A Comparison of Two Observed-Score Equating Methods That Assume Equally Reliable, Congeneric Tests

Robert G. MacCann
New South Wales Department of Education

For the external-anchor test equating model, two observed-score methods are derived using the slope and intercept assumptions of univariate selection theory and the assumptions that the tests to be equated are congeneric and equally reliable. The first derivation, Method 1, is then shown to give the same set of equations as Levine's equations for random groups and unequally reliable tests and the "*Z* predicting *X* and *Y*" method. The second derivation, Method 2, is shown to give the same equations as Potthoff's (1966) Method B and the "*X* and *Y* predicting *Z*" method. Methods 1 and 2 are compared empirically with Tucker's and Levine's equations for equally reliable tests; the conditions for which they may be appropriately applied are discussed. *Index terms: Angoff's Design V equations, congeneric tests, equally reliable tests, Levine's equations (equally reliable), linear equating, observed-score equating, test equating, Tucker's equations.*

A frequently used model for equating the scores on different tests involves the use of an external-anchor test. In this model, Group 1 takes test X, Group 2 takes test Y, and both groups take a common test, Z. The problem to be considered is how to relate the score scales of X and Y. If X is more difficult than Y, then Group 1 is generally at a disadvantage with respect to Group 2 if the raw scores are used for the comparison. To prevent this, the scores on one test are usually transformed to make them comparable to the scores on the other. When this is done, the scores are said to be equated. The groups' relative performance on the common test provides the information for this statistical adjustment. This model was termed the common-item nonequivalent-populations design by Woodruff (1986) and Kolen and Brennan (1987), while Cook and Petersen (1987) simply referred to such designs as anchor test designs.

The form of test equating to be considered in this paper is known as observed-score equating to distinguish it from another form known as true-score equating: In the latter, it is the true scores hypothesized to underlie the observed scores that are equated. Angoff (1971) and Braun and Holland (1982) gave detailed accounts of observed-score equating.

Angoff listed two requirements that need to be satisfied in order for the scores to be strictly equated. The first is that the tests to be equated, X and Y, must measure the same psychological function. Taking

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 3, September 1989, pp. 263–276
© Copyright 1989 Applied Psychological Measurement Inc.
0146-6216/89/030263-14$1.95

the analogy of converting from degrees Fahrenheit to degrees Celsius, he pointed out that this conversion involves two measures that are both units of temperature:

> It makes no sense to ask for a conversion of, say, grams to degrees of Fahrenheit or from inches to pounds. Similarly, it makes little sense to ask for a conversion from a test of, say, verbal ability to a test of mathematical ability, or indeed across any two tests of different functions. (p. 563)

If the two tests are quite different in the psychological functions that they measure, then the conversion of units from X to Y may not be unique, but may vary according to the group on which it is based (see also Flanagan, 1964; Lindquist, 1964).

For linear observed-score equating, it is assumed that the distribution shapes of the tests to be equated are approximately equivalent. If this is the case, then for highly reliable tests it seems plausible to infer that the underlying true-score distribution shapes are approximately equivalent. In the ideal case, where the true-score distribution shapes are identical, the requirement that the tests measure the same psychological function implies that the true scores are perfectly correlated. The tests are then said to be *congeneric* (Jöreskog, 1971). In practice, linear equating is applied to cases where the true-score distributions may differ somewhat and where the two tests may be measuring slightly different functions, yielding an approximation to the congeneric assumption. However, severe departures from these conditions imply that the linear "equating" would be ineffective.

Mathematically, the congeneric assumption implies

$$X_t = aY_t + b \quad , \tag{1}$$

where X_t and Y_t are the true scores on X and Y, and a and b are constants with $a > 0$.

The second requirement for the strict equating of observed scores is that the tests to be equated should be equally reliable. Tests that are unequally reliable cannot be strictly equated. High-level examinees would be best advised to take the more reliable test as it would be more likely to make their abilities evident (Lord, 1977). [In fact, Angoff (1971) suggested that the term "calibrating" should be used when the scores on unequally reliable tests are converted to the same psychological scale to distinguish this case from the equating of equally reliable tests.] This second requirement may be expressed as

$$r_{xx} = r_{yy} \quad . \tag{2}$$

Although these two requirements are implicitly assumed in an ideal linear equating procedure, none of the equating methods in the literature makes explicit use of both of these assumptions. In this paper, both are used explicitly to produce two new equating method derivations. It is then shown how these derivations relate to existing equating methods based on different assumptions.

Method 1 Derivation

Some of the standard assumptions made in deriving equating methods involve the equivalence of two regression lines. Consider the regression line for predicting X from Z for the total group (Group 1 and Group 2 combined) and the regression line for predicting X from Z for Group 1.

Among the assumptions of univariate selection theory with respect to these two lines are the assumptions of equal slopes and equal ordinate intercepts (Angoff, 1971). The slopes assumption for the total group and Group 1 regression lines gives

$$b_{xz} = b_{xz1} \quad , \tag{3}$$

where

$$b_{xz} = r_{xz} \frac{S_x}{S_z} \tag{4}$$

and

$$b_{xz1} = r_{xz1}\frac{S_{x1}}{S_{z1}} \quad . \qquad (5)$$

S_x and S_z are the standard deviations of the total group on tests X and Z respectively; S_{x1} and S_{z1} are the corresponding standard deviations for Group 1.

