How Well Do the Angoff Design V Linear Equating Methods Compare With the Tucker and Levine Methods?

Ronald T. Cope
The American College Testing Program

Comparisons were made of three Angoff Design V equating methods and the Tucker and Levine Equally Reliable methods with respect to common item linear equating with non-equivalent populations. Forms of a professional certification test were equated with these five methods using (1) single-link equating of selected pairs of forms and (2) cyclical equating of selected forms to themselves, by means of equating chains. In the single-link equatings, raw score equivalents given by the Design V methods tended to fall between those obtained by use of the Tucker and Levine methods. The chain equatings produced similar estimated bias and estimated root mean squared error of score equivalents for the five different methods.

Consider the following situation: Two forms of a test, Form X and Form Y, are constructed to be similar in content and difficulty level. A formula is desired that will convert a raw score on Form X to its equivalent score on Form Y. There is a set of items, collectively called the "anchor test," that appears on both forms. The anchor test closely represents both forms in content and difficulty. Form X is administered to one examinee group, and Form Y is administered to another group. The two groups may differ in their distributions of the ability measured by the test.

Now consider the following three ways to obtain a formula that equates Form X scores to Form Y scores:

1. Equate Form X to anchor test U and equate U to Form Y as follows: Convert X and U raw scores to z scores. Convert Y and U scores to z scores. Set z scores on X and U equal and derive the equation converting X scores to the scale of U (i.e., $U = AX + B$). Similarly, set z scores on Y and U equal and derive the equation converting Y scores to the scale of U (i.e., $U = CY + D$). Set these two equations equal and solve for Y in terms of X. The result will be a formula that converts Form X raw scores to Form Y raw scores.
2. Find score X_0 on Form X and score Y_0 on Form Y, such that both X_0 and Y_0 are predicted by a given score U_0 on U. X_0 and Y_0 are the equated scores.
3. Find score X_1 on Form X and score Y_1 on Form Y, both of which predict the same score U_1 on anchor test U. X_1 and Y_1 are the equated scores.

Angoff (1984) outlined these procedures under "Design V: Other methods involving score data." Design V comprises the following procedures: (1) Forms X and Y equated to a common test, (2) Forms X

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 2, June 1987, pp. 143–149
© Copyright 1987 Applied Psychological Measurement Inc.
0146-6216/87/020143-07$1.60

and Y predicted by a common test, and (3) Forms X and Y predicting a common test. Angoff also outlined procedures for ''Design IV: Nonrandom groups—one test to each group, common equating test administered to both groups.'' Design IV methods include the commonly used Tucker and Levine linear methods and an equipercentile method. Either the Design IV or Design V methods can be used to establish correspondence for scores on Forms X and Y for the described situation. The present study deals only with the linear methods.

The Design IV linear methods are intended to produce transformed Form X scores that have the same mean and standard deviation as the Form Y scores for a particular group of examinees. Braun and Holland (1982) indicated that the first Design V method, which involves equating Forms X and Y to a common test, will not result in this property unless certain unusual relationships hold, and they suggested that this is a theoretical limitation of the Design V method. Despite the possible theoretical limitations of the Design V methods, the statistical assumptions for the Design IV methods are very strong, and the Design V methods might prove useful in situations in which the Design IV assumptions do not hold.

The purpose of this study was to make empirical comparisons, under common item linear equating, of the Levine Equally Reliable and the Tucker methods with the Angoff Design V methods. For detailed discussions of the Levine and Tucker methods, see Levine (1955/1956) and Braun and Holland (1982, pp. 23–24).

Method

Examinees

The examinee pool consisted of five groups of applicants to a professional certification program. Each of these five groups took one of five forms of the certification test in a particular administration. None of the examinees had previously taken any of the forms. Table 1 gives the number of first-time examinees who took each of the five forms, as well as means and standard deviations of their raw scores.

Measures

Measures were five forms of a multiple-choice professional certification test with a raw score range of 0 to 280. The test is administered twice a year at six-month intervals. Forms B, D, and E were administered in the spring; Forms A and C were administered in the fall. For each administration a new form is constructed, administered, and then retired. Except for retired items used in a sample test, no future, current, or past test items are released to examinees or otherwise made public.

Table 1
Examinee Groups

Certification Test		Raw Score	
Form	N	Mean	SD
A	10,450	158.47	34.91
B	22,562	162.85	35.94
C	11,272	168.08	39.31
D	11,041	156.47	36.24
E	23,450	160.55	36.25

Procedure

The following two ways of comparing the different equating methods were used:

1. Single-link comparisons: Results were compared of equating pairs of forms of the professional certification test under the five different linear methods.
2. Equating-chain comparisons (equating forms to themselves): Equatings around three cyclical chains of links among different forms of the certification test were compared. These chains are shown in Figure 1.

