Common-Person Equating with the Rasch Model

Geofferey N. Masters
University of Melbourne

Two procedures, one based on item difficulties, the other based on person abilities, were used to equate 14 forms of a reading comprehension test using the Rasch model. These forms had no items in common. For practical purposes, the two procedures produced

equivalent results. An advantage of common-person equating for testing the unidimensionality assumption is pointed out, and the need for caution in interpreting tests of common-item invariance is stressed.

The traditional approach to equating two test forms X and Y is to administer both forms to the same group of persons and to define as "equivalent" two scores corresponding to the same percentile rank in that group (Angoff, 1971). Recently, a number of researchers have drawn attention to shortcomings of traditional procedures such as equipercentile equating, particularly for vertical equating tasks, and have investigated equating procedures based on latent trait theory methods (e.g., Divgi, 1981; Guskey, 1981; Holmes, 1982; Lord, 1975; Loyd & Hoover, 1980; Slinde & Linn, 1978, 1979a, 1979b; Wright, 1977).

A simple and effective latent trait method for equating scores on two psychological tests X and Y is to include in each test a set of items that can be calibrated in each form separately. These shared items provide a link between the two forms and so permit their relative difficulties to be estimated. "Common-item" equating is especially useful when there are no constraints on the way in which items can be selected and combined to construct different test forms; however, for a variety of equating tasks shared items are either impractical or inconvenient. Two standardized intelligence tests, or two reading comprehension tests based on different passages of writing, for example, would not usually contain shared items. In these situations, alternative methods must be used.

This paper describes and compares two latent trait methods for equating tests that do not contain shared items. Both methods depend on a group of persons who have taken both tests. The first method, Method A, uses the responses of this linking group to calibrate all $L_X + L_Y$ items simultaneously. The shift needed to equate the forms is then estimated as

$$t_{XY} = d_{\cdot X} - d_{\cdot Y} \quad , \quad (1)$$

where $d_{\cdot X}$ is the mean item difficulty estimate for Form X and $d_{\cdot Y}$ is the mean for Form Y.

APPLIED PSYCHOLOGICAL MEASUREMENT

Vol. 9, No. 1, March 1985, pp. 73-82

© Copyright 1985 Applied Psychological Measurement Inc.

0146-6216/85/010073-10\$1.75

The second method, Method B, first calibrates the items in each form using all available responses to that form. For each person n in the X-Y link, two ability estimates b_{nX} and b_{nY} are obtained from the separate calibrations of Forms X and Y. The shift needed to equate Forms X and Y is then estimated as

$$t_{XY} = b_{\cdot Y} - b_{\cdot X} \quad (2)$$

where $b_{\cdot X}$ and $b_{\cdot Y}$ are the mean ability estimates for persons in the X-Y link.

When data conform to the Rasch model, these two methods should produce equivalent results. If they do not, then it will be important to understand where they differ and why. If they do lead to the same conclusions, then it would be useful to know whether there are advantages or disadvantages to either method or circumstances under which one method is to be preferred to the other. The purpose of this study was to investigate the equivalence of these two methods in practice. Methods A and B were used to equate 14 forms of a reading comprehension test through 25 different common-person links. The results of these two equatings are then compared, and their relative merits discussed.

Method and Results

The 25 common-person links used in this study were part of a linking design developed for the calibration and equating of 14 *Tests of Reading Comprehension* (Education Department of Western Australia, 1983). These tests were constructed for use with students in their third to tenth years of school. Each test consists of a passage of writing approximately two pages in length accompanied by a "retelling" of this passage in different words. This retelling contains gaps corresponding to details contained in the original passage. Readers who understand the passage should be able to fill in these gaps using one or more of their own words. Each insertion is then scored either correct or incorrect.

The linking design is shown in Figure 1. The 14 forms are labeled A to N; the easier forms are on the left, the more difficult, on the right. Each line segment connecting a pair of forms represents a group of students who took that pair of forms. The tests were administered to 2,698 third- to tenth-grade students in Western Australia in November 1982. Most students took 2 of the 14 test forms.