The intercepts assumption gives

$$\overline{X} - b_{xz}\overline{Z} = \overline{X}_1 - b_{xz1}\overline{Z}_1 \quad . \qquad (6)$$

Substituting Equation 3 into Equation 6 and rearranging yields the estimated mean of the total group on test X:

$$\overline{X} = \overline{X}_1 + b_{xz1}(\overline{Z} - \overline{Z}_1) \quad . \qquad (7)$$

Substituting Equation 3 into Equation 4 and rearranging, the standard deviation of the total group on X is estimated:

$$S_x = \frac{b_{xz1}}{r_{xz}} S_z \quad . \qquad (8)$$

Similar assumptions are now made about the total group regression line for predicting Y from Z, and the Group 2 regression line for predicting Y from Z. These assumptions yield equations corresponding to Equations 7 and 8 (with Y terms replacing X terms) for Group 2.

Two scores are linearly equated if they correspond to equal standard-score deviates in the same population (Angoff, 1971). This gives the observed-score conversion line:

$$Y = \overline{Y} + \frac{S_y}{S_x}(X - \overline{X}) \quad . \qquad (9)$$

Equation 7, and its counterpart in Y terms, may be substituted for $\overline{X}$ and $\overline{Y}$ but the standard deviation ratio must still be estimated. This is obtained from Equation 8 and its counterpart in Y terms by dividing the latter by the former to obtain

$$\frac{S_y}{S_x} = \frac{b_{yz2}}{b_{xz1}}\frac{r_{xz}}{r_{yz}} \quad . \qquad (10)$$

The problem here is that the values of r_{xz} and r_{yz} are unknown because in practice the total group is not administered either X or Y. However, the ratio of these terms can be estimated using Equations 1 and 2.

Using the formula for the correction for attenuation of a correlation coefficient (Lord & Novick, 1968, p. 70) the following result is obtained:

$$r_{xz} = r_{x_tz_t}(r_{xx}r_{zz})^{1/2} \quad . \qquad (11)$$

A similar result is obtained for the correlation between Y and Z.

Dividing Equation 11 by its counterpart in Y terms and using Equation 2, the equal-reliabilities assumption, gives

$$\frac{r_{xz}}{r_{yz}} = \frac{r_{x_tz_t}}{r_{y_tz_t}} \quad . \qquad (12)$$

Equation 1, the congeneric assumption, implies that the true scores are linearly related. From Equation 1 it can be seen that the correlation of Z_t with X_t is equal to the correlation of Z_t with Y_t, as a linear scaling does not alter the value of a correlation, that is,

$$r_{x_tz_t} = r_{y_tz_t} \quad . \qquad (13)$$

Substituting Equations 12 and 13 into Equation 10 gives

$$\frac{S_y}{S_x} = \frac{b_{yz2}}{b_{xz1}} \quad . \tag{14}$$

Finally, substituting Equation 7, its counterpart in *Y* terms, and Equation 14 into Equation 9, and writing the expressions for the slopes in full detail, yields

$$Y = \left(\frac{r_{yz2}}{r_{xz1}}\frac{S_{y2}}{S_{x1}}\frac{S_{z1}}{S_{z2}}\right)X + \bar{Y}_2 + \left(r_{yz2}\frac{S_{y2}}{S_{z2}}\right)\left[(\bar{Z}_1 - \bar{Z}_2) - \frac{1}{r_{xz1}}\frac{S_{z1}}{S_{x1}}\bar{X}_1\right] \quad . \tag{15}$$

Method 2 Derivation

The above derivation has made assumptions about the regression lines for predicting *X* from *Z*. An alternate derivation is possible if the regression lines for predicting *Z* from *X* are used. For these regression lines it will be assumed that the Group 1 slope is equal to the total group slope and that the Group 1 ordinate intercept is equal to the total group intercept.

The slopes assumption gives

$$b_{zx} = b_{zx1} \quad , \tag{16}$$

where

$$b_{zx} = r_{zx}\frac{S_z}{S_x} \tag{17}$$

and so on. The intercepts assumption gives

$$\bar{Z} - b_{zx}\bar{X} = \bar{Z}_1 - b_{zx1}\bar{X}_1 \quad . \tag{18}$$

From Equations 16 through 18, the estimated mean and standard deviation of the total group on *X* are found in a manner similar to that for Method 1:

$$\bar{X} = \bar{X}_1 + \left(\frac{1}{b_{zx1}}\right)(\bar{Z} - \bar{Z}_1) \quad , \tag{19}$$

$$S_x = \left(\frac{1}{b_{zx1}}\right) r_{xz} S_z \quad . \tag{20}$$

A similar analysis, based on the regression lines for predicting *Z* from *Y* and using Group 2, yields the corresponding equations in *Y* terms for Group 2.

