An Application of the Three-Parameter IRT Model to Vertical Equating

Deborah J. Harris
The American College Testing Program

H. D. Hoover
The University of Iowa

This study examined the effectiveness of the three-parameter IRT model in vertically equating five overlapping levels of a mathematics computation test. One to four test levels were administered within intact classrooms to randomly equivalent groups of third through eighth grade students. Test characteristic curves were derived for each grade/test level combination. It was generally found that an examinee would receive a higher ability estimate if the test level administered had been calibrated on less able examinees. Practical implications for "out-of-level" and adaptive testing are discussed.

It is often considered desirable to test a student in a given subject matter area periodically throughout his/her formal schooling, and to compare the scores obtained across the various testings. Because knowledge in many subject areas is closely linked to school curricula, standardized achievement tests are usually developed in levels that attempt to mirror "typical" curriculum placement of different aspects of a subject area. This usually results in a standardized test battery with levels corresponding, at least roughly, to grades in school. In order to compare test scores across these levels, a scale must be developed that allows comparisons of raw scores obtained on tests differing in content and difficulty. This is the problem that vertical equating attempts to solve—how to develop a score scale across test levels which (1) differ in difficulty and (2) are designed for groups of examinees who differ in average ability level.

This study was designed to examine the effectiveness of the three-parameter item response theory (IRT) model in vertically equating the mathematics computation test of the *Iowa Tests of Basic Skills* (Hieronymus, Lindquist, & Hoover, 1977).

IRT methods are frequently suggested as the preferred vertical equating approach for two reasons: (1) It is recognized that problems exist with the classical test theory methods (see, e.g., Lord, 1977; Lord & Wingersky, 1984), and (2) IRT methods are usually conceived of as having "person-free" calibration and "item-free" measurement. These properties imply that the item parameters which are estimated are invariant for all subgroups of examinees, and that, once the items are calibrated, the same θ estimate would be obtained (except for errors of measurement) for an individual regardless of the subset of items he/she was administered. These properties, if they held, would essentially solve the problem of vertical equating.

The two IRT models that have been most prominent in the vertical equating literature are the one-parameter (Rasch) model and the three-parameter model (see, e.g., Hambleton & Swaminathan, 1984). Although the Rasch model possesses certain desirable properties, such as simplicity and a monotonic relationship between raw score and estimated examinee ability, there are indications that the model does not perform well in practice in vertical equat-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 2, June 1987, pp. 151–159
© Copyright 1987 Applied Psychological Measurement Inc.
0146-6216/87/020151-09$1.70

ing situations (see Kolen, 1981; Loyd & Hoover, 1980; Slinde & Linn, 1977).

IRT models require some strong assumptions. One of these, unidimensionality, requires that there be only one trait underlying the examinees' responses to the test items. Although this assumption is never strictly met in practice, there is evidence that IRT equating methods are somewhat robust to violations of it (Cook & Eignor, 1982; Forsyth, Saisangjan, & Gilmer, 1981; Petersen, Cook, & Stocking, 1983). 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 so much how well the data fit the model as how well the model will perform with real data in a real testing situation.

IRT equating methods "characterize equivalent scores on two test forms as those scores which correspond to the same estimated level of latent trait, ability, or skill underlying both tests" (Cook & Douglas, 1982, p. 12). The findings in the literature regarding the effectiveness of IRT applications in vertical equating are somewhat conflicting, though there appears to be some consensus that the Rasch model is not satisfactory for this application (see Divgi, 1981; Holmes, 1982; Kolen, 1981; Loyd & Hoover, 1980; Skaggs & Lissitz, 1986; Slinde & Linn, 1977, 1978, 1979). Some authors suggest that the difficulty with the Rasch model may be less an issue of dimensionality than one of guessing (e.g., Phillips, 1983; Slinde & Linn, 1979). This would provide some support for the use of the three-parameter model, though estimating the pseudo-chance parameter may be troublesome if the data are sparse at the lower end of the θ scale (see Cook & Douglas, 1982; Kolen, 1981). Despite potential estimation difficulties, the three-parameter model has received support for its use, or potential use, in recent literature (see Cook & Douglas, 1982; Kolen, 1981; Marco, Petersen, & Stewart, 1983; Slinde & Linn, 1977).

