A Comparison of Three Linear Equating Methods for the Common-Item Nonequivalent-Populations Design

David J. Woodruff
American College Testing Program

Three linear equating methods for the common-item nonequivalent-populations design are compared using an analytical method. The analysis investigated the behavior of the three methods when the true-score correlation between the test and anchor was less than unity, a situation that may occur in practice. The analysis is graphically illustrated using data from a test equating situation. Conclusions derived from the analysis have implications for the practical application of these equating methods. *Index terms: congeneric model, Levine equating method, linear equating, Tucker equating method.*

Linear equating methods for the common-item nonequivalent-populations design have been derived or discussed by several authors: Gulliksen (1950), Levine (1955), Angoff (1982, 1984), Braun and Holland (1982), Kolen (1985), Woodruff (1986), and Kolen and Brennan (1987). Angoff (1984) referred to this design as design IV—nonrandom groups. Under this design, a new test X is given to group 1, an old test Y is given to group 2, and a usually shorter anchor test V is given to both groups. The anchor test V may comprise a scorable part of the tests, or it may not contribute to examinees' scores; these are respectively referred to as the inclusive and exclusive anchor situations.

One method commonly used in practice for linear equating under the common-item nonequivalent-populations design is Tucker's equally reliable method (Angoff, 1982, 1984; Gulliksen, 1950; Kolen, 1985; Kolen & Brennan, 1987). Another commonly used method is Levine's major-axis equally reliable procedure (Angoff, 1982, 1984; Kolen & Brennan, 1987; Levine, 1955). There are two versions of this second procedure. The usual version presented by Angoff (1982, 1984) and denoted herein as the Angoff-Levine method incorporates Angoff's (1953) reliability formula, whereas the original version (Levine, 1955), denoted herein as the Congeneric-Levine method, does not.

Woodruff (1986) derived both versions from explicitly stated test score models. He showed that the Angoff-Levine method makes stronger assumptions than the Congeneric-Levine method. More specifically, the Congeneric-Levine method assumes that the true scores on X, Y, and V all correlate unity, but the Angoff-Levine method assumes in addition that the true-score variances and error-score variances of X, Y, and V satisfy a certain proportionality constraint. Under the Angoff-Levine model, a continuous version of the Spearman-Brown formula is true, but the formula does not hold under the Congeneric-Levine model. Angoff (1953) made the same assumption but expressed it differently. He assumed that there exists a parameter n_{yv} such that Y is composed of n_{yv} tests all parallel to V. As a consequence, the Angoff-Levine method is slightly easier to implement in

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 3, September 1989, pp. 257–261

0146-6216/89/030257-05$1.50

that it does not require an explicit estimate of the anchor's reliability.

Tucker's method makes assumptions about observed score regressions; the Angoff-Levine and Congeneric-Levine methods make assumptions about true-score relationships. Tucker's method is based on a linear regression model; the Angoff-Levine and Congeneric-Levine methods are based on linear structural models. According to Angoff (1984), who cited Levine (1955) and Lord (1960), Tucker's method, which does not explicitly assume that X, Y, and V have true-score correlations of 1, is most appropriate for situations in which the two groups show no more than small differences in mean and variance on the anchor, whereas the Angoff-Levine method (and the Congeneric-Levine method as well) may accommodate larger differences so long as the true scores on the tests and anchor correlate unity.

This paper compares through analytical and empirical means the performance of the three methods as the covariance between the tests and the anchor varies, with the goal of learning how the Congeneric-Levine and Angoff-Levine methods behave when the assumption underlying their derivations of unity of true-score correlations is violated. In the common-item nonequivalent-populations design, the tests and anchor are supposed to be constructed so that their true-score correlations are unity. In practice, however, situations may arise where this condition is not fulfilled but the tests must still be equated. This investigation therefore has important implications for practitioners attempting to select the most appropriate method for such situations.