A corresponding analysis to that yielding Equation 14 for Method 1 gives

$$\frac{S_y}{S_x} = \frac{b_{zx1}}{b_{zy2}} \quad . \tag{21}$$

Substituting Equation 19, its counterpart in *Y* terms, and Equation 21 into Equation 9, and writing the expressions for the slopes in full detail, yields

$$Y = \left(\frac{r_{xz1}}{r_{yz2}}\frac{S_{y2}}{S_{x1}}\frac{S_{z1}}{S_{z2}}\right)X + \bar{Y}_2 + \left(\frac{1}{r_{yz2}}\frac{S_{y2}}{S_{z2}}\right)\left[(\bar{Z}_1 - \bar{Z}_2) - r_{xz1}\frac{S_{z1}}{S_{x1}}\bar{X}_1\right] \quad . \tag{22}$$

A comparison of Equations 22 and 15 shows that Method 2 differs from Method 1 in that the correlations in the latter are replaced by their reciprocals.

Methods Identical to Method 1

Although these derivations are new, the resulting equations are not. It is not uncommon for different theoretical approaches to lead to the same set of equations. Angoff (1971) illustrated this point in his survey of equating methods. He indicated that Lord's (1955) derivation results in exactly the same set of equations as Tucker's equations (attributed to L. R. Tucker in Gulliksen, 1950), "although the derivations of the two sets of equations are entirely different" (p. 580).

Although Angoff noted the equivalence of Lord's and Tucker's equations, he did not note that two of the other methods in his survey also give the same set of equations. These are Levine's equations for random groups and unequally reliable tests (originally in Levine, 1955), listed under Design IIIB, and the method known as "*Z* predicting *X* and *Y*," listed under Design VB1. It will now be shown that these two methods and Method 1 derived above all result in the same set of equations.

First, consider the *Z* predicting *X* and *Y* method. This uses a definition of equating different from the traditional definition embodied in Equation 9. For this method, scores on X and Y are defined as equivalent if they are predicted by the same score on Z. For Group 1, *X* is predicted from *Z* by

$$\hat{X}_1 = \bar{X}_1 + b_{xz1}(Z_1 - \bar{Z}_1) \quad . \tag{23}$$

For Group 2, *Y* is predicted from *Z* by

$$\hat{Y}_2 = \bar{Y}_2 + b_{yz2}(Z_2 - \bar{Z}_2) \quad . \tag{24}$$

When $Z_1 = Z_2$, $\hat{X}_1$ and $\hat{Y}_2$ are assumed equivalent. Rearranging Equations 23 and 24 yields

$$Z_1 = \bar{Z}_1 + \left(\frac{1}{b_{xz1}}\right)(\hat{X}_1 - \bar{X}_1) \tag{25}$$

and

$$Z_2 = \bar{Z}_2 + \left(\frac{1}{b_{yz2}}\right)(\hat{Y}_2 - \bar{Y}_2) \quad . \tag{26}$$

Setting $Z_1 = Z_2$, dropping the carets and group subscripts from the *X* and *Y* terms, and solving for *Y* yields Equation 15.

Next consider Levine's equations for random groups and unequally reliable tests, as presented in Angoff (1971, p. 579). Expressed in the notation of this paper, Angoff gave these equations as

$$Y = AX + B \quad , \tag{27}$$

where

$$A = \frac{b_{yz2}}{b_{xz1}} \quad , \tag{28}$$

and

$$B = \bar{Y} - A\bar{X} \quad , \tag{29}$$

where $\bar{X}$ and $\bar{Y}$ are as given in Equation 7 and its counterpart in *Y* terms. When Equation 27 is written out in detail, it is seen to be identical to Equation 15.

Methods Identical to Method 2

Angoff (1971) also considered the method known as "*X* and *Y* predicting *Z*" in his Design VC1. It will now be shown that this method gives the same set of equations as a method derived by Potthoff

(1966) known as Potthoff's Method B, and that both of these methods yield equations equivalent to Method 2.

First, consider the X and Y predicting Z method. This method also does not use the traditional definition of observed-score equating represented by Equation 9. It defines scores on X and Y as equivalent if they predict the same score on Z. For Group 1, Z is predicted from X by

$$\hat{Z}_1 = \overline{Z}_1 + b_{zx1}(X_1 - \overline{X}_1) \quad . \tag{30}$$

For Group 2, Z is predicted from Y by

$$\hat{Z}_2 = \overline{Z}_2 + b_{zy2}(Y_2 - \overline{Y}_2) \quad . \tag{31}$$

When $\hat{Z}_1 = \hat{Z}_2$, X_1 and Y_2 are assumed equivalent. Setting $\hat{Z}_1 = \hat{Z}_2$, dropping the group subscripts from the X and Y terms, and solving for Y yields Equation 22, the equation for Method 2.

