overline{\beta}$ since $\delta_i \neq \overline{\beta}$.

An estimate of $\delta_i^{'}$ will depend on the person locations and their distribution, and it will also depend on information provided by persons with different locations. Accordingly, we define the approximate theoretical location of $\delta_i^{'}$ as the weighted mean

$$\delta_i^{'} = \frac{\left(\sum_n \delta_{i(\beta)}^{'} \mathrm{I}_n \right)}{\sum_n \mathrm{I}_n}. \qquad (6)$$

That is, the location is defined as the mean of the item locations $\delta_{i(\beta)}^{'}$ taken across relevant persons given their locations, β_n, weighted for Fisher information.

To examine the impact of differential discrimination on item estimates, values of $\delta_i^{'}$ were computed for values of δ_i across the range –3 to +3 in increments of 0.5, and for different values of ρ_s, based on Equation (6). The values of $\delta_i^{'}$ were obtained using a set of person locations with density simulated according to a normal distribution. The locations $\delta_i^{'}$ were then plotted against simulated locations δ_i. The resulting plots indicated that in general, the approximate linear relationship

$$\delta_i^{'} - \overline{\beta} \cong c\rho_s(\delta_i - \overline{\beta}) \qquad (7)$$

holds when ρ_s lies in the range $[0.5, 2]$, where c lies in the range $[0.733, 1.094]$. Values of c obtained using this process are shown for a selection of values of ρ_s in Table 2. As evident in this table, the resulting plots indicate that $c > 1$ when $\rho_s < 1$ and $c < 1$ when $\rho_s > 1$, for values of ρ_s in the specified range.

Table 2: Selected values of ρ and c obtained from simulated data

ρ	c	$b = c\rho$
0.50	1.094	0.547
0.67	1.076	0.721
1.00	1.000	1.000
1.50	0.863	1.295
2.00	0.733	1.465

A preliminary simulation study is used next to show that the relationship between estimates and theoretical locations is approximately linear as predicted from application of Equation (6) to obtain Equation (7).

Preliminary simulation

In the preliminary simulation study, response data were simulated for three sets of 13 items, as shown in Table 3. Persons were simulated according to a theoretical normal distribution with mean 0.2 and standard deviation 1.0; i.e. $\beta \sim N(0.2,1.0)$.

Table 3: Key features of the preliminary simulation

Item set	ρ_s	I_s	Min	Max	Increment
1	0.50	13	-3.0	3.0	0.5
2	1.00	13	-3.0	3.0	0.5
3	2.00	13	-3.0	3.0	0.5

A scatterplot showing simulated and estimated item locations for the preliminary simulation study is shown in Figure 1. The key feature noted in relation to this scatterplot is the compression and expansion of estimates for sets 1 and 3, respectively, around the mean simulated person location of 0.2. Accordingly, the lines of best fit intersect at approximately 0.2, consistent with Equation (7). For example, it is readily shown that the lines of best fit for sets 1 and 3 shown in Figure 1 intersect at 0.198 on the *x*-axis.

Thus, it is clearly evident in Figure 1 that the estimates of items in sets 1 and 3 contain systematic error.

Figure 1 shows that the relationship is somewhat less linear for item set 3, for which $\rho_3 = 2$, than for either of the other sets. It is important to note, however, that the line of reasoning employed in this paper does *not* require the relationship to be exactly linear. Rather, Equation (7) is an approximation which is convenient for exposition of the nature and implications of the systematic error evident in Figure 1.

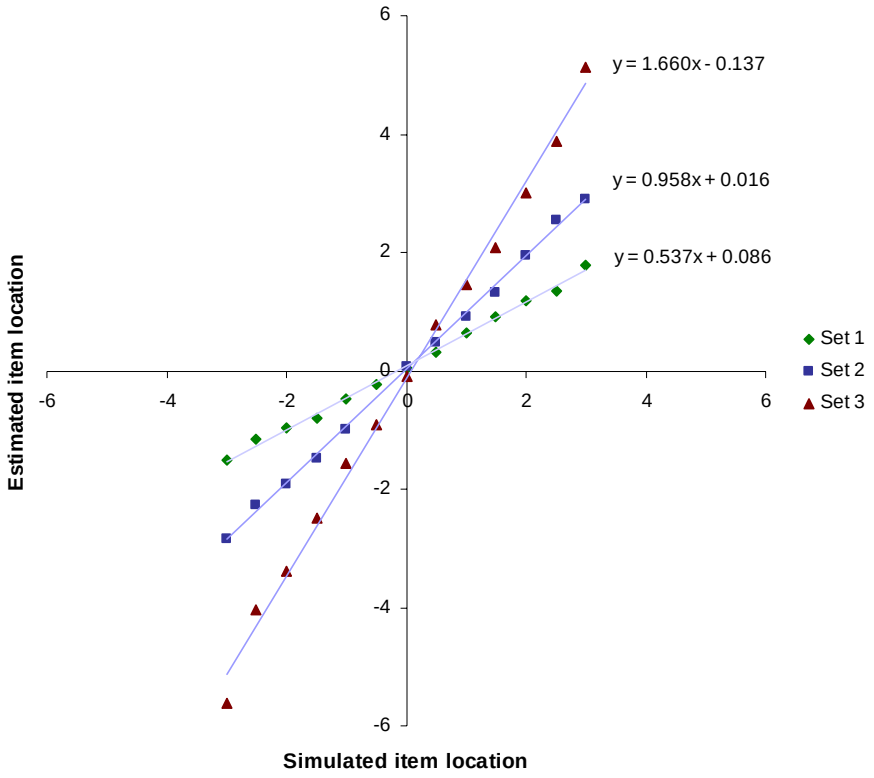


Figure 1: Simulated and estimated item locations for preliminary simulation

Theoretical analysis of the impact of error on vertical equating

In order to analyse the impact of the systematic error contained in item estimates on the results of common item equating, we suppose that item locations are scaled separately for different groups of students with different mean abilities. This method of vertical equating is used in both the main simulation study and the empirical illustration which follow. It is stressed, however, that the same items and information are instrumental to placing item estimates obtained from each assessment on a single scale, whether item locations are scaled simultaneously or separately. Consequently,

the implication of the violation of the model for equating holds whether locations are scaled simultaneously or scaled separately then equated, as has been verified in the main simulation study which follows.

