DOCUMENT RESUME

ED 375 152 TM 022 101

AUTHOR Morrison, Carol A.; Fitzpatrick, Steven J.
TITLE Direct and Indirect Equating: A Comparison of Four Methods Using the Rasch Model.
INSTITUTION Texas Univ., Austin. Measurement and Evaluation Center.
REPORT NO RB-91-3
PUB DATE May 92
NOTE 17p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Computer Simulation; *Equated Scores; Error of Measurement; *Item Response Theory; Scaling; Statistical Studies; Test Format; Test Theory
IDENTIFIERS Anchor Tests; Calibration; Major Axis Equating; *Rasch Model

ABSTRACT

An attempt was made to determine which item response theory (IRT) equating method results in the least amount of equating error or "scale drift" when equating scores across one or more test forms. An internal anchor test design was employed with five different test forms, each consisting of 30 items, 10 in common with the base test and 5 to 10 in common with one or more other forms. Simulated data were generated for each using the Rasch model. Using one form as the base test, each of the others was equated directly to the base test and equated through one or more others to the base test. Equating methods examined were: (1) concurrent calibration; (2) equating constant procedure; (3) major axis procedure; and (4) fixed bs procedure. When equating error was assessed, it was found that concurrent calibration resulted in the least amount of equating error overall. When concurrent calibration is not feasible, results indicate that major axis equating results in the least amount of equating error when equating across one or more forms. (Contains 5 references, 6 tables, and 1 figure.) (Author/SLD)

Direct and Indirect Equating: A Comparison of Four Methods Using the Rasch Model

Carol A. Morrison and
Steven J. Fitzpatrick

RB-91-3 May 1992

Measurement and Evaluation Center

Research Bulletin

The University of Texas at Austin

Direct and Indirect Equating: A Comparison of Four Methods Using the Rasch Model

Carol A. Morrison and
Steven J. Fitzpatrick

RB-91-3 May 1992

MEASUREMENT AND EVALUATION CENTER
The University of Texas at Austin

Direct and Indirect Equating: A Comparison of Four Methods Using the Rasch Model

Carol A. Morrison and Steven J. Fitzpatrick

In practice, it is often necessary to equate a new test back to a base test through a series of linking test forms. In this study, an attempt was made to investigate which item response theory (IRT) equating method results in the least amount of equating error or "scale drift" when equating across one or more forms. An internal anchor test design was employed with five different test forms. Each test form consisted of 30 items, with 10 items in common with the base test and 5-10 items in common with one or more other forms. Simulated data were generated for each test form using the Rasch model.

Using one test form as the base test, each of the four other forms were equated directly to the base test and equated through one or more other forms to the base test. Four equating methods were examined: concurrent calibration, equating constant procedure, major axis procedure, and fixed bs procedure. Equating error was assessed using the root mean square difference between the known true score on a given test form (obtained using the item difficulties used to generate the simulated data) and the estimated true score on a given test form (obtained using the adjusted item difficulties from a given equating procedure).

It was found that concurrent calibration resulted in the least amount of equating error overall. However, a concurrent calibration of several test forms is not always feasible given practical considerations. In cases like this, the results of this study indicate that major axis equating results in the least amount of equating error when equating across one or more forms.

In practice, it is often necessary to equate a new test back to a base test through a series of linking test forms. Normally, the various test forms are linked through common items. Scores on a new form are equated to scores on a previous form which have been equated to scores on a previous form which have been equated to scores on the base test, and so on. Conceivably, the equating chain could go on forever, as long as the chain is linked back to the base test through common items.

Although an indirect equating design of this type is widely used, it is likely to result in an increasing amount of equating error or "scale drift" as the number of test forms in the equating chain increases. Studies are needed to ascertain the magnitude of scale drift that results for different equating procedures for indirect equating.