Potthoff's Method B assumes that the conditional distributions of Z given X, and Z given Y, are normally distributed (Potthoff, 1966, p. 544). His equating relation for Group 1 is given by

$$h_1 = a_1 + b_1 X \quad , \tag{32}$$

where h_1 represents an "equated" score and the maximum likelihood parameter estimates are given by

$$b_1 = r_{xz1} \frac{S_{z1}}{S_{x1}} \tag{33}$$

and

$$a_1 = \overline{Z}_1 - r_{xz1} \frac{S_{z1}}{S_{x1}} \overline{X}_1 \quad . \tag{34}$$

Correspondingly, the "equated" score for Group 2 may be written

$$h_2 = \overline{Z}_2 + r_{yz2} \frac{S_{z2}}{S_{y2}} (Y - \overline{Y}_2) \quad . \tag{35}$$

Setting $h_1 = h_2$ and solving for Y, the resulting equation can be shown to be identical to Equation 22, the Method 2 equation.

Empirical Comparison

Previous Studies

Cope (1987) compared Angoff's Designs VB1 and VC1 with the two most commonly used equating methods, namely Tucker's equations and Levine's equations for equally reliable tests and dissimilar groups. In a single-link equating of pairs of test forms, he showed that Designs VB1 and VC1 often gave equated scores that fell between those obtained by the Tucker and Levine methods. In a second experiment involving cyclical equating of selected forms to themselves, he obtained bias and root mean squared error (RMSE) values that were very similar to the Tucker and Levine values. He concluded that these results were encouraging for the use of the Design V methods. These conclusions may be applied to the Method 1 and Method 2 equations derived in this paper, as it has been shown that Method 1 and Design VB1 share the same equations and that Method 2 and Design VC1 share the same equations.

Further evidence on the performance of Method 2 may be obtained from the fact that its equations have been shown to be the same as those of Potthoff's Method B. The latter has been extensively compared to the Tucker and Levine methods in Petersen, Marco, and Stewart's (1982) large-scale study. In their Table 10 (p. 94) the results are given for equating the verbal portion of the Scholastic Aptitude

Test (SAT-V) to itself through an external-anchor test similar in content and difficulty, an ideal equating design.

For random samples, Method 2 (Potthoff's Method B) performed at least as well as the Tucker and Levine methods, with 17 of the 18 equatings falling in the smallest error zone, compared to 16 for Tucker's equations and Levine's equations. For similar samples, Method 2 performed marginally better than Levine's equations, with both methods performing substantially better than Tucker's equations. For dissimilar samples, Levine's equations performed best, with Method 2 performing better than Tucker's equations. The results of the other external-anchor experiments, where the anchor test was dissimilar in content to the equated test, are difficult to interpret for the random and similar samples. However, for dissimilar samples, Levine's equations were clearly best, despite the lack of parallelism between the anchor and the equated test.

In the present research Methods 1 and 2 were further compared to Tucker's and Levine's equations for different types of samples, on tests that were not of the objective multiple-choice type, but that required some form of extended-response answer from the examinees. [The equations for the Tucker and Levine methods may be found in Angoff (1982), sections 5.1, 5.4.1, and 6.1 (Equation 53).]

Tests

The tests were assembled from questions developed for the Higher School Certificate Examination (HSC), a public examination that students may attempt in their final (sixth) year of secondary school in New South Wales, Australia. The courses on which the examination is based are studied in the fifth and sixth years of secondary school. Five different subject areas were chosen to provide a variety of question types: English, Mathematics, Geography, French, and Industrial Arts. In English and Geography the questions, in general, required answers in essay form. In Mathematics and Industrial Arts the questions usually required the solution to problems in an extended-response format, showing all steps involved in the solution. In French the response format varied considerably, including objectively scored responses, relatively unstructured essays in French, and assessments of conversations recorded on cassette tape.

Because the methods compared were linear equating methods, the tests in each experimental comparison were constructed to be as similar in difficulty as possible, as measured by the mean total group score expressed as a percentage. The resulting tests, though similar in difficulty, could not be considered parallel because corresponding pairs of parallel items were not available in the set of questions from which the tests were constructed.

Examinees

Two methods of examinee selection were employed. In the first (denoted "random groups"), the two groups were formed by random assignment. In the second (denoted "dissimilar groups"), the two groups differed substantially in average ability. For all subjects except Mathematics, the dissimilar groups were naturally occurring in that their members studied the subject at different levels, but shared a common syllabus and a common examination. For Mathematics, the two dissimilar groups were formed artificially by selection on a variable that was positively correlated with the Mathematics examination scores. An artificial variable was constructed by adding a normally distributed random error term to the sum of a further set of Mathematics questions; the resulting variable had a weak correlation with the tests in the equating experiments. Group 1 was then formed by taking approximately the top 10% of examinees on the selection variable; Group 2 comprised the remaining 90%. (These percentages were chosen to simulate the relative group sizes of the two highest level Mathematics courses examined in the HSC.)

The performance of the random groups is given in Table 1, where each test is denoted by a lower-case letter and a number. The tests are grouped in pairs of approximately equal difficulty. For example, e1 and e2 form a pair, e3 and e4 form a pair, and so on.