In the simulation and empirical studies to follow, the constraint

$$\bar{\delta}_{i(g)} = 0 \qquad (8)$$

is imposed in the analysis of response data for items attempted by group *g*, resulting in a different origin for estimates obtained from the data of each year group. In both the main simulation and empirical illustration which follow, there are two groups of students and two assessments. Consequently, each common item *i* has two locations, $\delta_{i(g)}$ and $\delta_{i(h)}$, which arise from imposing the constraint of Equation (8) separately to each of the sets of items attempted by groups *g* and *h*.

Because each common item is required to have a single location on a common metric, the difference between estimates of the same item provides an estimate of the translation constant required to equate two sets of item estimates. That is, the estimate of the translation constant $t_{g\to h}$ obtained from item i is

$$\hat{t}_{g\to h(i)} = \hat{\delta}_{i(h)} - \hat{\delta}_{i(g)}. \qquad (9)$$

The mean $\bar{\hat{t}}_{g\to h}$ provides a single estimate based on all common items. Clearly, this mean is equal to the difference between the mean of the estimates of the common items on the two metrics; i.e. $\bar{\hat{t}}_{g\to h} = \bar{\hat{\delta}}_{i(h)} - \bar{\hat{\delta}}_{i(g)}$ where the means are computed only for the common link items.

Now, let $\delta'_{i(g)}$ be the theoretical location of the item as defined in Equation (6), when this location is referenced to the origin that results from the constraint of Equation (8) on the set of items attempted by group *g*. It follows from Equation (7) given these definitions that

$$\delta'_{i(g)} - \overline{\beta}_{g(g)} \cong c\rho_s(\delta_{i(g)} - \overline{\beta}_{g(g)}), \tag{10}$$

where $\overline{\beta}_{g(g)}$ is the mean person location when the mean location of the items attempted by group *g* is constrained to 0. Equation (10) is included to emphasize that the systematic error shown in Figure 1 operates about the mean person location of the group *g*. The importance of this is that the same item is expected to contain a different error when administered to each of two groups with different mean abilities. That is, the error is dependent on the targeting of the item to each of the relevant groups, as shown in Figure 2 in the following section.

Now, to simplify matters, from Equations (9) and (10) it follows that when there is differential discrimination, the estimate of the translation constant obtained from relevant items is expected to contain systematic error that is largely independent of the specific locations of the items. Specifically

$$\begin{aligned} t'_{g\to h(i)} &= \delta'_{i(h)} - \delta'_{i(g)} \\ &\cong c\rho_s(\delta_{i(h)} - \delta_{i(g)}) + (\overline{\beta}_{h(h)} - \overline{\beta}_{g(g)})(1 - c\rho_s). \\ &\cong c\rho_s(t_{g\to h}) + (\overline{\beta}_{h(h)} - \overline{\beta}_{g(g)})(1 - c\rho_s) \end{aligned} \tag{11}$$

When the targeting of each of the tests to the persons is similar, the mean abilities on the scales obtained for each of the tests will be similar; that is, $\overline{\beta}_{g(g)}$ and $\overline{\beta}_{g(h)}$ will be similar. Consequently, when the targeting of each assessment to each group is similar, the error is largely dependent on the magnitudes of ρ_s and $t_{g\to h}$.

Thus, an estimate $\hat{t}_{g\to h(i)}$ is generally expected to contain *systematic error* when $\rho_s \neq 1$, except in very particular cases such as when $t_{g\to h} = 0$ and $\overline{\beta}_{g(h)} - \overline{\beta}_{g(g)} = 0$. In particular, in the presence of high levels of discrimination, the absolute magnitude of the estimate of $t_{g\to h}$ is expected to too large; that is, $\left|\hat{t}_{g\to h}\right| > \left|t_{g\to h}\right|$. Conversely, low discrimination is expected to result in estimates of the translation constant whose absolute magnitude is too small; that is, $\left|\hat{t}_{g\to h}\right| < \left|t_{g\to h}\right|$.

When conducting horizontal equating, the translation constant is generally expected to be small. Consequently, the impact of systematic error on the results of horizontal equating is generally smaller than for vertical equating.