In a study pertinent to the present study, Petersen, Cook, & Stocking (1983) investigated scale drift for the verbal and quantitative sections of the Scholastic Aptitude Test (SAT) using both traditional and item response theory (IRT) equating methods. It was found that the concurrent calibration method produced the most stable equating results overall. The three linear equating methods (Tucker, Levine Equally Reliable, and Levine Unequally Reliable

models) performed adequately for reasonably parallel tests. However, when the test forms to be equated differed in content and length, the three-parameter logistic IRT methods (concurrent calibration, fixed bs procedure, and characteristic curve transformation procedure) lead to more stable equating results.

Although previous research has addressed equating situations that have arisen in commercial tests, no research to date has investigated the relative merits of various equating procedures when equating across one or more test forms that differ considerably in difficulty. Thus, the purpose of the present investigation was to study four equating methods that have been recommended for use with the Rasch model in the context of test forms that differ from one another in terms of difficulty.

Description of Equating Procedures Used

Four equating procedures were used in this study: (1) concurrent calibration, (2) fixed bs procedure, (3) equating constant procedure, and (4) major axis procedure. The four procedures are described below.

Concurrent Calibration

If two or more test forms share common items, they may be combined into a single data set and calibrated simultaneously. Because all the items are calibrated at the same time, all the item parameters are estimated on a common scale and no further equating is necessary (Wright & Stone, 1979).

Fixed bs Procedure

With the fixed bs procedure, the base test form is calibrated first. The new test form is then calibrated, holding the item difficulty parameters for the anchor items (items in common on the two forms) fixed at the estimates obtained from the calibration of the base test. Because the item difficulty parameters of the anchor items are fixed, the scale on the new test form is said to be the same as that of the base test form (Wright, Rossner, & Congdon, 1985).

Equating Constant Procedure

In the equating constant procedure, the item difficulties for the anchor test items on the new test are subtracted from their corresponding anchor item difficulties on the base test.

These differences are summed and divided by the total number of anchor test items. The mean difference is then added to each item difficulty on the new test form to obtain the adjusted item difficulties (Wright & Stone, 1979).

Major Axis Procedure

The major axis procedure uses the regression of the new test anchor item difficulties on the base test anchor item difficulties and the regression of the base test anchor item difficulties on the new test anchor item difficulties to obtain the major axis equation to be used in finding the base test equivalent for a given item difficulty on the new test. Once the major axis equation has been determined, all the item difficulties on the new test are simply substituted into the equation to get the adjusted item difficulties. The equations to be used to find the major axis are as follows (E. Jennings, personal communication, 1990):

Given the item difficulty values (BASE) on the base test and paired item difficulty values (NEW) on the new test, consider the regression equations

$$NEW = a + b*BASE$$

and

$$BASE = c + d*NEW$$

where a and c are regression constants and b and d are slopes.

Let $e = (c-a/b)/2$

$f = (d+1/b)/2$

$g = (a-c/d)/2$

$h = (b+1/d)/2$

The equation for the major axis to be used in finding the value of the new test that is equivalent to a given value of the base test is:

$$NEW = g + h*BASE$$

The equation for the major axis to be used in finding the value of the base test that is equivalent to a given value of the new test is:

$$BASE = e + f*NEW$$

True Score Equating

In item response theory, an estimated true score for a test form can be calculated by summing the probabilities of correct responses over all items. Once the item parameter estimates for two tests measuring the same ability, θ, have been put on a common scale, the estimated true score on one test is said to be equated to the estimated true score on a second test form if each corresponds to the same ability level (Lord, 1980). For example, if the estimated true score on Form A is expressed as:

$$ETRUA = P_i(\theta)$$

and the estimated true score on Form B is expressed as:

$$ETRUB = P_j(\theta)$$

then for a particular ability level (θ):

$$ETRUA = P_i(\theta) \text{ is equivalent to } ETRUB = P_j(\theta)$$

Method

In practice, it is often necessary to equate a new test back to a base test through a series of linking test forms. In this study, an attempt was made to investigate which equating method results in the least amount of equating error or "scale drift" when equating across one or more forms. An internal anchor test design was employed with five different test forms. Each test form consisted of 30 items, with 10 items in common with the base test and 5-10 items in common with one or more other forms (no more than 5 items in common with any one form).

