Effect of Examinee Group on Equating Relationships

Deborah J. Harris and Michael J. Kolen
The American College Testing Program

Many educational tests make use of multiple test forms, which are then horizontally equated to establish interchangeability among forms. To have confidence in this interchangeability, the equating relationships should be robust to the particular group of examinees on which the equating is conducted. This study investigated the effects of ability of the examinee group used to establish the equating relationship on linear, equipercentile, and three-parameter logistic IRT estimated true score equating methods. The results show all of the methods to be reasonably independent of examinee group, and suggest that population independence is not a good reason for selecting one method over another.

Many educational testing programs make use of multiple test forms for reasons including test security and test disclosure. Multiple forms of a test are constructed from the same content specifications to be as similar as possible to one another in their content and statistical characteristics. These forms are then horizontally equated so that scaled scores from the forms can be used interchangeably. Angoff (1971) suggests that equating should result in a conversion of scores from one test form to the scale of another test form that is "independent of the individuals from whom the data were drawn to develop the conversion and should be freely applicable to all situations" (p. 563). Lord (1980, ch. 13) indicates that observed score equating relationships are dependent on the subpopulation of examinees, whereas true score equating relationships are subpopulation independent if a unidimensional item response theory (IRT) model holds. The subpopulation independence property has often been used as an argument for preferring IRT true score methods to observed score methods for equating test forms (Lord, 1980; Lord & Wingersky, 1984). However, in practice, unidimensional IRT models are not likely to hold exactly. Thus, the equating method which is least subpopulation dependent must be established empirically.

This study investigated the effects of examinee subgroup ability on equating functions resulting from the three-parameter logistic estimated true score IRT method, the linear observed score method, and the equipercentile observed score method. Using a random groups design and five forms of the ACT Assessment Mathematics Usage test, forms were equated using an examinee subgroup of high ability and then equated again using an examinee subgroup of low ability. The resulting high group function for a given equating method is compared to the low group function for that same method. The equating method with the most similar high and low group functions is considered to be least subpopulation dependent.

Background

The effects of examinee group on equating re-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 1, March 1986, pp. 35–43

0146-6216/86/010035-09$1.70

lationships have, in general, been studied in situations other than the horizontal equating situation. The Anchor Test Study (Loret, Seder, Bianchini, & Vale, 1972), probably the largest equating study ever conducted, investigated equating relationships between the reading tests of eight widely used standardized achievement tests for Grades 4, 5, and 6. Because the tests differed substantially in difficulty level and content specifications, it cannot be viewed as a typical horizontal equating between test forms; however, the results are still of interest in the present context. Pairs of tests were equated using both equipercentile and linear methods, and these results were compared in terms of estimated errors. Various subgroup equating analyses were conducted, including one based on three levels of IQ. It was concluded that "the equating results for the IQ subgroups generally parallel those of the total equating sample. No systematic differences between the equating lines of the IQ subgroups and the total equating group were found for all tests or for a given test across the three grades [fourth, fifth, and sixth]" (Loret et al., 1972, p. 194). The inconsistencies that were present in the tails were tentatively attributed to the scarcity of data in those regions.

Other studies have also examined effects of examinee group on equating relationships. In a vertical equating application of the Rasch model and using the Mathematics Computation test from the Iowa Tests of Basic Skills, Loyd and Hoover (1980) found that different ability subgroups (here defined by grade in school—sixth, seventh or eighth) established different equating relationships. Studies by Slinde and Linn (1977, 1978, 1979a, 1979b) and Goulet, Linn, and Tatsuoka (1975) taken together suggest that in vertical equating situations, linear, equipercentile, and Rasch method results are dependent on examinee group, although the dependence appears to lessen as either the groups or tests become more similar.

Marco, Petersen, and Stewart (1983; Petersen, Marco, & Stewart, 1982) examined a number of equating methods, including linear and curvilinear methods, under a variety of conditions using anchor tests; one of the conditions was the similarity of the examinee groups. Making partial use of equating a test to itself, and of the standardized weighted mean square difference and the squared bias to evaluate the effectiveness of the equating, Marco et al. found the similarity of the sample to have a relatively small effect on equating when the anchor test was similar to the total test.

The studies reviewed above provide some information on the dependency between examinee groups' differences in ability and the equating relationship, although most focus on vertical rather than horizontal equating. In addition, most use the Rasch model rather than the three-parameter logistic model, which was used in the present study.