Table 1 also gives the performance of the dissimilar groups on a different series of tests. Each test for the dissimilar groups is denoted by an upper-case letter and a number. These tests were designed to differ in ''relative length'' as indicated by the maximum possible score that was available for each test. Within each subject area, by varying the pairings of forms of equal difficulty, the relative length ratio of the equating test to the anchor test could be varied; this was used to test whether the effectiveness of the equating methods interacted with this ratio.

Equating Error

If test X is equated to itself, then the criterion conversion line $Y' = PX + Q$ has a slope P of unity and an intercept Q of 0. Each equating method will generate, for each X, an equated score Y. Let the difference between an equated score and its criterion score be given by

$$d_i = Y_i - Y'_i \quad . \tag{36}$$

Petersen et al. (1982) defined the total error for equating as

$$\text{Total Error} = \frac{\Sigma f_i d_i^2}{n_1} \quad , \tag{37}$$

where f_i is the frequency of occurrence of X_i, n_1 is the number of examinees in Group 1, and the summation is over the group whose scores are transformed, Group 1. Petersen et al. showed that this total error incorporates both a bias component and a variance of the difference scores component.

Cope (1987) took as his main measure of equating error the square root of the total error, which he termed the RMSE. In the present paper, the RMSE was used as the measure of equating error. However, following Petersen et al. (1982), this was standardized by dividing by the standard deviation of the criterion scores (which in this case equals S_{x1}).

Results

Random groups. Table 2 gives the results of 26 experiments in which a test was equated to itself through an external-anchor test of similar difficulty, for random groups. In general, the RMSEs are low for all four methods. As a way of classifying such results, Petersen et al. (1982) divided the error into zones: In zone A the RMSE is less than or equal to .05; in zone B the RMSE is greater than .05 but less than or equal to .1; in zone C the RMSE is greater than .1 but less than or equal to .15; and in zone D the RMSE is greater than .15 but less than or equal to .2. Using this classification gives the following summary:

Tucker's: 18 classified as A, 6 B, and 2 C
Levine's: 17 A, 6 B, 3 C
Method 1: 17 A, 8 B, 1 C
Method 2: 14 A, 7 B, 2 C, 3 D.

For this dataset, Tucker's equations performed best, but there was little difference between the results of Tucker's equations, Levine's equations, and Method 1. Method 2 gave a slightly inferior performance.

Table 1
Test Performance Statistics for Random and Dissimilar Groups

Subject	Maximum Score	Group 1			Group 2		
and Test	Possible	*n*	Mean	SD	*n*	Mean	SD
Random Groups							
English							
e1	100	2479	54.15	11.88	2521	53.85	11.88
e2	100		50.04	13.82		50.02	14.15
e3	53	2479	25.69	8.29	2521	25.35	8.42
e4	33		15.92	5.61		15.78	5.70
Mathematics							
m1	48	643	23.88	8.94	5602	23.83	8.75
m2	48		24.01	10.26		23.91	10.13
m3	96	643	60.78	16.83	5602	60.60	14.47
m4	96		60.85	14.38		60.58	13.98
m5	48	643	38.37	7.15	5602	38.23	6.90
m6	48		38.25	6.64		38.23	6.56
Geography							
g1	80	502	45.34	11.30	2355	45.14	11.90
g2	120		69.25	14.59		69.71	14.60
g3	80	502	48.19	9.95	2355	48.69	9.95
g4	40		24.15	7.04		23.86	7.34
French							
f1	42	322	20.72	7.67	1236	21.32	7.69
f2	32		15.60	5.96		16.60	5.76
f3	108	322	58.80	17.38	1236	61.80	16.43
f4	92		49.42	15.71		51.60	15.16
f5	64	322	37.68	11.82	1236	39.61	11.23
f6	42		23.67	7.51		25.00	6.85
Industrial Arts							
i1	24	431	12.65	4.12	1742	12.52	4.04
i2	24		13.22	4.64		13.08	4.47
i3	52	431	32.50	8.88	1742	32.01	8.68
i4	48		30.20	6.86		30.17	6.71
i5	28	431	19.68	4.46	1742	19.57	4.44
i6	24		17.15	4.21		17.01	4.16
Dissimilar Groups							
English							
E1	160	2000	101.23	19.71	5000	84.55	20.23
E2	140		86.61	18.71		71.69	18.85
E3	100		60.56	14.00		50.03	13.99
E4	40		26.05	6.77		21.66	6.71
E5	40		23.37	6.23		19.48	5.92
Mathematics							
M1	168	635	128.28	15.11	5610	106.84	24.21
M2	144		110.74	13.64		91.30	22.28
M3	108		84.42	11.57		68.33	18.52
M4	72		56.85	8.49		45.34	13.65
M5	36		29.65	5.44		22.60	8.61
M6	36		28.00	4.72		22.79	7.09
Geography							
G1	160	494	101.01	16.53	2363	88.79	18.49
G2	120		77.64	12.69		68.25	13.95
G3	40		26.54	6.82		23.36	7.26
G4	40		23.36	7.48		20.54	7.52
G5	40		23.55	7.11		20.42	7.40
French							
F1	160	326	112.91	18.25	1232	84.64	23.16
F2	120		84.96	13.58		64.89	17.49
F3	80		54.06	10.82		39.48	12.30
F4	40		26.11	5.83		19.73	6.34
F5	40		26.80	5.38		20.64	6.39
Industrial Arts							
I1	80	439	57.38	8.86	1734	47.83	11.99
I2	40		28.45	4.39		24.49	5.73
I3	24		17.60	3.35		14.38	4.36
I4	24		17.54	3.53		14.34	4.39
I5	20		14.42	3.54		12.05	4.20