Data Generation Procedure

Simulated data were generated for each test form using the Rasch model. The true item difficulties for each form were based on item difficulty parameter estimates obtained from a calibration of the actual test forms which had served as an impetus to this study. The mean true item difficulties on Forms B-E differed from the mean true item difficulty on Form A, the base test, by .4 to .8.

The true item difficulties for the base test (Form A) and an easier form (Form B) are presented in Table 1.

Table 1

True Item Difficulty Values for the Base Test, Form A, and an Easier Form, Form B

Form A Items	Common Items	Form B Items
	2.07	
	0.77	
	−1.61	
	1.50	
	−0.55	
	−1.00	
	−2.39	
	1.14	
	0.61	
	−0.73	
−1.37		−2.17
0.32		−0.48
−0.49		−1.29
0.33		−0.47
−0.18		−0.98
−1.02		−1.82
−1.08		−1.88
−2.32		−3.12
2.44		1.64
−1.44		−2.24
0.42		−1.38
−0.14		−1.94
0.59		−1.21
−1.28		−2.08
−0.91		1.71
1.05		0.25
0.33		0.47
−1.90		−1.70
0.02		−0.78
−0.40		0.20

The true item difficulties for the base test (Form A) and a more difficult form (Form E) are presented in Table 2.

Table 2

True Item Difficulty Values for the Base Test, Form A, and a More Difficult Form, Form E

Form A Items	Common Items	Form E Items
2.07		–0.22
0.77		–0.28
–1.61		–1.52
1.50		3.24
–0.55		–0.64
–1.00		1.20
–2.39		0.59
1.14		1.94
0.61		1.41
–0.73		1.07
–1.37		0.57
0.32		1.12
–0.49		1.31
0.33		1.13
–0.18		0.62
–1.02		1.22
–1.08		1.28
–2.32		–0.52
2.44		2.24
–1.44		1.64
	0.42	
	–0.14	
	0.59	
	–1.28	
	–0.91	
	1.05	
	0.33	
	–1.90	
	0.02	
	–0.40	

Two of the test forms were easier than the base test (Forms B & C) and two of the forms were more difficult than the base test (Forms D & E). Descriptive statistics for the true item difficulties for Forms A-E are presented in Table 3.

Table 3

Means and Standard Deviations of True Item Difficulties - Forms A-E

Form[a]	Mean	SD
A	–0.24	1.22
B	–0.65	1.40
C	–0.97	1.33
D	0.23	1.44
E	0.51	1.15

[a] n = 30 items for each form

The theta values used to generate response data for the five test forms were randomly drawn from a standard normal distribution using the RANNOR function in SAS. The probability of a correct response by each simulee to each item was calculated using the Rasch model and then compared to another number drawn randomly from a uniform distribution using the RANUNI function in SAS. If the probability of a correct response was greater than the uniform random number, the item was counted as correct (1). If the probability of a correct response was less than the random number, the item was counted as incorrect (0).

A total of 2,500 simulees were used, 500 for each test form. The data generation procedure resulted in perfect scores for some of the simulees. These simulees were eliminated because θ cannot be estimated in this case. Descriptive statistics for the simulated test scores for the five forms are presented in Table 4.

Table 4

Means and Standard Deviations of Raw Scores Forms A-E

Form	n	Mean	SD
A	500	16.33	5.79
B	499	17.75	5.28
C	499	19.72	4.88
D	500	13.42	5.18
E	496	12.02	5.39

Equating Design

Direct equating. Using Form A as the base test, Forms B, C, D, and E were each equated directly to Form A using four equating methods - concurrent calibration, equating constant procedure, major axis procedure, and fixed bs procedure.

Indirect equating. Form C was then equated to Form A through Form B using the equating constant procedure, major axis procedure, and fixed bs procedure. In other words, Form C was equated to the Form B item parameters which had already been equated to the Form A item parameters. Similarly, Form D was equated to Form A through both Forms C and B. Finally, Form E was equated to Form A through Forms D, C, & B. A diagram depicting the equating design used in this study is presented in Figure 1.