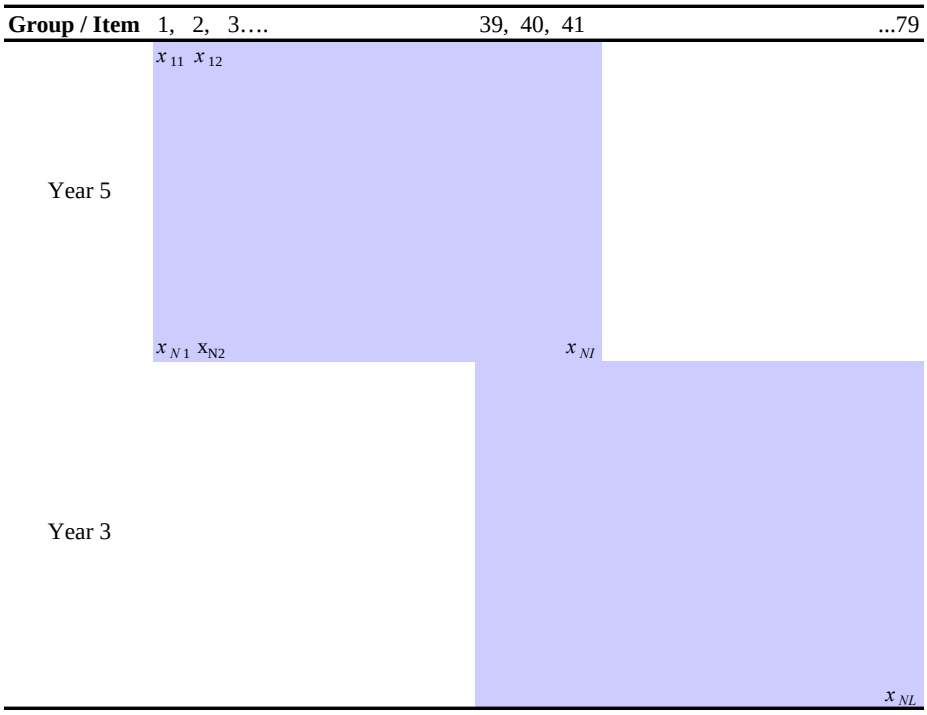
The implication of the systematic error contained in such estimates is illustrated in the main simulation study.

Main simulation study

In the main simulation study, response data for 2000 persons and 79 items were simulated as shown in Table 4. Data were simulated for $N = 1000$ students in each of the year groups 3 and 5.

In Table 4, the shaded region represents response data. It can be seen that there is structurally missing response data in the matrix due to the design of the study. For year 5 students, response data were simulated for items $i = 1,...,41$ and for year 3 students, response data were simulated for items $i = 39,...,79$. Thus, response data for both groups were simulated for items 39 to 41, as shown.

Table 4: Simulated data matrix



Person and item locations were simulated on a common metric using *SimsRasch* (Andrich & Luo, 1997-2003). The theoretical parameters are summarized in Table 5. The item parameters were simulated at equal intervals of 0.15 between the minimum and maximum. The person locations were simulated according to a theoretical normal distribution. Response data were simulated according to Equation (2) where $\rho_1 = 1$ for all items except for item 41, for which $\rho_2 = 2$. Thus, the item set 2 comprises a single item 41. The location of item 41 is $\delta_{41} = 0.7$.

Table 5: Summary of generated person and item location for simulation study

	Year 3				**Year 5**			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Persons	0.513	1.00	-	-	-0.499	1.00	-	-
Items	0.50	1.80	-3.50	2.50	-0.50	1.80	-2.50	3.50

Response data for each group were analysed separately in *RUMM2020* (Andrich et al, 1997-2005) which implements pairwise conditional maximum likelihood estimation for the Rasch model (Andrich & Luo 2003). Estimates of $\hat{\delta}'_{41(g)}, g = 3,5$ were obtained as shown in Table 6. Along side each estimate is the approximate theoretical location $\delta'_{41(g)}$ obtained from Equation (6) using the simulated person locations. It can be seen in Table 6 that there is a close correspondence between the estimates and the theoretical values.

Table 6: Comparison of estimates and approximated theoretical values

	$\hat{\delta}'_{41(g)}$	$\delta'_{41(g)}$
Year 3	1.793	1.694
Year 5	0.306	0.261

Now, from Table 6, the estimate of the translation constant based on information from item 41 is

$$\hat{t}_{5\to 3(41)} = \hat{\delta}_{i(3)} - \hat{\delta}_{i(5)} = 1.487 \qquad (12)$$

It is readily verified from Table 5 that the theoretical value of the translation constant given the relevant constraints is $t_{5\to3} = 1.0$. The error of this estimate is therefore $\hat{t}_{5\to3(41)} - t_{5\to3} = 1.487 - 1 = 0.487$.

From Equation (11) the translation constant is expected to be approximately

$$\begin{aligned} t'_{5\to3(i)} &\cong c\rho_s(\delta_{i(3)} - \delta_{i(5)}) + (\overline{\beta}_{h(3)} - \overline{\beta}_{g(5)})(1 - c\rho_s) \\ &\cong c\rho_s(t_{g\to h}) + (\overline{\beta}_{h(3)} - \overline{\beta}_{g(5)})(1 - c\rho_s) \\ &\cong 1.465(1) + (-0.001 - 0.013)(1 - 1.465) \\ &\cong 1.471 \end{aligned} \qquad (13)$$

which is close to the value of 1.487 based on the actual item estimates. The estimate of the translation constant is too large as predicted on theoretical grounds; that is, $\hat{t}_{5\to3(41)} > t_{5\to3}$. Thus, the estimate of the translation constant contains *systematic error* which is in the predicted direction, and of approximately the predicted magnitude.

The value of $c\rho_s = 1.465$ for $\rho_s = 2.0$ is shown in Table 2. The reader may note that this value differs from the coefficient of the line of best fit for item set 3 shown in Figure 1, which was 1.660. The variance of the estimates of item locations in each set in the preliminary simulation was greater than the variance of the simulated locations, resulting in an inflated estimate of the coefficients. This impact on the scale is another consequence of differential discrimination. The variance of estimates in the main simulation is very close to the variance of the simulated estimates. Consequently, the value 1.465 is used in Equation (13) rather than 1.660, as it is a better approximation.