Equating Methods

Linear Equating

In linear equating, the same scaled scores are assigned to a raw score on Form X and a raw score on Form Y if the raw scores correspond to identical standardized scores for a given group of examinees—that is, if

$$\frac{X - \bar{X}}{S_x} = \frac{Y - \bar{Y}}{S_y}, \quad (1)$$

where $\bar{X}$, S_x, $\bar{Y}$ and S_y are, respectively, the mean and standard deviation of Form X and Form Y.

Equipercentile Equating

Angoff (1971, p. 563) defines equating as follows: "Two scores, one on Form X and the other on Form Y (where X and Y measure the same function with the same degree of reliability) may be considered equivalent if their corresponding percentile ranks in any given group are equal." This definition is commonly accepted, and suggests the method of equipercentile equating.

This method requires the cumulative frequency distributions for each test, and assigns the same scaled score to the raw scores on Form X and Form Y if their percentile ranks are the same. When the two distributions differ only in their first and second moments, the equipercentile and linear equat-

ing methods coincide; they will yield similar results when the frequency distributions of raw scores on the two test forms are similar in shape.

Because, in a sense, the linear method is subsumed by the equipercentile method, it might be advantageous to use the latter. This, however, is not always the case. Although sparse data in the tails of the distribution may affect the linear method by influencing the value of the means and standard deviations, the problem is particularly acute for the equipercentile method, and may lead to large random errors in extreme scores. For this reason, methods for smoothing in equipercentile equating, such as those described by Angoff (1971) and Kolen (1984), are used.

IRT Equating

IRT equating methods "characterize equivalent scores on two test forms as those scores which correspond to the same estimated level of the latent trait, ability, or skill underlying both tests" (Cook & Douglas, 1982, p. 12). IRT models require some stringent assumptions, one of which is unidimensionality. This implies that there is only one trait underlying the given responses to the test items. Although this assumption is never strictly met in practice, there is some evidence that IRT equating methods are somewhat robust to violations of it (Cook & Eignor, 1983a). Other assumptions, such as a specified functional form for the item characteristic curves, are also required, though in a practical sense, the issue is not how well the data fit the model so much as how well the model will perform with real data in a real testing situation.

Several IRT models exist, including those for one, two, and three estimated item parameters. Only the three-parameter logistic IRT model was considered in this study. This model assumes that the probability of a correct response of a person of ability θ with item g ($g = 1, \ldots, n$) is

$$P_g(\theta) = c_g(1 - c_g)\frac{e^{Da_g(\theta - b_g)}}{1 + e^{Da_g(\theta - b_g)}}, \qquad (2)$$

where D is a scaling constant usually set equal to 1.7;
n is the number of items on the test;
b_g is the item difficulty parameter;
a_g is the item discrimination parameter; and
c_g is the lower asymptote parameter.

In estimated true score equating under this model, for a given θ the sum of the estimated item characteristic curves on Form X is considered to be equivalent to the sum of the estimated item characteristic curves on Form Y.

Method

Data

Test score data from five forms of the ACT Assessment (1980) Mathematics Usage test were used. The tests consist of 40 five-option multiple-choice items. Number-correct scoring was used; group means were approximately 50% correct. The test forms were designed to meet the same content specifications and to be as similar as possible in terms of difficulty and reliability. The five test forms were randomly assigned within test centers (Design I; Angoff, 1971). Sample sizes varied somewhat across forms, with each form administered to 3,869 to 3,967 examinees. In addition to responding to the test items, examinees also responded to a series of inventory items, one of which concerned high school grade point average (GPA). Although self-reported GPA may be somewhat suspect, a study concluded (ACT, 1973, p. 308): "In summary, considerable confidence may be placed in the accuracy with which students report their high school grades. In fact, 78.0% of all students report their grades accurately, and 97.8% agree within one letter grade of what is reported by school officials."

Examinees were divided into two groups: those with a self-reported GPA of B or better, and those with a self-reported GPA of less than B. Of the total 19,518 examinees, 494 were dropped from analyses for failing to report GPA. Test score statistics for the dropped examinees were similar to those of retained examinees. Test form moments, by examinee group, are shown in Table 1.