Klein and Jarjoura (1985) undertook a similar investigation using only empirical methods. They noted that the Angoff-Levine method was more sensitive than Tucker's to a lack of content balance between the tests and the anchor. The present study suggests an explanation for their finding which indicates that the performance of the Congeneric-Levine method should be more similar to the Tucker method than to the Angoff-Levine method as the covariance between the test and anchor decreases.

Analysis

The analysis begins with the exclusive anchor situation. Later, it is shown how the results for the exclusive situation easily generalize to the inclusive situation. For the three methods under consideration—Tucker's equally reliable method, the Angoff-Levine equally reliable method, and the Congeneric-Levine equally reliable method—if the two groups do not differ in either mean or variance on the anchor, the methods reduce to Angoff's (1984) design I (random groups, equal reliabilities) method because no adjustment for group differences is necessary. If the groups do differ in performance on the anchor, then the anchor differences are used to adjust for group differences on X and Y. The higher the correlation between V and X and between V and Y, the more likely this adjustment is appropriate (Angoff, 1987; Cook & Petersen, 1987). It may be shown (Kolen & Brennan, 1987; Woodruff, 1986) that the following three parameters determine how these anchor group differences are incorporated into the equating for the Tucker, Angoff-Levine, and Congeneric-Levine methods respectively:

$$\gamma_T = \frac{\sigma_{yv}}{\sigma_v^2} \quad , \tag{1}$$

$$\gamma_{AL} = \frac{\sigma_{yv} + \sigma_y^2}{\sigma_{yv} + \sigma_v^2} \quad , \tag{2}$$

and

$$\gamma_{CL} = \frac{\sigma_{yv}}{\sigma_v^2 \rho_{vv'}} = \frac{\gamma_T}{\rho_{vv'}} \quad . \tag{3}$$

The parameters γ_{CL} and γ_{AL} will be equal to each other when the assumptions required for their derivations are fulfilled. However, the subsequent analysis will consider the behavior of these parameters when the assumption $\rho(\tau_y,\tau_v) = 1$, which is required in the derivations of both γ_{AL} and γ_{CL}, is violated (τ denotes the true score on the test indicated by its subscript).

The above γ parameters pertain to the old test Y administered in group 2. If the weighted combination of group 1 and group 2, called the synthetic population by Braun and Holland (1982), is invoked, then the equating requires that the γ parameters be estimated for both the old and new tests.

If the synthetic population is ignored (Gulliksen, 1950; Kolen & Brennan, 1987; Woodruff, 1986),

then the γ parameters need only be estimated for the old test. For simplicity, this paper ignores the synthetic population, but its conclusions apply equally to equating with the synthetic population. In practice, these parameters are usually estimated by the method of moments (Angoff, 1982, 1984; Woodruff, 1986).

To simplify the analysis, certain assumptions are made that will always be satisfied in the practical application of these linear equating methods. They are $\sigma_y^2 > \sigma_v^2 > 0$ and $0 \leq \sigma_{yv} \leq \sigma_y\sigma_v$, the latter being equivalent to $0 \leq \rho_{yv} \leq 1$. In what follows, σ_{yv} is treated as a variable, but σ_y^2, σ_v^2, and $\rho_{vv'}$ are treated as constants. Under classical test theory, $\rho_{yv}^2 \leq \rho_{vv'}$. The present analysis allows the constant $\rho_{vv'}$ to assume any value between 0 and 1.

In the Tucker method, γ_T is a linear function of σ_{yv} with positive slope $1/\sigma_v^2$ and zero intercept. Its minimum value of 0 occurs when $\sigma_{yv} = 0$, and its maximum value of σ_y/σ_v occurs when $\sigma_{yv} = \sigma_y\sigma_v$. As σ_{yv} decreases, the Tucker method gives anchor group differences less weight in the equating process. This is a reasonable and desirable property because group differences between Y and X will usually be reflected by group differences on V, largely to the extent that V correlates with Y and X.