The first step in the equating was to calibrate each form separately using the responses of all 300 (approximately) students who took that form. Misfitting items were identified and removed, reducing the average test length from 25 to 22 items. Of particular concern in the fit analysis were possible violations of the assumption of local independence. It had been recognized during test development that because all items in each form related to the same passage, special care would be needed to minimize possible dependencies among items. A statistical test of fit based on an analysis of responses to pairs of items and related to the statistic described by van den Wollenberg (1982) and Molenaar (1983) was used to test for violations of local independence.

The second step was to consider each pair of linked test forms X and Y, and to recalibrate the $L_X + L_Y$ items in these two forms simultaneously using the responses of students who took both forms. The mean item difficulties $d_{\cdot X}$ and $d_{\cdot Y}$ were then used to estimate the relative difficulties of Forms X

Figure 1
The Linking Structure

and Y. This was done for each of the 25 links, and the results are shown on the left of Table 1. Most of the $d_Y - d_X$ differences are positive, indicating that the test forms on the left of Figure 1 tend to be easier than the forms to their right, as intended. The exceptions are Form A, which was estimated to be somewhat more difficult than Forms B and C; Form I, which is more difficult than Form J; and Form K, which is slightly more difficult than Form L.

The shift needed to equate each pair of linked test forms X and Y was next estimated using Method B. The earlier calibration of each of the 14 tests provided two ability estimates b_{nX} and b_{nY} for each person n with nonperfect scores on both forms. The number of persons used in each link is shown in the middle of Table 1, and the estimated shift $b_X - b_Y$ is shown towards the right of the table.

An inspection of the values of $d_Y - d_X$ and $b_X - b_Y$ for each of the 25 common-person links shows that they differ on average by less than one-tenth of a logit. In other words, Methods A and B agree to within one-tenth of a logit on the amount that must be added to each Form Y ability estimate to bring it to the same scale as the Form X abilities. For comparison, the minimum standard error of the ability estimates themselves is in the vicinity of .5 logits. The 25 shifts estimated by Methods A and B appear to produce equivalent estimates of the shifts needed to equate these 14 test forms.

Nevertheless, there is one important difference between Methods A and B. The recalibration of Forms X and Y using the responses of the X-Y link group does not provide a particularly powerful check

Table 1
Shifts Estimated by Methods A and B
and Internal Consistency

Link X-Y	Method A				Method B				Internal Consistency S^2/V
	Number Persons	d_X	d_Y	$d_Y - d_X$	Number Persons	b_X	b_Y	$b_X - b_Y$	
A-B	146	.22	-.22	-.44	123	.61	1.03	-.42	1.20
A-C	49	.07	-.07	-.14	48	.44	.63	-.19	2.26
B-C	65	-.14	.16	.30	60	.87	.64	.23	1.33
B-D	78	-.79	.83	1.61	57	1.70	.29	1.41	1.18
C-D	59	-.45	.43	.88	58	.47	-.22	.70	1.60
C-E	81	-.63	.55	1.18	76	1.40	.28	1.11	1.64
D-E	88	-.16	.14	.30	76	1.09	.93	.16	1.14
D-F	72	-.32	.27	.59	63	.92	.42	.50	1.67
E-F	66	-.19	.39	.58	57	1.16	.70	.46	1.33
E-G	92	-.06	.06	.12	72	1.47	1.37	.10	1.00
F-G	77	-.13	.14	.27	69	1.48	1.21	.27	1.12
F-H	58	-.13	.15	.28	58	.80	.54	.26	1.66
G-H	73	-.19	.20	.40	67	1.13	.92	.21	1.70
G-I	91	-.79	.72	1.51	81	1.63	.24	1.40	1.26
H-I	87	-.69	.60	1.30	83	1.71	.50	1.21	1.28
H-J	78	-.16	.14	.30	71	1.02	.70	.32	1.92
I-J	83	.47	-.49	-.96	72	.52	1.32	-.80	1.21
I-K	90	-.31	.32	.63	86	.89	.20	.70	1.71
J-K	64	-.67	.67	1.33	59	1.32	-.03	1.35	1.97
J-L	65	-.53	.47	1.00	60	1.42	.36	1.06	1.29
K-L	82	.09	-.10	-.20	77	.63	.85	-.22	1.60
K-M	67	-.45	.45	.90	63	.92	.02	.89	.98
L-M	48	-.64	.73	1.38	46	1.51	.17	1.34	1.32
L-N	65	-.19	.18	.37	65	1.00	.63	.37	1.63
M-N	145	-.14	.16	.30	142	.86	.57	.29	1.65