Table 2
Standardized RMSEs for the Random-Groups Experiments

Exper-	Tests					
iment	X	Z	Tucker's	Levine's	Method 1	Method 2
1	e1	e2	.0374	.0288	.0594	.0328
2	e2	e1	.0168	.0295	.0282	.0625
3	e3	e4	.0232	.0054	.0261	.0296
4	e4	e3	.0180	.0043	.0225	.0400
5	m1	m2	.0143	.0096	.0176	.0066
6	m2	m1	.0050	.0101	.0049	.0180
7	m3	m4	.0051	.0111	.0090	.0245
8	m4	m3	.0157	.0109	.0233	.0095
9	m5	m6	.0316	.0266	.0246	.0291
10	m6	m5	.0135	.0267	.0281	.0290
11	g1	g2	.0617	.0708	.0983	.0892
12	g2	g1	.0409	.0771	.0416	.1080
13	g3	g4	.0640	.1099	.0652	.1785
14	g4	g3	.0719	.0943	.0872	.1961
15	f1	f2	.0394	.1146	.0489	.1928
16	f2	f1	.1239	.1036	.1259	.0624
17	f3	f4	.0662	.0405	.0625	.0206
18	f4	f3	.0089	.0409	.0200	.0743
19	f5	f6	.0515	.0795	.0594	.1350
20	f6	f5	.1056	.0621	.0896	.0772
21	i1	i2	.0127	.0153	.0445	.0164
22	i2	i1	.0294	.0158	.0126	.0445
23	i3	i4	.0525	.0512	.0562	.0550
24	i4	i3	.0420	.0499	.0462	.0678
25	i5	i6	.0033	.0110	.0190	.0417
26	i6	i5	.0183	.0107	.0364	.0197

However, Method 2 (Potthoff's B) performed at least as well as Tucker's and Levine's methods in the Petersen et al. (1982) study for random groups.

Dissimilar groups. Table 3 gives the results of 38 experiments in which a test was equated to itself through an external anchor of similar difficulty, for groups that differed substantially in ability. The results are grouped by subject, with the ''length'' of the equating test to the anchor test being systematically varied.

For this dataset, the RMSEs were much higher than for the random groups case. In relative terms, the results were clear: Levine's equations gave the best results for all but two of the 38 experimental comparisons. However, the lack of parallelism between the equating test and the anchor test meant that for a number of the equatings, Levine's equations performed poorly. In general, the other methods performed very poorly when the groups were substantially different in ability. Tucker's equations and Method 1 gave a similar pattern of RMSE values, giving very high values when the equating test was much ''longer'' than the anchor test and giving relatively better results when the equating test was much ''shorter'' than the anchor test. The reverse held for Method 2, which performed more poorly when the equating test was much ''shorter'' than the anchor test.

Table 3
Standardized RMSEs for the Dissimilar-Groups Experiments

Exper-	Tests		"Length"				
iment	X	Z	Ratio	Tucker's	Levine's	Method 1	Method 2
1	E1	E5	4	.5325	.2187	.5337	.4801
2	E2	E5	3.5	.5075	.2476	.5033	.6052
3	E3	E5	2.5	.4586	.1781	.4517	.6563
4	E4	E5	1	.3961	.0598	.4416	.8010
5	E5	E4	1	.3819	.0526	.3791	.7713
6	E5	E3	0.4	.3010	.1042	.2923	.9071
7	E5	E2	0.29	.2898	.1259	.2700	.9716
8	E5	E1	0.25	.2507	.1028	.2266	.9235
9	M1	M6	4.67	.5839	.1456	.5839	.1530
10	M2	M6	4	.5909	.1229	.5909	.1581
11	M3	M6	3	.5943	.1265	.5947	.2085
12	M4	M6	2	.5619	.1032	.5625	.2376
13	M5	M6	1	.6002	.1139	.6126	.3761
14	M6	M5	1	.3360	.1069	.3422	.6616
15	M6	M4	0.5	.2214	.0892	.2115	.6094
16	M6	M3	0.33	.2125	.1006	.2012	.6563
17	M6	M2	0.25	.1665	.0960	.1469	.6296
18	M6	M1	0.21	.1761	.1125	.1582	.6400
19	G1	G3	4	.5861	.0403	.6426	.5544
20	G2	G3	3	.5922	.1043	.6436	.5760
21	G4	G5	1	.2474	.0681	.3872	.7949
22	G5	G4	1	.3265	.0689	.5057	.6314
23	G3	G2	0.33	.2369	.0718	.3086	1.4630
24	G3	G1	0.25	.2296	.0182	.2858	1.3744
25	F1	F5	4	.6122	.1059	.5978	.0819
26	F2	F5	3	.5797	.0367	.5688	.2243
27	F3	F5	2	.5225	.1355	.5103	.2217
28	F4	F5	1	.3811	.0898	.3737	.5544
29	F5	F4	1	.4468	.0963	.4444	.6078
30	F5	F3	0.5	.2041	.1164	.2127	.7102
31	F5	F2	0.33	.1695	.0329	.1637	.6522
32	F5	F1	0.25	.0716	.0870	.0733	.6893
33	I1	I5	4	.7374	.1242	.7303	.4186
34	I2	I5	2	.5715	.1373	.5632	.5420
35	I3	I4	1	.3587	.0533	.3793	.3752
36	I4	I3	1	.2856	.0513	.2885	.4157
37	I5	I2	0.5	.2396	.1208	.2516	.9553
38	I5	I1	0.25	.1618	.0910	.1911	1.1978