Equating Procedures

One test form, A, was chosen to function as the

Table 1
Raw Score Moments for Test Form by Examinee Group

Group and Form	N	Mean	Standard Deviation	Skewness	Kurtosis
High Group					
A	2003	23.37	7.68	-.05	2.28
B	2071	24.50	8.35	-.13	2.11
C	2088	23.88	7.87	-.03	2.16
D	2017	23.28	8.02	.03	2.20
E	2034	22.92	8.62	-.12	2.12
Low Group					
A	1823	16.27	6.25	.46	2.88
B	1723	17.28	7.04	.47	2.67
C	1782	17.24	6.54	.57	2.90
D	1742	16.20	6.49	.64	3.20
E	1741	15.04	7.31	.68	2.79

anchor form, and the other four test forms (B, C, D, and E) were equated to it using linear, equipercentile, and three-parameter estimated true score equating procedures. The equipercentile relationships were smoothed using the cubic spline procedure described by Kolen (1984). In this procedure, larger values of the *S* parameter allow for more smoothing, with a very large value of *S* resulting in a straight line. Three equipercentile relationships were used: (1) unsmoothed; (2) a smoothing value determined judgmentally on the basis of visual inspection; and (3) an *S* value equal to .50. (The *S* value reported here is equal to the *S* value described by Kolen, 1984, divided by the number of test items.) The varied *S* values were: Form B, high ability = .10; Form B, low ability = .15; Form C, high ability = .15; Form C, low ability = .20; Form D, high ability = .20; Form D, low ability = .10; Form E, high ability = .10; Form E, low ability = .10.

The IRT parameters were estimated using LOGIST V (Wingersky, Barton, & Lord, 1982). The estimated true score of an examinee is given as the sum, over items, of the estimated probability to correctly respond to an item for someone with his/her estimated ability, implying that estimated true scores below the "pseudo-chance" level (the sum of the lower asymptotes) are undetermined. For this study, zero scores on forms were considered, arbitrarily, to be equivalent, and estimated true score equivalents below the pseudo-chance level were arrived at by linear interpolation (see Kolen & Whitney, 1982). Due to the scarcity of scores in the zero to pseudo-chance region, the use of this procedure has minimal effect on the findings of this study.

Using each equating method, forms B, C, D, and E were equated to Form A for the high-ability group. A similar procedure was followed for the low-ability group. Thus two equating relationships were established per form using each of five methods (linear; equipercentile-unsmoothed; equipercentile-smoothed, *S* varied; equipercentile-smoothed, *S* = .50; three-parameter IRT estimated true score method). Each of these relationships was established using the same anchor test form, though half used high-ability examinee groups and half used low-ability groups.

Evaluative Indices

Several indices were examined to study the effect of examinee group differences on equating methods. The first index, the root mean squared error, was computed by squaring the difference between the score equivalent established in the high ability group and that established in the low ability group for each raw score on the anchor form;

weighting this by the frequency, over all examinees, of that raw score on the anchor form; summing over score points; dividing by the total number of examinees on the anchor form; and taking the square root. That is,

$$\text{Index 1} = \left[\frac{\sum f_i (H_i - L_i)^2}{\sum f_i}\right]^{1/2}, \quad (3)$$

where H_i is the equivalent of a raw score of i on the anchor form established by equating a non-anchor form to the anchor form using high-ability examinees;

L_i is the equivalent of a raw score of i on the anchor form established by equating a non-anchor form to the anchor form using low-ability examinees;

f_i is the frequency of the raw score i on the anchor form, using both high- and low-ability examinees; and

i ranges from 1 to 39, to aid in comparing the equating methods, as LOGIST V does not deal with zero or perfect raw scores.

Index 2 is similar to Index 1, with the absolute value of the high-low difference taken, instead of the square:

$$\text{Index 2} = \frac{\sum f_i |H_i - L_i|}{\sum f_i}. \quad (4)$$

Index 3, the mean difference, makes use of the signed difference:

$$\text{Index 3} = \frac{\sum f_i (H_i - L_i)}{\sum f_i}. \quad (5)$$

Index 1 and Index 2 can be viewed as measures of overall difference. Index 3 can be viewed as a measure of bias or average signed difference.

Although weighting by score frequencies in the above indices does give more emphasis to those score equivalents most likely to occur, it is also of interest to examine the unweighted indices:

$$\text{Index 1}' = \left[\sum (H_i - L_i)^2/K\right]^{1/2} \quad (6)$$

$$\text{Index 2}' = \sum |H_i - L_i|/K \quad (7)$$

$$\text{Index 3}' = \sum (H_i - L_i)/K \quad (8)$$

where K is the number of score points (39 in this study).