The differences between the estimates of the common items are shown in Table 7. Each difference provides an estimate of the translation constant. The last column in Table 7 is the standardized error of the estimate of the translation constant obtained from the relevant item, which is defined as

$$\hat{\sigma}_{\hat{t}-t} = \frac{\left(\hat{\delta}_{i(3)} - \hat{\delta}_{i(5)} - t_{3\to5}\right)}{\sqrt{\hat{\sigma}^2_{i(3)} + \hat{\sigma}^2_{i(5)}}}, \qquad (14)$$

where $\hat{\sigma}_{i(g)}$ is the standard error of the estimate of item *i* obtained from response data of group *g*. It can be seen that the estimate of the translation constant obtained from item 41 contains error substantially outside bounds expected given the standard error of the item estimate.

Table 7: Estimates of locations of common items in the simulation study

Item	$\hat{\delta}_{i(5)}$	$\hat{\delta}_{i(3)}$	$\hat{t}_{5\to3(i)}$	Error	Standardized error
39	-1.13	-0.05	1.08	0.08	0.80
40	-0.58	0.56	1.14	0.14	1.37
41	0.31	1.79	1.49	0.49	4.28

Thus, the estimates of the translation constant in the simulation study confirm the predicted systematic error contained within the estimate $\hat{t}_{5\to3(41)}$.

The implication of the systematic error shown in Table 7 is that item 41 provides for a poor estimate of the translation constant and should be omitted. Accordingly, the translation constant was estimated for the main simulation study from items 39 and 40, yielding an estimate $\bar{\hat{t}}_{5\to3} = 1.113$. The error using these items is 0.113. The translation constant estimated from all items was $\bar{\hat{t}}_{5\to3} = 1.237$. The reduction in the error when item 41 is omitted is 0.124.

Having omitted item 41 in the estimation of $t_{5\to3}$, locations of the year 5 items for the common items were transformed onto the year 3 metric according to the transformation formula

$$\hat{\delta}_{i(5\to3)} = \delta_{i(5)} + \bar{\hat{t}}_{5\to3}. \qquad (15)$$

A scatterplot showing the correspondence between the year 3 item estimates and the year 5 transformed estimates is shown in Figure 2 against the reference line *y* = *x*.

Although it was not used in the estimation of the translation constant, item 41 is included in Figure 2, after transformation of the year 5 estimate according to Equation (15), to highlight the systematic error contained within the two estimates of the item's location. For item 41, the errors are $\hat{\delta}_{41(3)} - \delta_{41(3)} = 0.593$ and $\hat{\delta}_{41(5)} - \delta_{41(5)} = 0.106$. Thus, the error is larger for the year 3 estimate than for the year 5 estimate, consistent with Equation (11), due to the fact that the item's location is further from the year 3 mean person location than from the year 5 mean person location. Both errors are in a positive direction, but in Figure 2 the error in the *x*-axis is larger than the error in the *y*-axis as shown.

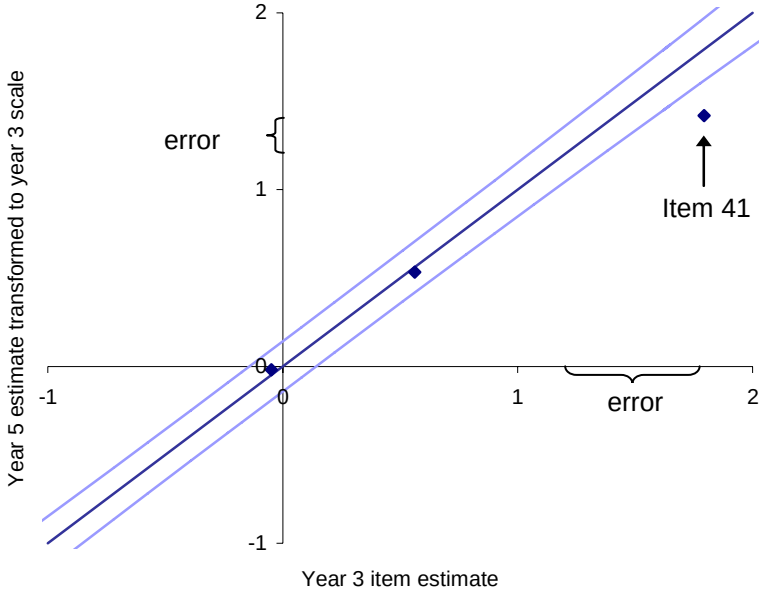


Figure 2: Scatterplot for simulation study after transforming year 5 estimates with 95% confidence bands

In analogous fashion, it is readily shown that in the presence of a low level of discrimination, an opposite impact on the translation constant is expected. For example, from Table 2, $c = 1.094$ when $\rho_s = 0.5$. Inserting $c\rho_s = 1.094 \times 0.5 = 0.547$ in Equation (13) gives $t'_{5\to3} \cong 0.540$ which, in turn, gives an approximate error of $t'_{5\to3} - t_{5\to3} \cong 0.540 - 1 \cong -0.459$. Hence, the expected magnitude of the error for an item belonging to a set for which $\rho_s = 0.5$ is very similar to that observed in the main simulation, but in the *opposite direction*.

In the simulation study, the magnitude of ρ_s is known. In empirical contexts, relevant evidence must be used to identify items for which there is a violation of the requirement of uniform discrimination. Such evidence is discussed in terms of the empirical illustration which follows, and is further discussed in the section following the illustration.

Empirical illustration

The data used in the empirical illustration were collected in the Western Australian Literacy and Numeracy Assessment (WALNA) testing program in 2004. The program involves reading, writing, mathematics, and spelling assessments in years 3, 5, and 7 administered by classroom teachers based on detailed administrative instructions. Approximately 30,000 students in each year group participate in the testing program. RUMM2020 (Andrich, Sheridan, & Luo, 1997-2005) was again used to analyse the data.