Figure 1

Equating design, Forms B-E equated directly to Form A and Forms C-E equated indirectly to Form A.

Direct Equating	Indirect Equating
B--->A	
C--->A	C--->B--->A
D--->A	D--->C--->B--->A
E--->A	E--->D--->C--->B--->A

An equating design such as the design used in this study is likely to occur when a test publisher, university testing center, or other testing organization must equate one or more new test forms to an existing test form through a series of linking forms.

Assessing Equating Error

Calculation of estimated true scores. Once the adjusted item difficulties were obtained for each form (using the different equating methods) for a direct equating and an equating through one or more forms, estimated true scores were obtained using the adjusted item difficulties from each equating procedure and thetas ranging from –3.5 to +3.5 in .10 intervals.

Calculation of known true scores. Known true scores were then calculated for each form using the item difficulties that were used to generate the simulated data (known item difficulties) for each form and thetas ranging from –3.5 to +3.5 in .10 intervals.

Based on true score equating, the known true score for a particular ability level (θ), obtained using the known item difficulties for the base test, Form A, is equivalent to the known true scores for that same ability level obtained using the known item difficulties for Forms B-E. This relationship holds because the known item difficulties for Forms B-E are on the same scale as the known item difficulties for Form A because the common items have exactly the same item difficulties.

Therefore, the estimated true scores for various ability levels obtained using the adjusted item difficulties from a particular equating procedure for a particular form may be compared to the corresponding known true scores for that form to see how well the equating procedure recovered the known true scores for that form, which are perfectly equivalent to the known true scores on the base test.

Statistic to assess scale drift. A root mean square difference statistic (RMSD) was computed to assess the amount of equating error for each procedure for direct equating and equating across one or more forms. This statistic, or a variation, is frequently employed in the statisical literature to represent total error. The RMSD results in values that are on the same scale as the estimated true scores. The formula for the RMSD statistic is as follows:

$$RMSD = \sqrt{\frac{\Sigma (t'_i - t_i)^2}{s_t^2}}$$

Where:

t'_i is the estimated criterion score
t_i is the criterion score
s_t^2 is the variance of the criterion score

In this case, t'_i is the estimated true score for a given ability level, θ_i, obtained using the adjusted item difficulties obtained from an equating procedure for a particular test form, and t_i is the known true score for the same ability level obtained using the known item difficulties used to generate the simulated data for the same test form.

A comparison was also made between the root mean square difference obtained from a direct equating to Form A vs. the root mean square difference from an equating through one or more forms for each test form. The direct equating RMSD was simply subtracted from the RMSD from equating across forms to obtain a measure of the effects of scale drift.

Results and Discussion

Direct Equating

Table 5 presents the unweighted mean square differences for Forms B-E equated directly to Form A and Forms B-E equated through one or more forms to Form A.

Table 5
Root Mean Square Differences for Various Equating Procedures:
Direct and Indirect Equating
(n=71)[a]

Direct Equating

	Equating Constant	Major Axis	Fixed bs	Concurrent Calibration
Form B onto A	1.0003	0.8586	0.9961	0.6656
Form C onto A	0.1814	0.1183	0.1609	0.0424
Form D onto A	0.0656	0.4297	0.0600	0.0755
Form E onto A	1.0859	1.0794	1.0929	0.3708

Equating Across One Form

	Equating Constant	Major Axis	Fixed bs
Form C onto A (Through B)	1.5037	1.1674	1.4207

Equating Across Two Forms

	Equating Constant	Major Axis	Fixed bs
Form D onto A (Through C & B)	1.6236	1.1820	1.5256

Equating Across Three Forms

	Equating Constant	Major Axis	Fixed bs
Form E onto A (Through D, C & B)	1.5002	0.9378	1.4797

[a] n = 71 because there are 71 thetas between –3.5 and +3.5 in .10 intervals

Based on the results reported in Table 5, concurrent calibration resulted in the least equating error of the four equating methods examined in this study. These results are consistent with the results reported by Petersen, Cook, & Stocking (1983). Although concurrent calibration was included under the heading of direct equating, it really should be in a category of its own since it is neither direct equating nor equating across forms. A concurrent calibration puts all test forms on the same scale automatically by calibrating the forms in a single calibration run.

