DOCUMFNT RESUME

ED 292 878	TM 011 329

AUTHOR Woodruff, David J.

TITLE An Analytical Comparison among Three Linear Equating Methods for the Common Item Nonequivalent Populations Design.

INSTITUTION American Coll. Testing Program, Iowa City, IA. Research Div.

PUB DATE 1 Apr 88

NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Correlation; *R'search Methodology; *Statistical Analysis; Statistical Data

IDENTIFIERS Levine Equating Method; *Linear Equating Method; Tucker Common Item Equating Method

ABSTRACT

Three linear equating methods for the common item non-equivalent populations design, a design often used in practice, are compared by an analytical method. The models include: (1) Tucker's equally reliable method; (2) Levine's equally reliable methods; and (3) the congeneric method recently introduced by Woodrull (1986). Tucker's method makes assumptions about observed score regressions and is based on a linear regression model. The Levine and congeneric methods make assumptions about true score regressions and are based on linear structural models. The analysis is graphically illustrated using data from actual test administrations. If groups differ greatly as shown by their performance in the anchor and if application of Tucker's equating method is not tenable, the disattenuated correlation between Y and V should be computed. If this disattenuated correlation is significantly less than unity, the Levine method should not be used, and the congeneric method becomes an appealing alternative. (SLD)

An Analytical Comparison Among Three Linear Equating Methods for the Common Item Nonequivalent Populations Design

David Woodruff, Psychometrician
The American College Testing Program
P.O. Box 168
Iowa City, IA 52243
319/337-1073

Running Head: LINEAR EQUATING COMPARISONS

An Analytical Comparison Among Three Linear Equating Methods for the Common Item Nonequivalent Populations Design

Running Head: LINEAR EQUATING COMPARISONS

Abstract

Three linear equating methods for the common item nonequivalent populations design, a design commonly used in practice, are compared using an analytical method. The analysis is graphically illustrated using data from actual test administrations. Conclusions derived from the analysis which have implications for the practical application of these equating methods are discussed.

Key Words: Linear Equating, Tucker's Equating Method, Levine's Equating Method, Congeneric Model

Introduction

Linear equating methods for the common item nonequivalent populations (CINEP) design are derived or discussed by several authors: Gulliksen (1950), Levine (1955), Angoff (1971,1982), Braun and Holland (1982), Kolen (1985), Wood''uff (1986), and Kolen and Brennan (1987). Angoff (1971) refers to this design as design IV--nonrandom groups. Under this design, a new test X is given to group 1 and an old test Y is given to group 2, while a usually shorter anchor test V is given to both groups. The anchor test V may comprise a scoreable part of the tests and this is referred to as the inclusive anchor situation, or test V may not contribute to examinees' scores and this is referred to as the exclusive anchor situation. Two methods commonly used in practice for linear equating under the CINEP design are Tucker's equally reliable method (Gulliksen, 1950; Angoff, 1971, 1982; Kolen, 1985) and Levine's equally reliable method (Levine, 1955; Angoff, 1971, 1982; Woodruff, 1986). A third method recently introduced by Woodruff (1986) is called the congeneric method. Tucker's method makes assumptions about observed score regressions, while the Levine and congeneric methods make assumptions about true score regressions. Tucker's method is based on a linear regression model, while the Levine and congeneric methods are based on linear structural models. The congeneric method is less restrictive in its assumptions than is Levine's method, but as a consequence the congeneric method is slightly more difficult to implement in that it requires an estimate of the anchor's reliability. According to Angoff (1971), who cites Levine (1955) and Lord (1960), Tucker's method is most appropriate for situations in which the two groups show no more than small differences in mean and ariance on the anchor, while Levine's method (and the congeneric method also) may accommodate larger differences so long as the true scores on the tests and anchor correlate unity. The purpose of the present paper is to compare through analytical and empirical means the performance of the three methods as the covariance between the tests and the anchor varies. Klein and Jarjoura (1985) undertook a similar investigation using only empirical methods. They noted that Levine's method was more sensitive than Tucker's to a lack of content balance between the tests and the anchor. The present study will suggest an explanation for their finding which indicates that the performance of the congeneric method should be more similar to the Tucker method than to the Levine method as the covariance between the test and anchor decreases. This result has practical implications for equating and these will be discussed. Real test data will be used to graphically illustrate these conclusions.