Nine items were included in both the year 3 and 5 Numeracy assessments in 2004 as a basis for vertical equating. Based on all items, the estimated translation constant was $\bar{\hat{t}}_{5\to 3} = 1.083$. This constant was used to transform the year 5 estimates onto the year 3 metric. The resulting estimates are shown in Table 8.

Table 8: Estimates of the locations of common items in the empirical study

Y3 item label	$\hat{\delta}_{i(3)}$	Fit residual	Y5 item label	$\hat{\delta}_{i(5\to 3)}$	Fit residual
N04304	-0.798	**-2.58**	N04503	-1.614	**-1.53**
N04308	-2.068	-2.23	N04505	-1.401	-1.26
N04312	-0.462	0.32	N04507	-0.382	-1.41
N04325	2.375	-1.48	N04515	1.856	-0.58
N04326	0.192	4.08	N04517	0.616	0.16
N04328	1.056	2.31	N04519	1.740	-0.80
N04329	1.634	**2.13**	N04526	1.919	**5.77**
N04309	-0.603	**-3.94**	N04530	-0.632	**-3.46**
N04324	1.449	**-3.21**	N04536	0.677	**-0.72**
Mean	0.308			0.308	
SD	1.423			1.381	

The fit residual shown in Table 8, which is produced in RUMM2020, is defined as

$$Z_i = \frac{\log\left(Y_i^2 / F_i\right)}{\sqrt{V\left[Y_i^2 / F_i\right]}} \qquad (16)$$

where $Y_i^2 = \sum_n z_{ni}^2$, $z_{ni} = \left(x_{ni} - E[X_{ni}]\right) / \sqrt{V[X_{ni}]}$, $E[X_{ni}]$ is the expected score for person *n* on item *i*, x_{ni} is the observed response and F_i is the degrees of freedom. The calculation of this fit residual involves a log transformation of the Outfit statistic in the numerator. The expected value of Z_i is 0 and its variance is 1. The log transformation makes the theoretical distribution of the fit residual symmetric. Due to the sensitivity of *Z* to sample size, a random sample of 1000 of the total of 30,000 cases was used to estimate the fit residuals shown in Table 8.

A scatterplot of the item estimates in Table 8 is shown in Figure 3. Items for which $|Z_i| > 2.5$ for one or both year groups were targeted for omission, other than the item labelled N04326. The reason for retaining the item N04326 is that graphical evidence indicated the magnitude of the residual is mainly attributable to the effects of guessing at the lower end of the continuum rather than differential discrimination across the range of the continuum. The importance of graphical evidence is discussed in the following section. Where there is evidence of guessing, the relevant persons generally have locations which are considerably lower than the item location. As a person's location becomes further from the item location, less information is provided about the item's location. Thus, the effects of guessing on the item's location are generally less than the effects of differential discrimination operating across the range of the continuum. Other problems are associated with such misfit, but these are beyond the scope of this paper.

The items targeted for omission are highlighted in bold in Table 8, and are identified by their labels in Figure 3.

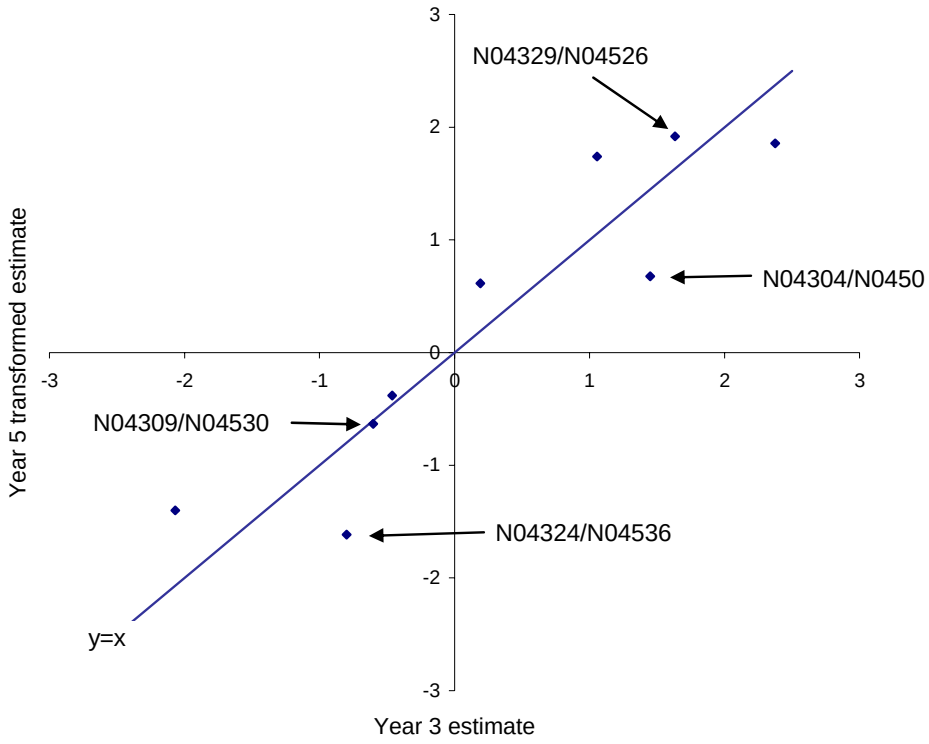


Figure 3: Estimates using all common items for equating

In Figure 3, two highlighted items have estimates which appear to contain systematic error. These items have the labels N04304/N04503 and N04324/N04536. Both items have negative fit residuals for both year groups, indicating a high level of discrimination.