Discussion

Method 1 has been derived from assumptions concerning the slopes and intercepts of regression lines, and from the assumptions that X and Y are congeneric and equally reliable. It has been shown that the same equations are obtained from the *Z* predicting *X* and *Y* method and Levine's equations for random groups and unequally reliable tests. Of the latter two methods, *Z* predicting *X* and *Y* makes no overt assumptions, but uses an unorthodox definition of equating (i.e., scores on X and Y are equivalent if they are predicted by the same score on Z) that is not generally accepted in the test equating literature. As a general definition, there would seem to be no compelling reason why this should be more appropriate than a definition stating that X and Y are equivalent if they predict the same score on Z, the latter definition being the basis of the *X* and *Y* predicting *Z* method. However, the equations resulting from the *Z* predicting *X* and *Y* method can now be reconciled to the traditionally accepted definition of observed-score equating, as given in Equation 9, through the derivation of Method 1.

The identity of the equations resulting from Method 1 (which assumes that X and Y are equally reliable) and Levine's equations (for random groups and unequally reliable tests) was unexpected. Levine's equations, which were derived for the purpose of "equating" unequally reliable tests (Angoff, 1971, 1982; Levine, 1955) have now been derived for equally reliable tests, provided that X and Y are congeneric.

For Method 2, the equations have been shown to be equivalent to those resulting from the *X* and *Y* predicting *Z* method or Potthoff's Method B. As considered above, the *X* and *Y* predicting *Z* method rests on an unorthodox definition of equivalent scores; the Method 2 derivation enables the use of these equations on a more theoretically sound basis. Potthoff's method makes the somewhat restrictive assumptions that the conditional distributions of *Z* given *X* and *Z* given *Y* are normally distributed. Method 2 and Method 1 do not make distributional assumptions, but rather assumptions based on the equivalence of regression lines. These assumptions will now be considered and will be used to relate Methods 1 and 2 to two other closely associated methods, Tucker's equations and Lord's equations (Lord, 1950). A matrix showing the assumptions underlying these four methods is given in Table 4.

Method 1 assumes that the regression lines for predicting *X* from *Z* will have the same slopes for Group 1 and the total group, and the same intercepts for these groups. These assumptions are also made for Tucker's equations. The difference between the assumptions required for Tucker's equations and for Method 1 is that the former also require the regression line assumption that the standard errors of estimate be equal for Group 1 and the total group, whereas the latter requires that X and Y be congeneric and equally reliable. Thus, Tucker's equations and Method 1 are closely related through sharing both the slopes and intercepts assumptions.

Table 4
Assumptions Underlying Methods 1 and 2 and Two Other Closely Related Methods

	For subgroup and total group regression lines (i) the slopes are equal (ii) the intercepts are equal	
	Regression of Equating Tests on Anchor Test	Regression of Anchor on Equating Tests
Standard errors of estimate are equal for the subgroup and total group	Tucker's Equations	Lord's (1950) Equations
Equating tests (i) equally reliable (ii) congeneric	Method 1	Method 2

Method 2 assumes that the regression lines for predicting *Z* from *X* will have the same slopes for Group 1 and the total group, and the same intercepts for these groups. It is closely related to a method suggested by Lord (1950, p. 20) that makes the above two assumptions and a third assumption that the standard errors of estimate are equal for Group 1 and the total group. Method 2 replaces this third assumption with the assumption that X and Y are congeneric and equally reliable.

The slopes assumption for Tucker's equations (and Method 1) was investigated mathematically by Levine (1955). He showed that if Group 1 was selected on the basis of an external variable that was correlated with the common test, then the slopes assumption generally would not hold. However, the slopes assumption would hold if the correlation were 0, giving an effectively random selection. Levine, following Pearson (1903), also pointed out that all three regression line assumptions underlying Tucker's equations would hold if the selection variable was identical with the common test, Z (p. 59).