Analysis

The analysis will begin with the exclusive anchor situation. Later, it will be 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, Levine's equally reliable method, and the congeneric method, if the two groups do not differ in either mean or variance on the anchor, then all three methods reduce to Angoff's (1971) design I: random groups equal reliabilities method since 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 V and Y the more likely that this adjustment is appropriate (Cook and Peterson, 1987; Angoff, 1987). It may be shown (Kolen and Brennan, 1987; Woodruff, 1986) that the following three parameters determine how these anchor group differences are incorporated into the equating for the Tucker, Levine, and congeneric methods respectively:

$$\gamma_T = \sigma_{yv}/\sigma_v^2 \quad ,$$

$$\gamma_L = (\sigma_{yv} + \sigma_y^2)/(\sigma_{yv} + \sigma_v^2) \quad , \text{ and}$$

$$\gamma_C = \sigma_{yv}/\sigma_v^2\rho_{vv'} = \gamma_T/\rho_{vv'} \quad .$$

The above gamma parameters pertain to the old test Y administered in population 2. If the synthetic population (Braun and Holland, 1982) is invoked, then the equating requires that the gamma parameters be estimated for both the old and new tests. If the synthetic population is ignored (Gulliksen, 1950; Woodruff, 1986; Kolen and Brennan, 1987), then the gamma parameters need only be estimated for the old test. For simplicity, this paper will ignore 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, 1971, 1982; Woodruff, 1986).

To simplify the analysis, certain assumptions will be made which 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} will be treated as a mathematical variable, but σ_y^2, σ_v^2, and $\rho_{vv'}$ will be treated as mathematical constants. Under classical test theory, $\rho_{yv}^2 \leq \rho_{vv'}$. The present analysis allows the constant, $\rho_{vv'}$, to assume any value between zero and one.

Focusing first on the Tucker method shows that γ_T is a linear function of σ_{yv} with positive slope $1/\sigma_v^2$ and zero intercept. Its minimum value of zero occurs when $\sigma_{yv} = 0$, and its maximum value is σ_y/σ_v which 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, since, as was previously mentioned, 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.

The second method to be analyzed is Levine's. The first derivative of γ_L is $d\gamma_L/d\sigma_{yv} = (\sigma^2_v - \sigma^2_y)/(\sigma_{yv} + \sigma^2_v)^2 < 0$. Its second derivative is $d\gamma^2_L/d\sigma^2_{yv} = 2(\sigma^2_y - \sigma^2_v)/(\sigma_{yv} + \sigma^2_v)^3 > 0$. Hence, γ_L is a decreasing function of σ_{yv} with upward concavity. Furthermore, γ_L has a minimum value of σ_y/σ_v when $\sigma_{yv} = \sigma_y\sigma_v$, and a maximum value of σ^2_y/σ^2_v when $\sigma_{yv} = 0$. Since the minimum value of γ_L coincides with the maximum value of γ_T, $\gamma_L \geq \gamma_T$. As σ_{yv} decreases, the Levine method gives anchor group differences more weight in the equating process. This is an unreasonable and undesirable property, but recall that the Levine method assumes that $\rho(T_y,T_v) = 1$ 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 the Levine method will perform poorly when this assumption is violated.

Focusing, finally, on the congeneric method, γ_C has behavior similar to γ_T . It is a linear function of σ_{yv} as is γ_T, but it has a steeper positive slope given by $1/\sigma_v\rho_{vv'}$. Its minimum is also zero 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, $\gamma_C \geq \gamma_T$ with equality holding only when $\rho_{vv'} = 1$, as can also be seen from an inspection of the formulas for γ_T and γ_C. Like the Tucker method, the congeneric method has the desirable property of giving less weight to anchor group differences as σ_{yv} decreases. However, the congeneric method, like the Levine method, assumes that $\rho(T_y,T_v) = 1$ or equivalently that $\sigma_{yv} = \sigma_y\sigma_v(\rho_{yy'}\rho_{vv'})^{1/2}$. The above analysis reveals that the congeneric method, in contrast to the 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 γ's 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 between the three procedures.

Illustration

The previous analysis is illustrated under the exclusive anchor situation for four different test administrations in Figures 1 through 4. Though the data are real, the exact details of the present application are not reflective of the actual equating situations and are illustrative only.