The other highlighted items do not show clear evidence of systematic error. Because empirical data are used, there are likely to be other sources of error such as uniform *differential item functioning* (DIF). Such sources can be expected in some cases to confound evidence of systematic error of the nature described here. The key point is that the results shown in this paper provide a basis for identifying items that would be expected to contain systematic error based on evidence for differential item discrimination even if no other factors result in systematic error. This provides a clear theoretical rationale for selection and rejection of items so as to minimize systematic error expected based on results of analyses. Clearly, however, other evidence should also be taken into account in making decisions regarding which items to omit and retain for the purpose of vertical equating.

Omitting the four identified items from the calculation, the estimate $\bar{\hat{t}}_{5\to3} = 0.816$ was obtained in the present empirical illustration. This is substantially smaller than the estimate of 1.083 based on all common items. The scatterplot after applying this constant, and showing only the common items used in its computation, appears in Figure 4.

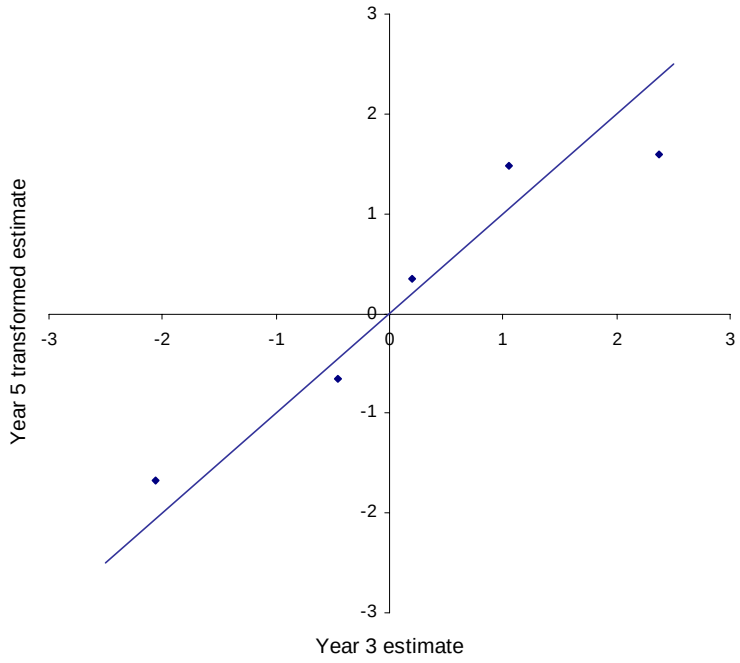


Figure 3: Estimates after omitting identified items for equating

By omitting the items highlighted in Figure 2, the correlation between the estimates increased from 0.914 to 0.961 and the root mean squared difference (RMSD) fell from 0.549 to 0.450, where

$$\text{RMSD} = \sqrt{\sum_i \left(\hat{\delta}_{i(5\to3)} - \hat{\delta}_{i(3)}\right)^2 / I_c} \qquad (17)$$

and I_c is the number of common items.

The key points to be noted are that (i) omission of the items resulted in improved correspondence between equated item estimates and (ii) the *change* in the translation constant with the removal of items on theoretical grounds is *systematic*. In contrast, if

decisions regarding which items to retain are based on which items are outliers in the scatterplot, it is unlikely that systematic changes in the estimate of the translation constant will be correctly identified when they are warranted. For example, suppose the standardized difference

$$s_{i(5-3)} = \frac{\hat{\delta}_{i(3)} - \hat{\delta}_{i(5\to 3)}}{\sqrt{\sigma^2_{i(3)} + \sigma^2_{i(5)}}} \qquad (18)$$

were computed for each item shown in Table 7 and the items having the four largest standardized differences omitted for the purpose of estimating the translation constant. The items identified in this way are items 1, 4, 6 and 9 in sequential order in the table. The resulting estimate is $\bar{\hat{t}}_{5\to 3} = 1.130$, very close to the estimate of 1.083 obtained from all items. Adopting such an approach would therefore result in little difference in $\bar{\hat{t}}_{5\to 3}$, when the evidence indicates that a change in the translation constant is warranted in order to reduce systematic error expected on theoretical grounds.

Similarly, the use of other kinds of fit indices without regard to their sensitivity to differential discrimination will also generally result in less effective reduction of systematic error in translation constants than is possible.

Graphical evidence of differential discrimination

It was mentioned earlier that computation of the fit residual defined in Equation (16) involves a logarithmic transformation of the Outfit statistic, or mean squared standardized residual. Fit indices such as the Outfit and Infit (Wright & Stone, 1979; Wright & Masters, 1982) are also sensitive to differential discrimination.

While such indices are sensitive to differential discrimination, the systematic error described in this paper carries important implications for the degree of *sensitivity* of such indices to this violation of the model. From Equation (7), for a given magnitude of ρ_s the degree of error in an estimate is expected to increase as $\left|\delta_i - \bar{\beta}\right|$ increases. In the calculation of a fit index such as Outfit, Infit, or *Z*, the *item estimate* is used to

compute the expected score for each person on a given item. That is, the expected score used in the computation of all of these fit indices is computed as

$$E[X_{ni}] = \frac{\exp(\hat{\beta}_n - \hat{\delta}_i)}{1 + \exp(\hat{\beta}_n - \hat{\delta}_i)}. \quad (19)$$

In the presence of differential discrimination the item estimate contains error which is systematic, and which increases as a function of $|\delta_i - \overline{\beta}|$. Because the estimate is used in the computation, the fit residual is *necessarily affected by the systematic error*.

Simulations show that the fit residual tends to become less sensitive to the violation of the model associated with differential discrimination as $|\delta_i - \overline{\beta}|$ increases. For example, in Figure 4 a scatterplot of item fit residuals by item locations is shown for items in set 3 of the preliminary simulation.