For the Angoff-Levine method, the first derivative of γ_{AL} is

$$\frac{\partial\gamma_{AL}}{\partial\sigma_{yv}} = \frac{\sigma_v^2 - \sigma_y^2}{(\sigma_{yv} + \sigma_v^2)^2} < 0 \quad . \tag{4}$$

Its second derivative is

$$\frac{\partial^2\gamma_{AL}}{\partial\sigma_{yv}^2} = \frac{2(\sigma_y^2 - \sigma_v^2)}{(\sigma_{yv} + \sigma_v^2)^3} > 0 \quad . \tag{5}$$

Hence γ_{AL} is a decreasing function of σ_{yv} with upward concavity. Furthermore, γ_{AL} has a minimum value of σ_y/σ_v when $\sigma_{yv} = \sigma_y\sigma_v$, and a maximum value of σ_y^2/σ_v^2 when $\sigma_{yv} = 0$. Because the minimum value of γ_{AL} coincides with the maximum value of γ_T, $\gamma_{AL} \geq \gamma_T$. As σ_{yv} decreases, the Angoff-Levine method gives anchor group differences more weight in the equating process. This is disadvantageous, but recall that the Levine method assumes that $\rho(\tau_y,\tau_v) = 1$, which implies that $\rho_{yv} = (\rho_{yy'}\rho_{vv'})^{1/2}$, which in turn implies that $\sigma_{yv} = \sigma_y\sigma_v(\rho_{yy'}\rho_{vv'})^{1/2}$. The above analysis reveals that when this assumption is violated, the behavior of the Angoff-Levine method is inappropriate.

Finally, in the Congeneric-Levine method, γ_{CL} behaves similarly to γ_T. It is a linear function of σ_{yv}, as is γ_T, but it has a steeper positive slope given by $1/(\sigma_v^2\rho_{vv'})$. Its minimum is also 0 when $\sigma_{yv} = 0$, but its maximum of $\sigma_y/(\sigma_v\rho_{vv'})$ when $\sigma_{yv} = \sigma_y\sigma_v$ is greater than γ_T's maximum. Consequently, $\gamma_{CL} \geq \gamma_T$ with equality holding only when $\rho_{vv'} = 1$, as can also be seen from an inspection of the formulas for γ_T and γ_{CL}. Like the Tucker method, the Congeneric-Levine method has the desirable property of giving less weight to anchor group differences as σ_{yv} decreases. However, the Congeneric-Levine method, like the Angoff-Levine method, assumes that $\rho(\tau_y,\tau_v) = 1$ or equivalently that $\sigma_{yv} = \sigma_y\sigma_v(\rho_{yy'}\rho_{vv'})^{1/2}$. The above analysis reveals that the Congeneric-Levine method, in contrast to the Angoff-Levine method, performs reasonably when this assumption is violated.

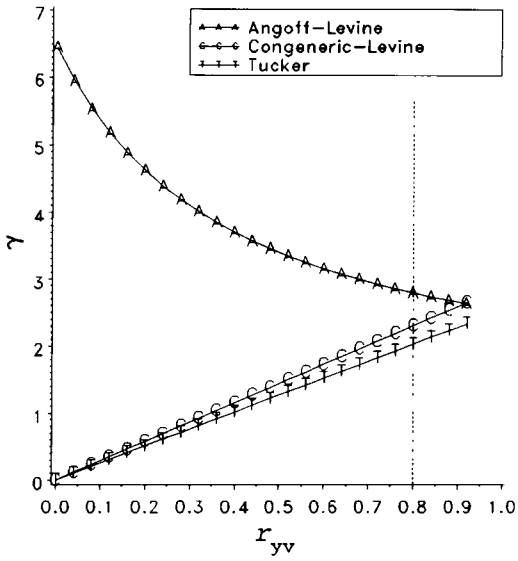
The previous analysis has focused on the exclusive anchor situation. It can be shown that the γ parameters for all three methods in the inclusive anchor situation equal their respective exclusive-situation γs plus unity (Woodruff, 1986). Hence the above results for the exclusive anchor situation apply to the inclusive anchor situation with only slight modification which does not alter comparative performance among the three procedures.