Insert Figures 1 through 4 about here

To facilitate comparisons among the graphs, gamma has been rescaled by the multiplication of $1 = s_y s_V / s_y s_V$ so that each graph has its horizontal axis on the scale of r_{yV} from 0 to 1. At the bottom of the figures are the values of the statistics from which the graphs were derived. The reliabilities are alpha coefficients. For both groups, the number of test items and anchor items for Figures 1 through 4 are, respectively, (295, 105), (190, 60), (55, 20), and (32, 13), while the number of examinees in group 1 and group 2 are, respectively, (326, 305), (748, 1625), (700, 4093), and (1111, 4093). The statistics are for the old test Y administred in group 2.

The figures are presented in order of test length with the longest and most reliable test presented in Figure 1 and the shortest and least reliable test presented in Figure 4. As a consequence, the figures are similarly ordered by the actual sample value of the correlation between the test and its anchor, r_{YV}, as can be seen from the vertical dashed line in each graph. The graphs indicate that as the reliability of the anchor decreases the discrepancy between the congeneric and Tucker gammas increases as their formulas indicate.

For Figures 1 and 2 the disattenuated correlation between the test and its anchor is between .99 and 1.01; so, for these figures, the actual sample correlation indicated by the vertical dashed line is the maximum attainable Y-V correlation given the unreliability of the measures. In these figures, the congeneric and Levine plots intersect at the sample value of the Y-V correlation which is appropriate since both methods assume unity for the value of the disattenuated Y-V correlation. In Figure 3, the intersection occurs to the right of the correlation, while in Figure 4 it is slightly to the left. The disattenuated correlation for the data in Figure 3 is .92. The disattenuated correlation for the data in Figure 4 is 1.02. Figure 3 will be discussed in the next section since it so clearly demonstrates the central point of this paper. Conversely, Figure 4 suggests a limitation. Here, where the 13-item anchor is quite short, it is probable that the anchor's reliability is slightly under-estimated with the result that the disattenuated Y-V correlation and the congeneric gamma are slightly over-estimated.

Alpha coefficients were used in the estimation of the disattenuated test-anchor correlations. These were judged to be appropriate reliability estimates for the tests used here and for the illustrative nature of this paper. Careful consideration is necessary for selecting an appropriate reliability index to use in estimating disattenuated correlations and gamma under the congeneric method. This topic is discussed by Lord and Novick (1968, sec. 6.5).

Discussion

The preceding analysis offers an explanation for the empirical results of Klein and Jarjoura (1985), and it also has implications for the application of these equating methods. If the groups differ greatly as evidenced by their performance on the anchor, and as a consequence application of the Tucker method is untenable, then before applying the Levine method the disattenuated correlation between Y and V should be computed. If this disattenuated correlation is significantly less than unity, then the Levine method should also not be used. An appealing alternative is the congeneric method since it permits large group differences and performs reasonably when $\rho(T_y,T_v) < 1$. This situation is illustrated in Figure 3. Here $\hat{\gamma}_L$ is about 2.4 and $\hat{\gamma}_C$ is about 2.1. The congeneric method gives about 12.5% less weight to the anchor information on mean differences and about 23% less weight to the information on anchor differences in variances (gamma is squared when applied to variances). This reduction seems appropriate since the disattenuated Y-V correlation is only .92 suggesting that the anchor may not be a perfect representation of the test.

The present analysis which has lead to the above conclusion is based on a comparison of parameter values. In practice, these parameters will have to be estimated from sample statistics as was illustrated in the four examples. This does not compromise the above conclusion, however, since in all practical applications of equating there is at least several hundred examinees in each group and more usually many thousand. 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 insure that decisions based on the sample values are reasonably accurate.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600. Washington, DC: American Council on Education.

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. (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: John 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 sampl·s of different ability (ETS Research Bulletin No. 55-23). P''inceton, NJ: Educational Testing Service.

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

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.