Procedure 1 assesses where equated raw scores produced by the three Design V methods stand in relation to those obtained through the Tucker and Levine methods. It permits no comparison of relative accuracy of the five methods because the raw score scales of two forms are expected to differ in an unknown way. If two forms were somehow known to have identical raw score scales, then there would be no need to try to equate the forms.

Procedure 2 compares the accuracy of the different methods by equating a selected form to itself, using a circular chain or cycle of equatings among forms. If an equating method were free of systematic and random error, the equating function $Y = AX + B$ would have a slope (A) of 1 and an intercept (B) of 0.

Equating functions for the three Design V methods follow; see Angoff (1984) for their derivation.

Method V1. Linearly equate Form X to anchor test U; then equate U to Form Y.

$$Y = \frac{s(Y)s_x(U)}{s(X)s_y(U)}X + M(Y) - \frac{s(Y)s_x(U)M(X)}{s(X)s_y(U)} + \frac{s(Y)M_x(U)}{s_y(U)} - \frac{s(Y)M_y(U)}{s_y(U)} \quad , \tag{1}$$

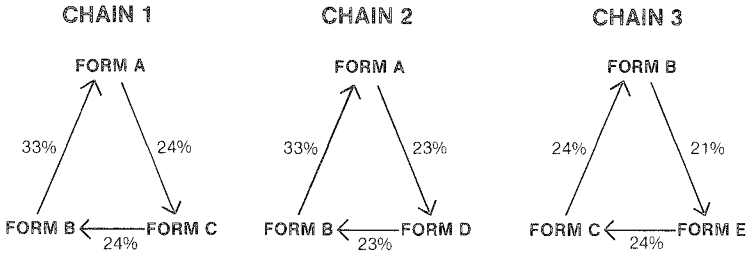
where $M_x(U)$ is the mean on anchor test U of the group taking Form X, and $s_x(U)$ is the standard deviation of this group on U. These quantities are similarly defined for the group taking Form Y.

Method V2. (Forms X and Y predicted by a common test.) For this method and method V3, correlations of the anchor test U with Forms X and Y are involved:

$$Y = \frac{r(Y,U)s(Y)s_x(U)}{r(X,U)s(X)s_y(U)}X + M(Y) + \frac{r(Y,U)s(Y)s_x(U)}{r(X,U)s(X)s_y(U)}\left(\frac{r(X,U)s(X)M_x(U)}{s_x(U)} - M(X)\right) - \frac{r(Y,U)s(Y)M_y(U)}{s_y(U)} \quad , \tag{2}$$

where $r(X,U)$ and $r(Y,U)$ are the respective correlations of anchor test U with Form X and Form Y.

Figure 1
Chains Used to Compare Accuracy of Equating
Using Tucker, Levine, and Angoff Design V Equating Methods
Arrows Indicate Equating Links; a Percentage Adjacent to an Arrow Gives the Number of Equating Items as a Percentage of the Number of Items in Each of the Equated Forms



Method V3. (Forms X and Y predicting a common test.)

$$Y = \frac{r(X,U)s(Y)s_x(U)}{r(Y,U)s(X)s_y(U)} X + M(Y) + \frac{-M_y(U) - \dfrac{r(X,U)s_x(U)M(X)}{s(X)} + M_x(U)}{\dfrac{r(Y,U)s_y(U)}{s(Y)}} \quad . \tag{3}$$

Comparison of Equations 1, 2, and 3 reveals the following relationships:

1. If anchor test U correlates perfectly with both Form X and Form Y, then $r(X,U) = r(Y,U) = 1$, and Method V3 reduces to Method V1.
2. If U correlates equally with X and Y, then

$$\frac{r(X,U)}{r(Y,U)} = \frac{r(Y,U)}{r(X,U)} = 1 \quad , \tag{4}$$

and all equating functions have the same slope.