The value of the item fit residuals shown in Figure 4 is less than 0 across the range of item locations as expected in the presence of high discrimination, where the data conform more closely to a Guttman structure than expected (Andrich, 1988). While all values are in the expected direction, however, they range between –0.988 and –6.678. A strong tendency is apparent in Figure 4 for the magnitude of the fit residual to become smaller as $|\delta_i - \overline{\beta}|$ increases. In turn, this implies the magnitude of the fit residual tends to decrease as the degree of systematic error increases.

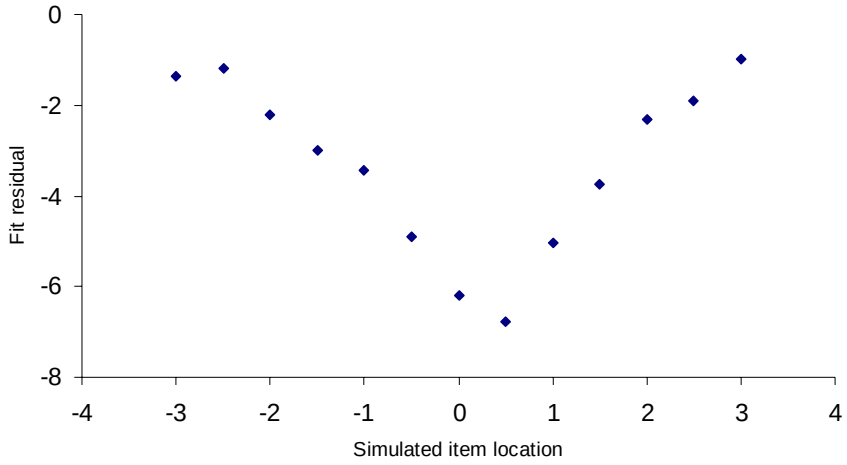


Figure 4: Fit residuals plotted against item location
for item set 3 in the preliminary simulation

For item set 3 in the preliminary study, the simulated level of discrimination was $\rho_3 = 2$ and the sample size was 1000. Due to the high level of discrimination, combined with the sensitivity of the fit index to sample size, the change in the degree of sensitivity of the fit residual as a function of $\left|\delta_i - \overline{\beta}\right|$ is strongly apparent in Figure 4. It is clear that fit residuals do not provide invariant information about the level of discrimination associated with items for which there is a uniform level of discrimination.

It is likely that the variation in the sensitivity of the fit residual results from a combination of the systematic error and the nature of the maximum likelihood estimation process. By definition, maximum likelihood estimation yields estimates which maximise the likelihood of an observed data matrix given the resulting estimates. The greater the genuine likelihood of an observed data matrix given estimates, the greater the fit of data to the model. Therefore, *systematic error* arising from maximum likelihood estimation due to violations of the model should tend to operate so as to *minimize misfit* of data to the model in order to produce estimates which maximize the likelihood of the observed data matrix. It is apparent in Figure 4 that this is the case, although factors other than systematic error also affect the magnitude of the fit residual.

The degree of sensitivity of indices such as Outfit and Infit statistics similarly vary as a function of $\delta_i - \overline{\beta}$ in the presence of differential discrimination, since expected scores as shown in Equation (19) are used in the computation of such indices.

Due to the variation of the sensitivity of these kinds of fit index, it is advisable to use graphical information in order to identify items likely to contain the greatest systematic error. To illustrate, the item characteristic curve (ICC) of item 21 is shown in Figure 5. The simulated location of this item was $\delta_{21} = 2.0$ and the discrimination of the item set was $\rho_3 = 2$. The fit residual for this item was -2.313. The absolute value of this fit residual is considerably smaller than that of items 11, 13, 15, and 17, for which the degree of misfit was the same. However, the ICC shown in Figure 5 clearly shows there is substantial misfit.

It is apparent in Figure 5 that the empirical curve would intersect the line $E[X_{ni}] = 0.5$ at a location on the *x*-axis near the simulated item location. The error of the estimate for this item is $\hat{\delta}_{21} - \delta_{21} = 1.018$. The magnitude of error can be ascertained approximately from the ICC, as shown by the vertical lines which intersect the *x*-axis at the simulated and estimated item locations.

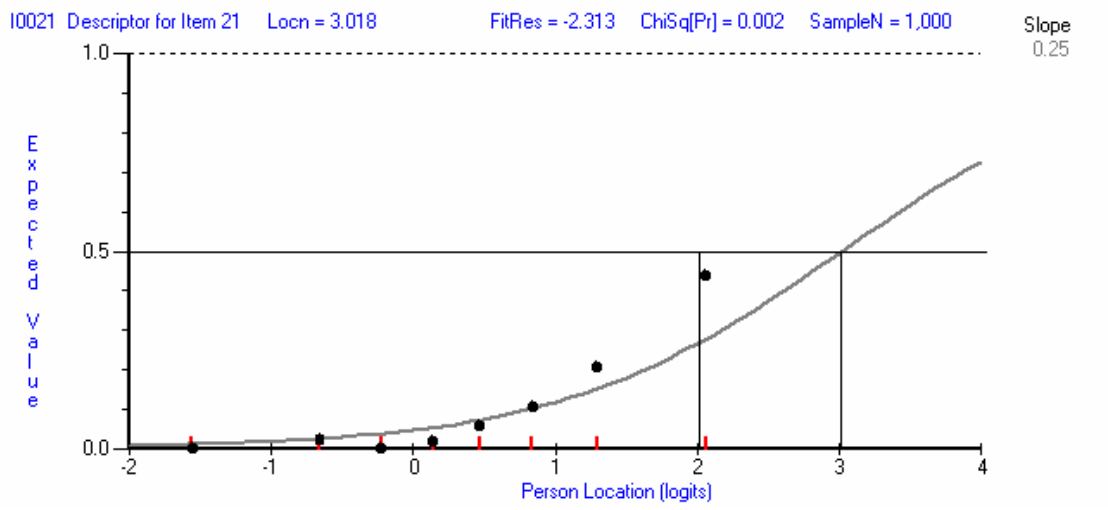


Figure 5: Item characteristic curve for item 21 in the preliminary simulation

A corollary of the variation of sensitivity of the fit residual as a function of $\delta_i - \overline{\beta}$ is that the order of fit residuals will not necessarily indicate the degree of misfit, and therefore the degree of systematic error.