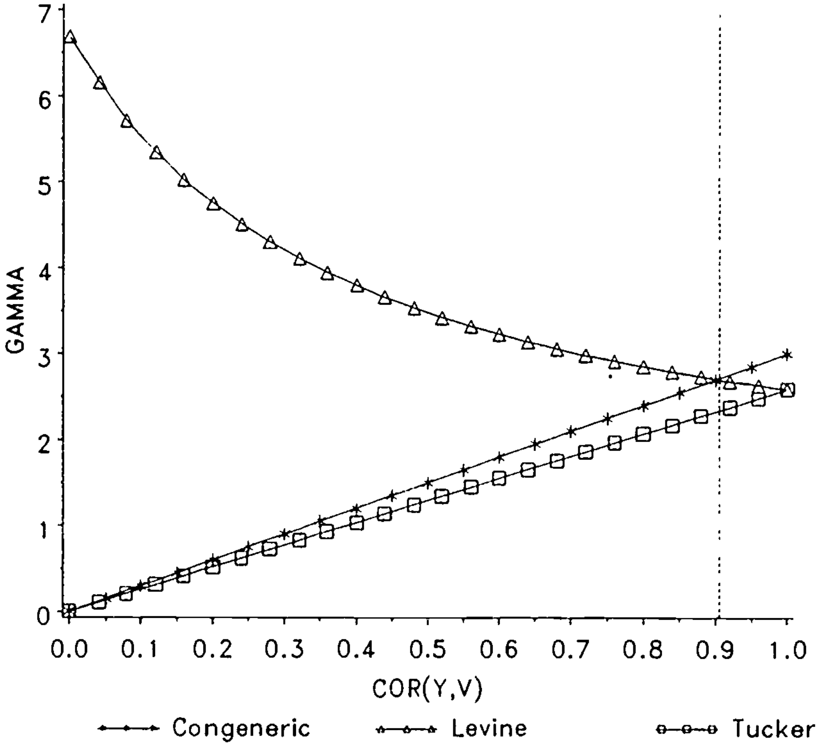

FIGURE 1
Plot of Gammas for SD(Y)=32.837, SD(V)=12.691, and REL(V)=.86176.
The vertical dashed line indicates the actual value of COR(Y,V).

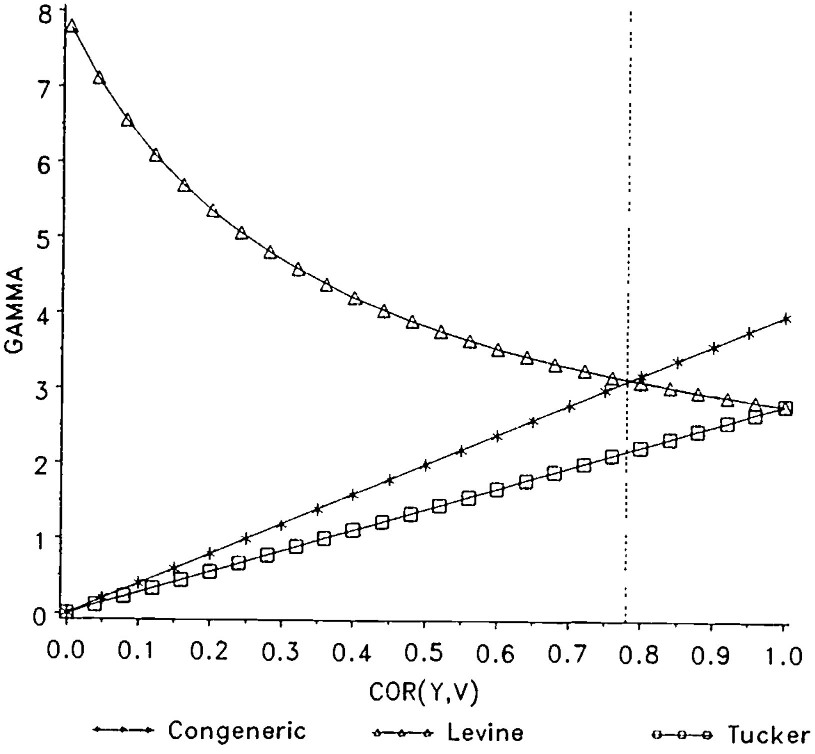

FIGURE 2

Plot of Gammas for SD(Y)=16.497, SD(V)=5.9077, and REL(V)=.70032.
The vertical dashed line indicates the actual value of COR(Y,V).