Measures of Equating Error

Two statistics, estimated bias and estimated weighted root mean squared error of equating, were used to compare results under the different methods. Estimated bias (BIAS) is defined as

$$\text{BIAS} = \overline{X}' - \overline{X} \quad , \tag{5}$$

in which $\overline{X}$ is the mean raw score of a given form, and $\overline{X}'$ is the mean of the raw score equivalents obtained by equating the form to itself around a chain. Estimated root mean squared error (RMSE) is defined by the expression

$$\text{RMSE} = \left[\frac{\sum N_i (X_i' - X_i)^2}{\sum N_i} \right]^{1/2} \quad , \tag{6}$$

in which X_i is the ith raw score on the form, X_i' is its raw score equivalent, and N_i is the number of examinees who obtained the score X_i. When there are as many examinees as in this study, RMSE is calculated more easily by the expression

$$\text{RMSE} = [(m-1)^2 \, \text{Var}(X) + (\text{BIAS})^2]^{1/2} \quad , \tag{7}$$

where m is the slope of the equating function $X' = mX + b$, and $\text{Var}(X)$ is the variance of X. Note that BIAS contributes to RMSE.

Results

Single-Link Equating

Table 2 gives the results of equating Form A to Form B, Form B to Form C, and Form B to Form D under the five different methods. Raw score equivalents under the Angoff Design V methods showed a strong tendency to be intermediate in value between the corresponding Tucker and Levine equivalents. Also, Design V equating function slopes and intercepts fell between the respective Tucker and Levine slopes and intercepts.

Chain Equating

Table 3 shows the results of equating Form A to itself around Chains 1 and 2 and equating Form B to itself around Chain 3. Figure 1 depicts these chains, and Table 4 gives estimates of bias and root mean squared error. The departures of equivalent raw scores from original raw scores were quite similar in the

Table 2
Raw Scores and Raw Score Equivalents from Equating
Form A to Form B, Form B to Form C, and Form B to Form D

	Raw Score Equivalent				
Raw Score	Tucker	Levine	Angoff V1	Angoff V2	Angoff V3
		Form A Equated to Form B			
Form B					
0	9.33	11.48	**10.35**	**10.17**	**10.55**
40	47.17	48.97	**48.02**	**47.79**	**48.28**
80	85.00	86.46	**85.69**	**85.42**	**86.00**
120	122.83	123.96	**123.37**	**123.04**	**123.72**
160	160.67	161.45	**161.04**	**160.67**	161.45
200	198.50	198.94	**198.72**	198.29	199.17
240	236.33	236.43	**236.39**	235.92	236.90
280	274.17	273.92	274.07	**273.54**	274.62
		Form B Equated to Form C			
Form C					
0	-4.06	-9.48	**-6.51**	**-5.02**	**-8.05**
40	34.69	30.33	**32.71**	**33.95**	**31.42**
80	73.43	70.14	**71.93**	**72.92**	**70.90**
120	112.17	109.95	**111.15**	**111.89**	**110.37**
160	150.92	149.76	**150.38**	**150.86**	**149.85**
200	189.66	189.57	**189.60**	189.83	189.32
240	228.41	229.37	**228.82**	**228.80**	**228.80**
280	267.15	269.18	**268.04**	**267.77**	**268.27**
		Form B Equated to Form D			
Form D					
0	-1.47	-6.24	**-3.59**	**-4.96**	**-2.25**
40	40.29	36.68	**38.68**	**37.69**	**39.65**
80	82.05	79.59	**80.96**	**80.34**	**81.55**
120	123.81	122.51	**123.23**	**122.99**	**123.45**
160	165.57	165.43	**165.50**	165.64	165.35
200	207.33	208.34	**207.77**	**208.29**	**207.25**
240	249.10	251.26	**250.05**	**250.94**	**249.15**
280	290.86	294.18	**292.32**	**293.59**	**291.05**

Note. Entries in boldface are values that fall between corresponding Tucker and Levine raw score equivalents.

Tucker, Levine, and Design V methods. In Chain 1, RMSE ranged from .67 for the V1 method to .70 for the Tucker method. BIAS of the Levine and V3 methods were somewhat lower in Chain 1 than those of the other methods. In Chain 2, RMSES of the five methods ranged from .67 (Levine and V1) to .73 (Tucker). As in Chain 1, the Levine and V3 methods in Chain 2 showed somewhat lower estimated bias than the other three methods. In Chain 3, however, the Levine and V3 methods showed somewhat higher BIAS and RMSE than the other methods.

In these three chains, the tendency in the single-link equatings for Design V equivalents to lie in value between corresponding Tucker and Levine equivalents appeared strongly only for method V1. Method V3 produced a majority of such intermediate equivalents only in Chain 2.