To illustrate this point, a third set of data was simulated in the same fashion as the preliminary simulation, except with $\rho_1 = 0.67$ and $\rho_3 = 1.5$, rather than $\rho_1 = 0.5$ and $\rho_3 = 2.0$. The deviation of the location of item 15 of set 3 from the mean person location is $\delta_{i(15)} - \overline{\beta} \cong 0.3$. When the discrimination of set 3 items was simulated as $\rho_3 = 1.5$, the fit residual for item 15 of set 3 in RUMM2020 was reported as $Z_{15} = -4.604$. The absolute value of this residual is considerably larger than the residual for item 21 in the preliminary study in which $\rho_3 = 2.0$. The fit residual for this item in the preliminary study was –2.313 as shown at the top in Figure 5. This example illustrates that the ordering of fit residuals is not necessarily indicative of the degree of discrimination. In turn, this means that the order of fit residuals does not necessarily indicate the relative degree of systematic error contained within the item estimates. Consequently, visual representation of the accord between data and the model is an essential facet of evaluating fit and degree of systematic error.

Discussion

The results show that interrogation of data for evidence of differential discrimination is vital in order to minimize error resulting from the process of vertical equating. They also show the importance of using graphical evidence to complement fit indices in order to diagnose misfit associated with differential discrimination. It should be noted simulations show that, as would be expected, differential discrimination manifests as DIF when assessments are scaled jointly and ICCs or fit statistics are inspected, consistent with the manifestation of differential discrimination as DIF in Figure 2.

An implication of the results shown in this paper is that when items are trialled, information should be used to select common items least likely to have estimates that contain systematic error. It has been found in the WALNA program that deliberate

selection of link items in this fashion has provided considerably more effective vertical equating, as gauged by indices such as correlation coefficients and root mean squared differences between item estimates subsequent to equating.

It is stressed that it is not recommended post hoc adjustments are made based on the approximate linear relationship shown in Equation (7). This relationship is dependent on the distribution of persons and also the impact of violation of the model on the dispersion of all item scale locations. However, it may be useful to explore avenues for obtaining fit indices which are more stable across the range of the continuum for the purpose of more accurately diagnosing misfit associated with differential discrimination.

Conclusion

It has been shown that in the presence of differential discrimination, relevant items contribute information, for the purpose of vertical equating, which contains *systematic error*. Systematic error that occurs in the presence of high levels of discrimination operates in the opposite direction to systematic error which occurs in the presence of low levels of discrimination. The implication of these results is that differential discrimination carries particularly important consequences for the results of vertical equating. Accordingly, tests of fit sensitive to this violation should be given precedence over most other kinds of misfit for the purpose of vertical equating. Another kind of misfit which carries similar importance is DIF that occurs in the absence of differential discrimination; i.e. DIF that is not a manifestation of differential discrimination.

It has also been shown that indices such as Outfit and Infit are affected by systematic error in item estimates associated with differential discrimination. Specifically, the sensitivity of these indices to misfit varies as a function of the systematic error such that the magnitudes of indices tend to be smaller when there is a large distance between an item location and the mean person location. Consequently, it is vital that graphical evidence is used to complement indices to ascertain the degree of misfit and systematic error contained within item estimates.

References

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills: Sage Publications.

Andrich, D. & Luo, G. (1997-2003). *SimsRasch*. RUMM Laboratory, Perth, Australia.

Andrich, D. & Luo, G. (2003). Conditional Pairwise estimation in the Rasch model for ordered response categories using principle components. *Journal of Applied Measurement*, 4, 205-221.

Andrich, D., Sheridan, B. & Luo, G. (1997-2005). *RUMM2020*. RUMM Laboratory, Perth, Australia.

Baker, F. (1983). Comparison of ability metrics obtained under two latent trait theory procedures. *Applied Psychological Measurement*, 7(1), 97-110.

Baker, F. (1984). Ability metric transformations involved in vertical equating under item response theory. *Applied Psychological Measurement*, 8(3), 261-271.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Humphry, S.M. (2005). *Maintaining a common arbitrary unit in social measurement*. Ph.D. Thesis: http://wwwlib.murdoch.edu.au/adt/browse/view/adt-MU20050830.95143

Humphry, S.M. (2006). Maintaining a common unit in social measurement. Paper submitted to *Applied Psychological Measurement* in June 2006.

Kim, J. & Hanson, A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement*, 26, 255-270.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology, pp. 321-334 in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV*. Berkeley: University of Chicago Press, 1980.

Rasch, G. (1977). On Specific Objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *The Danish Yearbook of Philosophy*, 14, 58-93.

Suanthong, S., Schumacker, R.E., & Beyerlein, M. (2000). An investigation of factors affecting test equating in latent trait theory. *Journal of Applied Measurement*, 1(1), 20-27.

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

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis*. Chicago: MESA Press.

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