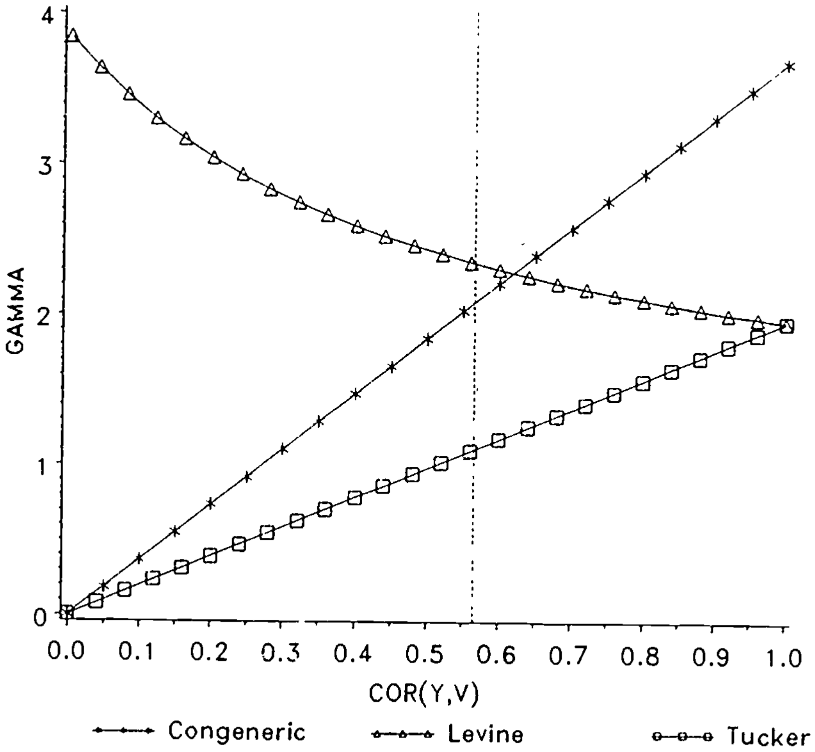

FIGURE 3
Plot of Gammas for SD(Y)=4.9067, SD(V)=2.5022, and REL(V)=.53138.
The vertical dashed line indicates the actual value of COR(Y,V).

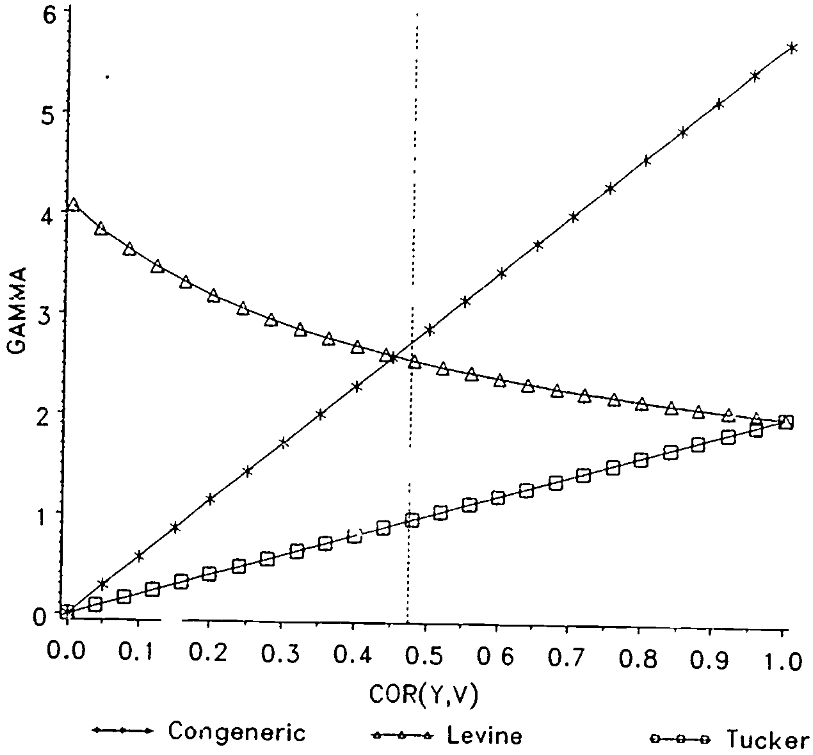

FIGURE 4
Plot of Gammas for SD(Y)=3.4624, SD(V)=1.7160, and REL(V)=.35030.
The vertical dashed line indicates the actual value of COR(Y,V).