Figure 2
Shifts Estimated by Methods A and B

on the extent to which these two forms define the same latent variable. This is especially true for tests such as these which are intended to measure the same latent variable and which have been constructed to fit the Rasch model individually. Additional global or item tests of fit at this stage are unlikely to reveal subtle differences between forms. Method B, on the other hand, provides an obvious and explicit check on the unidimensionality assumption underlying the attempt to equate these 14 forms. Because each person n in the X-Y link has two ability estimates, the set of $b_{nX} - b_{nY}$ differences provides a basis for assessing the extent to which Forms X and Y define the same latent variable.

For each of the 25 common-person links, abilities measured on Form X were plotted against abilities measured on Form Y. Figure 3 is a typical plot and provides a graphic check on the internal consistency of link J-L. The diagonal line in Figure 3 has been shifted 1.06 logits to the right of the identity line—the shift estimated for link J-L in Table 1—and a 95% confidence band based on the estimated errors of measurement has been added. Several students are outside this band, indicating that significantly different estimates of their reading abilities have been obtained on Forms J and L.

A statistical check on the internal consistency of each common-person link can be carried out using the observed and modeled variances of the $b_{nX} - b_{nY}$ differences. The observed variance of this difference is calculated as

$$S_{XY}^2 = \left[\frac{\sum_{n=1}^N (b_{nX} - b_{nY})^2 / W_{nXY}}{\sum_{n=1}^N 1 / W_{nXY}} \right] - \left[\frac{\sum_{n=1}^N (b_{nX} - b_{nY}) / W_{nXY}}{\sum_{n=1}^N 1 / W_{nXY}} \right]^2, \quad (3)$$

Figure 3
Abilities Estimated on Form J and Form L
for Students Taking This Pair of Forms

where $W_{nXY} = s_{nX}^2 + s_{nY}^2$ gives more weight to more precisely estimated differences, s_{nX} and s_{nY} are the errors of measurement associated with b_{nX} and b_{nY} , and N is the number of persons in the X-Y link. The modeled variance of this difference is calculated as

$$V_{XY} = \sum_{n=1}^N W_{nXY}/N \quad (4)$$

The ratio of the observed variance S^2 to modeled variance V can then be inspected for each common-person link to assess the internal consistency of that link. This internal consistency index is shown on the far right of Table 1 for each of the 25 links.

Values of this index are consistently greater than their expected value of 1, indicating that the observed $b_{nX} - b_{nY}$ differences are consistently more dispersed about their expected value, 0, than has been modeled. During the construction of these tests, items that displayed poor fit to the model were either removed or revised. This should have resulted in greater item homogeneity *within* each form. However, because each test is based on a different passage of writing, and because these tests cover different levels and types of writing, the question of unidimensionality *across* tests is especially relevant. Values of the internal consistency index suggest that these 14 tests do not define exactly the same reading comprehension variable.

An inspection of the plots for each of the X-Y links reveals a number of students with very discrepant ability estimates on the two forms that they took. These students performed significantly better on one

reading test than on the other. Their very discrepant performances may be a reflection of their different interests in these topics, their different approaches to different types of writing (e.g., informative, narrative) or, more likely perhaps, to their motivation on the day of the test (the two tests were administered on consecutive days). Whatever the reason, their very different levels of performance on the two tests make a meaningful estimate of their reading ability impossible.

The removal of students with particularly implausible test performances improved the internal consistency of each link. In most cases, the removal of two or three outliers resulted in an index very close to its expected value of one. Even for link A-C, the removal of the four most extreme outliers reduced the internal consistency index from 2.26 to 1.36. Although the removal of outliers to achieve better fit may be a questionable activity, it does make clear that much of the misfit evident from the large values of the internal consistency index can be traced to a few individuals with highly improbable test performances.