This study explores the use of the three-parameter IRT model to vertically equate five levels of a mathematics computation test by using elementary school grade as the ability grouping factor. The analyses investigate the item-free and person-free estimation properties. If the equating relationship is dependent upon the examinee sample by which it is estimated, this must be recognized in the interpretation of θ estimates because an examinee's estimate will be a function of the ability of the equating subsample, as well as his/her own ability. In large school systems, where significant proportions of students are frequently tested "out-of-level," the effect on distributions of within-grade derived scores may be substantial.

Method

Data Source

The data for this study were gathered using the *Iowa Tests of Basic Skills* (Hieronymus et al., 1977). Levels 10 through 14 of the Mathematics Computation test were used, corresponding to material typically covered in Grades 4 through 8, respectively. The levels consist of 42 (Level 10) or 45 (Levels 11, 12, 13, and 14) four-option multiple-choice items. Adjacent levels (e.g., Levels 12 and 13) contain 30 common items, while levels one step apart (e.g., Levels 12 and 14) contain 15 common items. Levels two or more steps apart contain no items in common. The examinees were a representative sample of third through eighth grade students in Iowa. From one to four test levels were administered within intact classrooms in a spiraled format to randomly equivalent groups. The grade/test level combinations with corresponding *N* counts are shown in Table 1.

Procedure

The three-parameter IRT model was implemented using LOGIST 5 (Wingersky, Barton, & Lord, 1982), modified for omitted and not-reached items. An initial run, based on all examinees at all test levels, was completed in order to simultaneously obtain θ estimates and item parameter estimates for each of the 3,652 examinees and 102 items. These θ estimates were then used for all subsequent item parameter estimation runs in order to establish a common scale. The item parameter estimates from this

Table 1
Number of Examinees by Grade and Test Level

	Test Level				
Grade	10	11	12	13	14
3	236				
4	232	233			
5	232	230	232		
6		251	254	252	
7			240	250	256
8		79	89	293	293

run were used to obtain test characteristic curves for all test levels, and were viewed as the standard to which comparisons of test characteristic curves based on subsamples of the total examinee sample could be made.

Item parameter estimates, using the total group run or "fixed" θ estimates, were obtained and test characteristic curves were computed for each of the grade/level combinations shown in Table 1. There were some problems in the estimation of item parameters with some of the subsamples. LOGIST 5 failed to converge on six of the runs; in five cases, fixing one item's parameters to the values obtained on the overall criterion run was sufficient to achieve convergence, and in the other case (Grade 4, Level 11) fixing parameters for four of the items was sufficient. One grade/test level combination (Grade 8, Level 11) also contained an item which was answered correctly by all examinees, and therefore the item parameter values were established based on only 44 items. To achieve comparability of test score range, item parameters from the master run were added for the missing item.

In the master run of 102 items, 50 items had lower asymptote values fixed at a common value of approximately .12. The majority of these items occurred in the lowest level, which presumably would contain the easiest items, as the items are arranged on a developmental continuum. For the individual grade × test level runs, the number of fixed lower asymptotes ranged from 12 to 28, with a median of 21. The lower asymptote was fixed at values ranging from .10 to .21, with a mean of .17.

After the item parameters were estimated, test characteristic curves were computed for each grade/test level combination by summing the item characteristic curves over all items in the test at selected θ levels. Because the θs from the overall calibration run were used in all subsequent item estimation runs, the resulting estimates are all on the same scale, and are thus directly comparable.

Results

The means and standard deviations for the θ estimates from the overall LOGIST 5 run are given in Table 2. The within-grade samples are randomly equivalent; therefore, if the item-free measurement property holds, the θ estimates should be comparable across all test levels administered within a given grade. As can be seen from Table 2, the differences among within-grade groups are relatively small with the exception of Grade 8, where the largest difference exceeds .5 standard deviation. This result is hardly surprising, in that a test designed to function on a lower level would not be expected to allow eighth grade examinees to show their full potential. This is not a problem unique to the three-parameter latent trait model, but is rather a question of whether the concept of item-free measurement has meaning across levels in a multilevel achievement test, such as was considered here.