An Example

The previous analysis and its implications for equating practice are illustrated for an exclusive anchor situation in Figure 1. The example is based on an actual equating, but the data have been slightly modified to clarify and simplify the presentation. The number of items in both the old and new tests is 300, and the number of items in the exclusive anchor is 100. The two groups each consisted of approximately 300 examinees. The statistics reported below are derived from the performance of group 2 on the old test, Y, and anchor, V.

In Figure 1, γ has been rescaled by the multiplication of $1 = s_y s_v / s_y s_v$ so that the graph has its horizontal axis on the scale of r_{yv}. The vertical dashed line indicates the obtained sample value for r_{yv}. The values of the sample statistics used to con-

Figure 1
Plot of Gammas for $s_y = 33$, $s_v = 13$, and $r_{vv'} = .88$ (Vertical Dashed Line Indicates the Observed Sample Value of $r_{yv} = .80$)



struct the graph were $s_y = 33$, $s_v = 13$, $r_{vv'} = .88$, and $r_{yv} = .80$. Coefficient alpha was used to estimate reliability, and alpha for Y was .96. Under the assumptions of classical test theory, ρ_{yv} cannot exceed the square root of $\rho_{yy'}\rho_{vv'}$ which in this situation was estimated to be .92. Hence, the γ curves run from 0 to .92, and the graph demonstrates the behavior of the γs as r_{yv} takes all values permissible in this situation.

For the actual observed value of .80 for r_{yv}, the three γs have different values. If the assumptions required for the derivations of γ_{CL} and γ_{AL} were satisfied, then the Congeneric-Levine γ curve, the Angoff-Levine γ curve, and the vertical dashed line would all intersect at the maximum permissible value of $r_{yv} = .92$, which would imply a disattenuated Y-V correlation equal to 1. However, the disattenuated correlation between Y and V is not 1 but only .87, indicating that the anchor is not a perfect representation of the test. As a consequence, the Angoff-Levine method should not be used because the sample estimate of γ_{AL} is spuriously inflated.

If the two groups of examinees show large differences in either mean or variance on the anchor, then the Tucker method should also not be used, especially if the reliability of the anchor is much less than 1. In such a situation, the Congeneric-Levine method may be used because it makes no assumptions about group differences, and it performs reasonably when the disattenuated correlation between the test and the anchor is less than unity.

Coefficient alpha was used in the estimation of the disattenuated test-anchor correlation and γ_{CL}. It was judged to be an appropriate reliability estimate for the test and anchor used here. Careful consideration is necessary for selecting an appropriate reliability index to use in estimating disattenuated correlations and γ under the Congeneric-Levine method. This topic is discussed by Lord and Novick (1968, sec. 6.5).

Conclusions

The preceding analysis offers an explanation for the empirical results of Klein and Jarjoura (1985), who noted that the Angoff-Levine method was more sensitive than Tucker's method to a lack of content balance between the tests and anchor. It also has implications for the application of these equating methods. If the groups differ greatly in ability as evidenced by their performance on the anchor, and application of the Tucker method is consequently untenable, then the disattenuated correlation between Y and V should be computed before applying the Angoff-Levine method. If this disattenuated correlation is significantly less than unity, then the Angoff-Levine method may also be inappropriate. An appealing alternative is the Congeneric-Levine method because it permits large group differences and performs reasonably when $\rho(\tau_y,\tau_v) < 1$.

This conclusion is based on a comparison of parameter values. In practice, these parameters must be estimated from sample statistics, as was illustrated in the example. This does not compromise the above conclusion, however, because in all practical applications of equating there are at least several hundred examinees (more often, many thousand) in each group. The parameter estimates will

be derived from sample first- and second-order moments and first-order cross-product sample moments. Hence, the sample statistics will be consistent estimators of the parameters, and the large sample sizes met with in practice will ensure that decisions based on the sample values are reasonably accurate.