Results

Values of the indices for the four non-anchor forms equated to the anchor form, using each of the five equating methods, are shown in Table 2. Smaller values of Indices 1, 1′, 2, and 2′ are indicative of more similar equating functions across groups. Values closer to zero for Indices 3 and 3′ are indicative of more similar equating functions. As an examination of Table 2 shows, no method is clearly favored over another for Indices 1 and 2. All methods on all forms, with the single exception of the IRT method on Form D, show a negative bias on Index 3, which results from the low ability equivalent being larger than the high ability equivalent. The IRT method produces Index 3 values closest to zero.

Similar results are found for the unweighted results, with the IRT method producing smaller Index 3′ values than the other methods. No method shows consistent superiority for Indices 1′ and 2′, as the IRT, linear, and unsmoothed equipercentile methods all range from the smallest value to the largest value for these indices.

The high ability equivalent minus low ability equivalent differences are graphed for each form in Figure 1. Each graph shows results from all five equating methods for that form. The horizontal line at zero indicates no difference between the equivalents established on the high-ability and the low-ability groups for that particular form when equated to the anchor form. Again, no clear preference is illustrated for a particular method, especially when considered across the graphs and throughout the score scale.

As can be seen from Figure 1, most of the score equivalent differences are negative throughout the score scale range for the linear and equipercentile methods, while the IRT method tends to produce negative values in the middle and positive values at the ends of the score scale. This trend provides some insight as to why the IRT method produces values closer to zero for Indices 3 and 3′ than the other methods, while not consistently producing smaller values for Indices 2 and 2′ or 1 and 1′.

The standard errors of the unsmoothed equipercentile equating method provide a baseline to com-

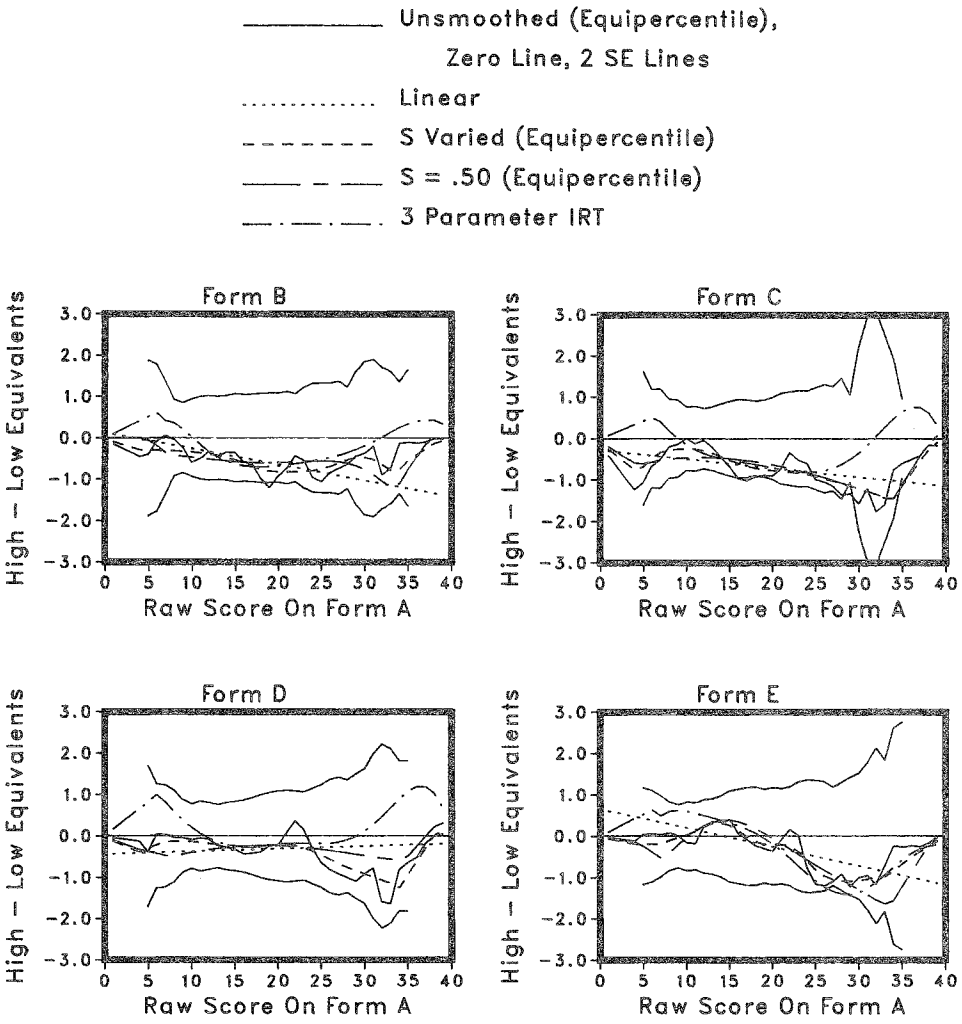
Table 2
Weighted and Unweighted Index Values