With one exception (Form D onto A), the other three equating procedures resulted in more equating error than did the concurrent calibration procedure. The equating constant procedure and the fixed bs procedure resulted in virtually identical amounts of equating error when equating one form directly onto another form.

The results from the major axis procedure were more variable. In two cases (Form C onto A and Form E onto A), the results for the major axis procedure were very similar to the results reported for the equating constant procedure and the fixed bs procedure. However, in one case (Form B onto A), the major axis procedure performed somewhat better than the equating constant procedure and the fixed bs procedure and in one case (Form D onto A), the major axis procedure performed somewhat worse than these two methods.

Equating Across One or More Forms

The major axis procedure resulted in the least amount of equating error when equating across one or more forms, as reported in Table 5. Further, the major axis method also resulted in the smallest difference between the RMSD of a direct equating and the RMSD of an equating across forms, as reported in Table 6.

Table 6

Differences Between Root Mean Square Difference of a Direct Equating Versus Equating Across Forms for Various Equating Procedures (n=71)[a]

	Equating Constant	Major Axis	Fixed bs
Equating Across One Form			
Form C onto A (Through B)	1.5037	1.1674	1.4207
Form C onto A (Directly)	0.1814	0.1183	0.1609
Difference	1.3223	1.0491	1.2598
Equating Across Two Forms			
Form D onto A (Through C & B)	1.6236	1.1820	1.5256
Form D onto A (Directly)	0.0656	0.4297	0.0600
Difference	1.5580	0.7523	1.4656
Equating Across Three Forms			
Form E onto A (Through D, C & B)	1.5002	0.9378	1.4797
Form E onto A (Directly)	1.0859	1.0794	1.0929
Difference	0.4143	-0.1416	0.3868

[a] n = 71 because there are 71 thetas between -3.5 and +3.5 in .10 intervals

The equating constant procedure and the fixed bs procedure resulted in very similar amounts of equating error when equating across forms. Further, these two procedures also resulted in very similar differences between the RMSD of a direct equating and the RMSD of an equating across forms.

Discussion

In all cases but one (the major axis procedure for Form E onto A vs. Form E onto A through D, C, & B), equating across forms resulted in considerably more equating error than did equating one form directly onto another form. It may seem strange that the equating across three forms resulted in less equating error than did a direct equating, but it is possible that the equating errors cancelled each other out as Form E was equated to Form A (the base test) through two easier forms and one more difficult form.

Based on the results of this study, it appears that concurrent calibration produces the least amount of equating error or "scale drift" of the four equating procedures examined for test forms that differ in difficulty. These results are consistent with the results reported by Petersen, Cook, & Stocking, 1983, for the SAT.

Unfortunately, it is not always feasible to calibrate several test forms concurrently because different forms are given at different times of the year and calibrations must be done immediately to give examinees their scores. If practical considerations make a concurrent calibration impossible, it appears that the major axis procedure results in the least amount of equating error when equating across one or more forms that differ in difficulty. The major axis, equating constant, and fixed bs procedures result in very similar amounts of equating error when equating one test form directly onto another form.

In conclusion, the results of this study suggest that concurrent calibration results in the least amount of equating error when equating one or more test forms to a base test form and the major axis procedure results in the least amount of equating error when equating a test form through one or more other test forms to the base test form. Therefore, it is recommended that several test forms that differ in difficulty be calibrated concurrently if practical considerations permit; otherwise, the major axis procedure appears to be the best alternative.

References

Lord, F. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Petersen, N., Cook, L., & Stocking, M. (1983). IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 8, 137-156.

SAS Institute Inc. (1985). SAS user's guide: Basics, version 5 edition. Cary, NC: SAS Institute Inc.

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

Wright, B. D., Rossner, M., & Congdon, R. T. (1985). MSTEPS: A Rasch program for ordered item categories [Computer program manual]. Chicago: Benjamin Wright.