The next step in the equating is to bring together the 25 shift estimates $d_Y - d_X$ obtained from Method A to estimate the locations of these 14 tests on one scale of increasing difficulty. The validity of attempting to equate the 14 tests in this way is now open to question given the internal consistencies of the 25 links; however, to complete the comparison of Methods A and B, it was assumed that the 14 tests could be meaningfully equated. The equating procedure is perhaps best explained with a small example.

At the top of Table 2, shift estimates are shown for five hypothetical test forms V, W, X, Y, and Z. Form V has not been linked to either Form Y or Form Z, and Form W has not been linked to Form Z. This is similar to the design for the 14 reading tests in that the easier forms are not linked to the more

Table 2
Shift Estimates and Translation Constants

		Form				
	Form	V	W	X	Y	Z
Initial Estimates						
	V	-	.21	.40		
	W	-.21	-	.23	1.20	
	X	-.40	-.23	-	1.01	1.29
	Y		-1.20	-1.01	-	.34
	Z			-1.29	-.34	-
First Iteration						
	V	.00	.21	.40	1.41	1.72
	W	-.21	.00	.23	1.20	1.54
	X	-.40	-.23	.00	1.01	1.29
	Y	-1.41	-1.20	-1.01	.00	.34
	Z	-1.72	-1.54	-1.29	-.34	.00
Translation						
	Constant:	-.75	-.55	-.33	.65	.98
Second Iteration						
	V	.00	.21	.40	1.40	1.73
	W	-.21	.00	.23	1.20	1.53
	X	-.40	-.23	.00	1.01	1.29
	Y	-1.40	-1.20	-1.01	.00	.34
	Z	-1.73	-1.53	-1.29	-.34	.00
Translation						
	Constant:	-.75	-.55	-.33	.65	.98

difficult forms. The difference is that in the shift matrix for the reading tests there are many more missing links.

Initial estimates are now obtained for the missing links. From the available links, Form W is estimated to be .21 logits more difficult than Form V, and Form Y is estimated to be 1.20 logits more difficult than Form W. An initial estimate for the V-Y link is $.21 + 1.20 = 1.41$ logits. This has been inserted into the updated (first iteration) shift matrix in the middle of Table 2. Similarly, initial estimates have been obtained for missing links V-Z and W-Z. The mean of the five estimates in each column of this matrix is now obtained and provides the first set of form difficulty estimates (translation constants) $T_V, T_W, \dots, T_Z$. Form V is estimated to be the easiest of these five forms with difficulty $T_V = -.75$ logits, and Form Z the most difficult with difficulty $T_Z = .98$ logits. The five form difficulty estimates are automatically centered on zero.

These five estimates are now used to improve the estimates for the three missing links. The estimate for link V-Y, for example, is calculated as $T_Y - T_V = .65 - (-.75) = 1.40$. This improved estimate has been inserted into the revised (second iteration) matrix at the bottom of Table 2 together with similarly improved estimates for links V-Z and W-Z. The five estimates in each column are averaged again to obtain a new set of form difficulty estimates. This process is continued until values of the form difficulties $T_V, T_W, \dots, T_Z$ become stable to two decimal places. In this example, because the invented shift estimates at the top of Table 2 are reasonably consistent with one another, convergence has already occurred.

This iterative procedure was carried out on the 14×14 matrix of shifts ($d_{.Y} - d_{.X}$) to estimate a form difficulty for each of the 14 *Tests of Reading Comprehension*. The entire procedure was then repeated, this time using the matrix of shifts ($b_{.X} - b_{.Y}$) from Method B. This provides a second set of form difficulty estimates for the 14 reading tests. These two sets of estimates are plotted against each other in Figure 4.

It is clear from Figure 4 that, for practical purposes, Methods A and B produce equivalent estimates of the difficulties of these 14 reading tests. This is the result that is expected when data conform to the model. For these data, it seems not to matter whether the equating is performed using the difficulties (Method A) or using the abilities (Method B). Although there is still a doubt about the extent to which these 14 forms can be equated meaningfully, the observation that Methods A and B have produced more or less equivalent estimates of their form difficulties is encouraging.