Index and Form	IRT	Linear	Equipercentile S = 0	Equipercentile S Varied	Equipercentile S = .50
Index 1					
B	.51	.71	.66	.63	.61
C	.58	.74	.86	.79	.76
D	.37	.31	.57	.48	.34
E	.73	.45	.58	.55	.66
Index 2					
B	.47	.65	.59	.60	.59
C	.53	.72	.73	.72	.71
D	.28	.30	.42	.38	.33
E	.61	.35	.42	.42	.51
Index 3					
B	-.37	-.65	-.58	-.60	-.59
C	-.42	-.72	-.73	-.72	-.70
D	.01	-.30	-.37	-.38	-.32
E	-.25	-.26	-.27	-.29	-.29
Index 1′					
B	.45	.78	.56	.55	.58
C	.54	.76	.87	.80	.78
D	.54	.31	.59	.53	.36
E	.72	.58	.55	.56	.73
Index 2′					
B	.41	.66	.46	.50	.50
C	.48	.72	.74	.71	.69
D	.42	.30	.41	.41	.34
E	.60	.49	.38	.42	.54
Index 3′					
B	-.12	-.65	-.46	-.50	-.49
C	-.15	-.72	-.74	-.71	-.69
D	.25	-.30	-.38	-.40	-.31
E	-.24	-.27	-.29	-.35	-.42

pare the indices. This method was chosen as it was the most general of the observed score methods used in this study. Although similar data for the three-parameter IRT method would be interesting to examine for comparisons, adequate expressions for the standard errors unfortunately do not exist.

The standard error estimates shown in Table 3 at selected score points for the equatings are based on those derived for discrete scores and random groups by Lord (1982). The high and low group equatings are independent (different examinees), so the standard error of the difference is the square root of the sum of the two squared standard errors. Estimates of standard errors can be quite variable (Jarjoura & Kolen, 1985), as evidenced in Table 3. For this reason, the standard errors for Figure 1 were smoothed using the Cureton and Tukey procedure (see Angoff, 1982, p. 68), and bands of two standard errors are shown. Standard error estimates for scores of 5 through 35 appear in Figure 1 because the estimates are poor for extreme scores. Because the differences for unsmoothed equipercentile equating sometimes approximate two standard errors of the difference in the figure, it appears that the unsmoothed equipercentile equations are group dependent to some extent. How-

Figure 1
High Group Minus Low Group Equivalents on Four Forms for Linear, Equipercentile, and IRT Equating Methods



ever, the differences are not, in general, substantial.

Conclusions

An examination of the results given in Tables 2 and 3 and Figure 1 leads to the conclusion that none of the methods examined is clearly superior for these data. In general, the methods show negative bias as indicated in Indices 3 and 3′, meaning score equivalents based on low-ability subgroups were higher than the corresponding score equivalents based on high-ability subgroups. Values for Indices 1 through 3 were generally more similar within form than across forms, again pointing to the comparability of the methods. Figure 1 also illustrates the similarity of methods within form. Deviations from a baseline of zero difference be-

Table 3
Estimated Standard Errors of Equating for Forms B, C, D, and E Using Unsmoothed Equipercentile Method