Finally, this paper has demonstrated that the Congeneric-Levine method performs reasonably in a recalcitrant situation that may occur in practice. However, if the assumptions required for either the Tucker or Angoff-Levine methods are satisfied, then one of these methods may be preferable because both are easier to implement than the Congeneric-Levine method.

References

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

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 55–70). New York: Academic Press.

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.). Washington DC: American Council on Education, 1971.]

Angoff, W. H. (1987). Technical and practical issues in equating: A discussion of four papers. *Applied Psychological Measurement, 11,* 291–300.

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. 9–50). 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.

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

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22,* 197–206.

Kolen, M. J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement, 9,* 209–223.

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. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (ETS Research Bulletin 55-23). Princeton NJ: Educational Testing Service.

Lord, F. M. (1960). Large sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55,* 307–321.

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

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 David Woodruff, ACT, P.O. Box 168, Iowa City IA 52243, U.S.A.

Computer Program Exchange
Continued from page 256

Description

The program collection ADCOM is written in C and runs under MS-DOS 3.3 and UNIX. It consists of the filters INDEP, CONEQ, SOLIN, and EQDELTA.

INDEP is a program for testing the independence condition of the binary relation of an additive conjoint structure, where a binary relation $\succsim$ on the Cartesian product $A_1 \times A_2 \times \ldots \times A_n$ is *independent* if and only if for each $M \subset N$ $(N = 1, 2, \ldots, n)$, the ordering $\succsim_M$ which is induced by $\succsim$ on $\times_{i \in M} A_i$ for fixed choices $a_i \in A_i$, $i \in N - M$ is not affected by those choices (see Krantz, Luce, Suppes, & Tversky, 1971).

The conjoint structure may consist of two or three components. INDEP reads the number of elements of each of the three components, and then the rank order of the items. Output includes the formatted (as matrices) rank order. If the order is not independent, a message is provided.

CONEQ computes the coefficient matrix **A** from a given rank order of an additive conjoint structure of up to three components and outputs the coefficient matrix, with a column added to distinguish between inequalities (0) and equalities (1), respectively. SOLIN computes the extreme points of a given positive convex polyhedral cone, using as input the output of the program FINEQ.

EQDELTA computes the equal delta solution of a given solution matrix. The input may be the output of SOLIN. EQDELTA writes the equal delta solution to output, that is, the sums of rows of the solution matrix.

ADCOM may be used in different ways. For instance, given a rank order of items of a conjoint measurement experiment, one possibility for obtaining scale values may be to run
indep < ord.dat|coneq|solin|eqdelta.

Availability

A listing of the programs and a manual are available without charge from the author: Ronald Hübner, Universität Regensburg, Universitätsstraße 31, D-8400 Regensburg, Federal Republic of Germany. To obtain the programs on diskette, send a formatted 5.25-inch MS-DOS diskette to the author.

References

Chernikova, N. V. (1965). Algorithm for finding a general formula for non-negative solutions of a system of linear inequalities. *U.S.S.R. Computational Mathematics and Mathematical Physics, 5,* 228–233.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. I). New York: Academic Press.

Lehner, P. E., & Noma, E. (1980). A new solution to the problem of finding all numerical solutions to ordered metric structures. *Psychometrika, 45,* 135–137.

Lukas, J. (1985). COMESCAL: A microcomputer program for testing axioms and finding scale values for conjoint measurement data. *Behavior Research Methods, Instruments, & Computers, 17,* 129–130.

McClelland, G. H., & Coombs, C. H. (1975). ORDMET: A general algorithm for constructing all numerical solutions to ordered metric structures. *Psychometrika, 40,* 269–290.

Motzkin, T. S., Raiffa, H., Thompson, G. L., & Thrall, R. M. (1953). The double description method. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games, II. Annals of Mathematics Studies*. Princeton NJ: Princeton University Press.