Discussion

Results of the single-link and cyclical equating procedures show that, in the situations studied here, the Angoff Design V methods yield raw score equivalents close to those of the Tucker and Levine

Table 3
Equivalent Raw Scores Obtained from Applying the Tucker Method, Levine Method, and Angoff Designs V1, V2, and V3, to Equating Chains 1, 2, and 3

	Raw Score Equivalent				
Raw Score	Tucker Method	Levine Method	Angoff V1	Angoff V2	Angoff V3
Form A		Chain 1			
0	-0.49	-1.76	**-1.04**	-0.35*	-1.76
40	39.80	38.80	**39.37**	39.90*	**38.81**
80	80.09*	79.37	**79.77**	80.16	79.37
120	120.37	119.93	**120.18**	120.41	119.94*
160	160.66	160.50*	**160.59**	160.66	160.50*
200	200.95	201.06	**201.00**	200.91*	201.07
240	241.23	241.63	**241.40**	241.16*	241.64
280	281.52	282.19	**281.81**	281.41*	282.20
Form A		Chain 2			
0	-0.29	-1.95	**1.00**	-0.27*	**-1.75**
40	39.96*	38.64	**39.40**	39.96*	**38.81**
80	80.21	79.24	**79.79**	80.20*	**79.37**
120	120.46	119.83	**120.19**	**120.43**	119.93*
160	160.71	160.43*	**160.59**	**160.67**	**160.49**
200	200.96	201.03	**200.99**	200.90*	201.05
240	241.21	241.62	**241.38**	241.14*	**241.61**
280	281.46	282.22	**281.78**	281.37*	**282.17**
Form B		Chain 3			
0	2.65	1.80	**2.34**	3.27	1.43*
40	43.04	42.62	**42.90**	43.51	42.33*
80	83.42	83.43	83.47	83.75	83.22*
120	123.81*	124.25	**124.04**	**123.99**	**124.12**
160	164.19*	165.06	**164.60**	**164.23**	**165.02**
200	204.58	205.88	**205.17**	204.47*	205.91
240	244.96	246.70	**245.74**	244.71*	246.81
280	285.34	287.51	**286.31**	284.96*	287.70

Note. Entries in boldface are values that fall between corresponding Tucker and Levine equivalents. The symbol "*" denotes equivalents closest in value to corresponding original raw scores.

methods. In the cyclical or chain equatings of forms to themselves, the five methods yield comparable estimates of bias and root mean squared error. In the single-link equatings, Design V raw score equivalents tend to fall between the Tucker and Levine equivalents.

Obtaining similar results for three single-link equatings and three equating chains gives some indication that comparable results would be obtained for forms of other tests. If the five equating methods had been compared for only one link or one chain, the generality of the findings would be questionable. However, the five methods could be tried on forms of other tests as a check on the current findings.

These results are encouraging for use of the Design V linear equating methods, which use fewer restrictive assumptions than the Tucker and Levine methods. The Tucker method, for example, requires the untestable assumption that a single line describes the linear regression of test form scores on anchor test scores both for the group taking one form and the group taking the other form (Braun & Holland, 1982). The Levine (1955/1956) method assumes a perfect correlation between anchor test true scores and true scores of at least one of the forms to be equated. Often in practice there is reason to doubt that these assumptions hold. In such cases practitioners may prefer to use one of the Design V methods.

Table 4
Estimates of Bias and Root-Mean-Squared Error

	Equating Method				
Estimate	Tucker Method	Levine Method	Angoff V1	Angoff V2	Angoff V3
			Chain 1		
BIAS	0.65	0.47	0.57	0.65	0.48
RMSE	0.70	0.68	0.67	0.69	0.69
			Chain 2		
BIAS	0.70	0.41	0.57	0.66	0.47
RMSE	0.73	0.67	0.67	0.69	0.69
			Chain 3		
BIAS	4.22	5.12	4.64	4.25	5.08
RMSE	4.23	5.17	4.67	4.25	5.14

References

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton NJ: Educational Testing Service. [Reprint of chapter in R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 508–600). Washington DC: American Council on Education, 1971.]

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.

Levine, R. (1956). Equating the score scales of alternate forms administered to samples of different ability (Doctoral dissertation, Syracuse University, 1955). *Dissertation Abstracts International, 16*, 1842A.

Acknowledgments

The author thanks the Editor, anonymous reviewers, Bradley Hanson, Deborah Harris, Michael Kane, and Michael Kolen for comments and suggestions for revision of earlier versions of this article.

Author's Address

Send requests for reprints or further information to Ronald T. Cope, The American College Testing Program, P.O. Box 168, Iowa City IA 52243, U.S.A.