Form and Selected Scores on Anchor Form	Standard Error Using High Ability Examinees	Standard Error Using Low Ability Examinees	Standard Error of the Difference
Form B			
5	1.04	.53	1.17
10	.43	.25	.50
15	.44	.31	.54
20	.41	.38	.56
25	.52	.65	.83
30	.35	.65	.74
35	.29	.90	.95
Form C			
5	.79	.51	.94
10	.33	.25	.42
15	.37	.31	.49
20	.33	.33	.47
25	.38	.53	.65
30	.38	1.10	1.17
35	.36	.40	.54
Form D			
5	.61	.19	.64
10	.44	.24	.50
15	.34	.25	.42
20	.45	.37	.58
25	.32	.49	.59
30	.41	.61	.73
35	.36	.60	.70
Form E			
5	.50	.26	.56
10	.28	.20	.34
15	.37	.32	.48
20	.36	.38	.53
25	.45	.55	.71
30	.42	.86	.96
35	.39	.59	.71

tween high and low ability group equivalents do not appear excessive, and are quite comparable, with no method establishing a clear superiority. In general, all models studied performed similarly, and for all models nonsubstantial differences were found in comparing equatings performed using high-ability and low-ability subgroups. This finding suggests that any of the equating methods studied are robust to even fairly large group differences. However, less test form similarity or greater examinee subgroup disparity than that considered here could lead to more substantial differences among the methods.

One advantage often mentioned for using IRT methods over traditional methods is the "sample-free" specification of item parameters which, if true, would eliminate the concern of the effect of diverse examinee groups on equating. According to Hambleton and Cook (1977, p. 91): ". . .item parameters are invariant across sub-groups of examinees chosen from an examinee population. In principle, item parameters should remain the same regardless of the sub-group tested." Others concur that, as group differences become more marked, IRT methods of equating may offer better results than more traditional methods (Cook & Eignor, 1983b; Marco et al., 1983). However, the results from this study fail to support the assertion that IRT

methods are less population dependent than traditional equipercentile observed score and linear observed score methods in a practical horizontal equating situation. Or, to state this more positively, non-IRT methods were found to be as population invariant as the IRT methods. Thus, population independence may not be a good reason for choosing IRT methods over more traditional methods. Other criteria that might be relevant include the appropriateness of the statistical assumptions for the test forms equated, the properties of the equated scores, and the ease in implementing the method and reporting scores.

References

American College Testing Program. (1973). *Assessing students on the way to college: Volume 1*. Iowa City IA: Author.

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 Holland, P. W., and Rubin, D. B. (Eds.), *Test equating* (pp. 55–69). New York: Academic Press.

Cook, L. L., & Douglas, J. B. (1982). *Analysis of fit and vertical equating with the three-parameter model*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Cook, L. L., & Eignor, D. R. (1983a). *An investigation of the feasibility of applying item response theory to equate achievement tests*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Cook, L. L., & Eignor, D. R. (1983b). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 175–195). Vancouver, British Columbia: Educational Research Institute of British Columbia.

Goulet, L. R., Linn, R. L., & Tatsuoka, M. M. (1975). *Investigation of methodological problems in educational research—longitudinal methodology* (Project No. 4–1114). Urbana-Champaign IL: University of Illinois.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14,* 75–93.

Jarjoura, D., & Kolen, M. J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. *Journal of Educational Statistics, 10,* 143–160.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9,* 25–44.

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of General Educational Development. *Journal of Educational Measurement, 19,* 279–293.

Lord, F. M. (1980). Equating. In F. M. Lord (Ed.), *Applications of item response theory to practical problems* (pp. 193–211). Hillsdale NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics, 7,* 165–174.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT observed-score and true-score "equatings." *Applied Psychological Measurement, 8,* 453–461.

Loret, P. G., Seder, A., Bianchini, J. G., & Vale, C. A. (1972). *Anchor test study: Final report*. Princeton NJ: Educational Testing Service.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179–193.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In Weiss, D. J. (Ed.), *New horizons in testing*. New York: Academic Press.

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

Slinde, J. A., & Linn, R. L. (1977). Vertically equated tests: Fact or phantom? *Journal of Educational Measurement, 14,* 23–32.

Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement, 15,* 23–35.

Slinde, J. A., & Linn, R. L. (1979a). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement, 16,* 159–165.

Slinde, J. A., & Linn, R. L. (1979b). The Rasch model, objective measurement, equating, and robustness. *Applied Psychological Measurement, 3,* 437–452.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST users guide: LOGIST V version 1.0*. Princeton NJ: Educational Testing Service.

Author's Address

Send requests for reprints or further information to Deborah J. Harris, The American College Testing Program, P.O. Box 168, Iowa City IA 52243.