Thus, excluding the possibility that the selection variable is identical to Z, or closely related to Z, the slope and intercept assumptions would hold approximately only if Group 1 and Group 2 are similar in ability. Indeed, Angoff (1971) categorized Tucker's equations as being appropriate for "groups not widely different in ability" (p. 579). This would also seem a suitable categorization for Methods 1 and 2. The experimental results confirm this categorization, with Methods 1 and 2 performing well for random groups but having relatively large RMSEs for dissimilar groups.

As Table 4 shows, the difference between Tucker's equations and Method 1 is that the standard error assumption is replaced by the assumption that the tests to be equated are equally reliable and congeneric. Which of these assumptions is the most likely to be satisfied obviously depends on the data at hand. The standard error assumption is one of a set of coherent selection theory assumptions that are heavily dependent on the nature of the subgroups. If the subgroups are randomly formed, or very similar in average ability, then it seems likely that all three selection theory assumptions will be approximately satisfied.

The additional assumptions required by Method 1 (and Method 2) involve an extra characteristic: the nature of the tests to be equated. However, it could be argued that these assumptions are simply the embodiment of good equating practice. As discussed earlier, Angoff's (1971) criteria for equating are that the tests should be equally reliable and measure the same psychological function. The first of these criteria is directly assumed, while the second is closely related to, although not strictly identical with, the congeneric assumption for linear observed-score equating. Clearly, it would be unwise to apply Methods 1 and 2 in cases where the tests to be "equated" are of different psychological functions or are unequally reliable.

Methods 1 and 2 share the congeneric assumption for tests X and Y with Levine's equations for equally reliable tests (Woodruff, 1986). It is argued above, however, that for good equating practice, this is a rather weak assumption. Methods 1 and 2 differ from Levine's equations in that the latter also require that the anchor test be congeneric to both X and Y [or parallel if Angoff's (1953) formula is to be incorporated]. In contrast, none of the methods in Table 4 put any restrictions on the nature of the anchor test. It is this relatively strong assumption of Levine's equations that should enable the Levine method to deal more effectively with groups that differ substantially in ability.

A minor feature of the dissimilar-groups experiments concerns the interaction of Method 1, Method 2, and Tucker's equations with the "relative length" of the equating test to the anchor test. Because the groups were not selected on either of these tests, the selection theory assumptions would not be satisfied. However, for cases where Z was much "longer" (and hence more reliable) than X, it would be expected to correlate more highly with the selection variable than X would. In these cases, Method 1 and Tucker's equations would be expected to perform much better than for the reverse case where the equating test, X, was much "longer" than Z. This expectation was confirmed experimentally. Correspondingly, Method

2, which is based on the regression of *Z* on *X*, would perform relatively better when X was ''longer'' than Z and hence more closely related to the selection variable. This was also confirmed experimentally.

In summary, both theory and experiment indicate that Methods 1 and 2 will be effective only for random groups or groups similar in ability. When applied to such groups, the pooled evidence from a number of studies suggests that there is not a great deal of difference between Methods 1 and 2 and the more widely used methods, Tucker's equations and Levine's equations for equally reliable tests. For dissimilar groups, the equating error was much higher, in general, than for similar groups. Of the four methods, Levine's equations consistently gave the best results for dissimilar groups. However, the lack of parallelism between the anchor and equating tests caused large RMSEs for Levine's equations on several of the equating experiments.

References

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika, 18,* 1–14.

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.

Angoff, W. H. (1982). Summary and derivations of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 55–69). New York: Academic Press.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11,* 225–244.

Cope, R. T. (1987). How well do the Angoff Design V linear equating methods compare with the Tucker and Levine methods? *Applied Psychological Measurement, 11,* 143–149.

Flanagan, J. C. (1964). Obtaining useful comparable scores for non-parallel tests and test batteries. *Journal of Educational Measurement, 1,* 1–4.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 11,* 263–277.

Levine, R. S. (1955). *Equating the score scales of alternative forms administered to samples of different ability* (Research Bulletin 55-23). Princeton NJ: Educational Testing Service.

Lindquist, E. F. (1964). Equating scores on non-parallel tests. *Journal of Educational Measurement, 1,* 5–9.

Lord, F. M. (1950). *Notes on comparable scales for test scores* (Research Bulletin 50-48). Princeton NJ: Educational Testing Service.

Lord, F. M. (1955). Equating test scores—a maximum likelihood solution. *Psychometrika, 20,* 193–200.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14,* 117–138.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society, London, Series A, 200,* 1–66.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71–135). New York: Academic Press.

Potthoff, R. F. (1966). Equating of grades or scores on the basis of a common battery of measurements. In P. R. Krishnaiah (Ed.), *Multivariate analysis* (pp. 541–559). New York: Academic Press.

Woodruff, D. J. (1986). Derivations of observed score linear equating methods based on test score models for the common item nonequivalent populations design. *Journal of Educational Statistics, 11,* 245–257.

Author's Address

Send requests for reprints or further information to Bob MacCann, Leader, Systems Development & Research, Statutory Board Directorate, N.S.W. Department of Education, Box 460, P.O., North Sydney, N.S.W. 2059, Australia.