Finally, because the items in Form B have been calibrated on link group A-B and also on link group B-C, Form B can be thought of as a common-item link between Forms A and C. The two difficulty estimates obtained for each item in Form B from these separate calibrations can be subtracted and the ratio of the observed to modeled variance of this difference can be used to check on the internal consistency of this common-item link. This has been done for each of the links in this design, and the results are shown in Table 3. Values of this ratio are approximately centered on their expected value of 1. About half the values in Table 3 are below their expected value, indicating that for these links, the two sets of difficulty estimates are *more* similar than has been modeled.

Discussion

The observation that, for practical purposes, Methods A and B produce equivalent estimates of the difficulties of these 14 test forms is significant. Equating procedures based on latent trait methods usually work with item difficulties rather than with person abilities because, in general, difficulties are more precisely estimated. However, common-item equating is not always possible or practical. Often, the tests being equated include one or more existing standardized tests; sometimes, responses to individual items have been lost and only total test scores are available. In these situations, common-person equating may

Figure 4
Relative Difficulties of the 14 Forms
Estimated by Methods A and B

be the only procedure that can be used. Even in circumstances in which there is a small number of link items, common-person equating may be preferable if the number of persons taking both forms is large.

This study shows that an equating based on Rasch abilities can produce the same results as an equating based on Rasch item difficulties. It also brings out an advantage of common-person equating: When form equating is based on abilities (Method B), an explicit and rigorous test of the underlying assumption of unidimensionality is a natural extension of the equating procedure. For these 14 reading comprehension tests, this test raised a question about the validity of the equating—a question that did not arise in the present application of Method A.

Finally, in this particular linking design, each form can be treated as a “common-item” link between two neighboring forms, and the internal consistency of this common-item link can be analyzed. However, the usual test of invariance for link items is not in itself a test of the extent to which the forms being equated define the same latent variable. The within-link fit statistics in Table 3, for example, are simply checks on the fit of items within each form to the model, and provide no indication of the extent to which different forms define the same variable. In general, when each form has been constructed to conform to the model prior to equating, further tests of item fit are unlikely to expose violations of the unidimensionality assumption. The most explicit check on this assumption is to test the invariance of the *ability* estimates for persons with measures on more than one form.

Table 3
Ratio of Observed to Modeled Variance for
"Common-Item" Analysis

Forms Linked	Linking Form													
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A,B			1.29											
A,C		1.04												
B,C	.94			1.69										
B,D			1.13											
C,D		.88			.64									
C,E				1.37										
D,E			1.18			.88								
D,F					.78									
E,F				.83			1.79							
E,G						.66								
F,G					1.01			.86						
F,H							1.10							
G,H						1.45			1.70					
G,I								1.44						
H,I							1.02			.74				
H,J									1.32					
I,J								1.42			1.04			
I,K										.47				
J,K									1.24			.91		
J,L											1.46			
K,L										.82			1.12	
K,M												.96		
L,M											1.36			1.23
L,N													1.01	
M,N														1.00

References

- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington DC: American Council on Education.
- Divgi, D. R. (1981). Model-free evaluation of equating and scaling. *Applied Psychological Measurement*, 5, 203-208.
- Education Department of Western Australia. (1983). *The Tests of Reading Comprehension: Users' guide*. Perth, Western Australia: Author.
- Guskey, T. R. (1981). Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. *Applied Psychological Measurement*, 5, 187-201.
- Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement*, 19, 139-147.
- Lord, F. M. (1975). *A survey of equating methods based on item characteristic curve theory* (ETS RB 75-13). Princeton NJ: Educational Testing Service.
- Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, 17, 179-193.
- Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika*, 48, 49-72.
- Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement*, 15, 23-35.
- Slinde, J. A., & Linn, R. L. (1979a). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement*, 16, 159-165.

- Slinde, J. A., & Linn, R. L. (1979b). The Rasch model, objective measurement, equating and robustness. *Applied Psychological Measurement*, 3, 437-452.
- van den Wollenberg, A. L. (1982). A simple and effective method to test the dimensionality axiom of the Rasch model. *Applied Psychological Measurement*, 6, 83-91.
- Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-116.

Acknowledgment

The author thanks Peter Hill, Research Branch, Education Department of Western Australia, for his assistance in this project.

Author's Address

Send requests for reprints or further information to Geoffrey Masters, Centre for the Study of Higher Education, University of Melbourne, Parkville, Victoria, Australia 3052.