Table 3 presents values of test characteristic curves at selected θ levels based on examinees at each grade/test level combination, and based on the overall run of simultaneous estimation of examinee and item parameters across all levels and examinees. It can be seen that the person-free property of the latent trait model does not hold for these data. An examination of Table 3 reveals that, in general, an examinee will receive a higher θ estimate if the level he/she is administered has been calibrated on younger (i.e., less able) examinees. For example, an examinee with θ = −3 who is administered Level 13, calibrated on sixth grade examinees, would have an expected raw score of 7.8, but if the test level were calibrated using eighth grade examinees, this same examinee would have an expected raw score of 10.1,

Table 2
Means and Standard Deviations for Ability Estimates
by Grade and Test Level

Grade		Test Level 10	11	12	13	14
3	M	-1.78				
	SD	.81				
4	M	-1.04	-1.00			
	SD	.80	.68			
5	M	-.50	-.48	-.42		
	SD	.78	.78	.86		
6	M		-.01	.09	.08	
	SD		.65	.79	.64	
7	M			.46	.51	.58
	SD			.76	.74	.82
8	M		.60	.82	.84	1.03
	SD		.66	.75	.92	.77

Table 3
Values of Test Characteristic Curves
at Selected Ability Levels

Test Level	Calibration Group	Ability Level -3	-2	-1	0	1	2
10	Grade 3	12.0	18.6	30.5	38.9	41.1	41.7
	Grade 4	13.1	19.5	30.9	38.7	41.0	41.3
	Grade 5	14.0	19.7	30.7	39.0	41.3	41.7
	Total	13.0	19.5	30.8	38.9	41.2	41.7
11	Grade 4	11.3	15.9	25.4	36.2	42.0	43.8
	Grade 5	12.3	15.8	25.1	36.0	42.3	44.2
	Grade 6	13.0	17.1	25.7	36.0	42.6	44.5
	Grade 8	17.0	20.3	26.7	35.7	42.4	44.1
	Total	11.3	16.1	25.8	35.8	42.6	44.5
12	Grade 5	9.7	12.6	19.4	30.2	39.9	43.6
	Grade 6	10.5	13.2	20.1	30.1	40.3	43.9
	Grade 7	9.8	13.4	20.3	29.9	40.3	44.1
	Grade 8	13.4	15.5	21.2	30.2	39.9	43.7
	Total	8.7	12.3	19.9	30.1	40.2	44.1
13	Grade 6	7.8	9.5	13.7	21.8	34.6	42.2
	Grade 7	9.2	11.1	14.8	22.2	35.2	43.0
	Grade 8	10.1	11.7	15.3	22.7	35.1	43.0
	Total	7.6	9.6	14.0	22.2	35.3	43.3
14	Grade 7	7.6	8.9	12.1	18.9	30.6	40.3
	Grade 8	9.6	10.7	13.5	19.7	30.5	40.2
	Total	7.3	8.9	12.4	18.9	30.6	40.5

a difference of about 2.3 raw score points. Admittedly, not all examples contained in the table are this extreme, but the trend is pervasive.

Figure 1 illustrates the equatings of Level 12 to Level 13 using five different grade combinations. Levels 12 and 13 were designed to measure material appropriate for sixth and seventh graders, respectively. Figure 1 shows the equating of Level 12 to Level 13 based on sixth graders taking the former level and seventh graders taking the latter. Also shown in Figure 1 are the equatings of Level 12 to Level 13 based on sixth graders taking both levels, and on seventh graders taking both levels. The equatings using seventh graders on Level 12 and sixth graders on Level 13, and the more extreme ''off-level'' case of fifth graders taking Level 12 and eighth graders taking Level 13, are also given.

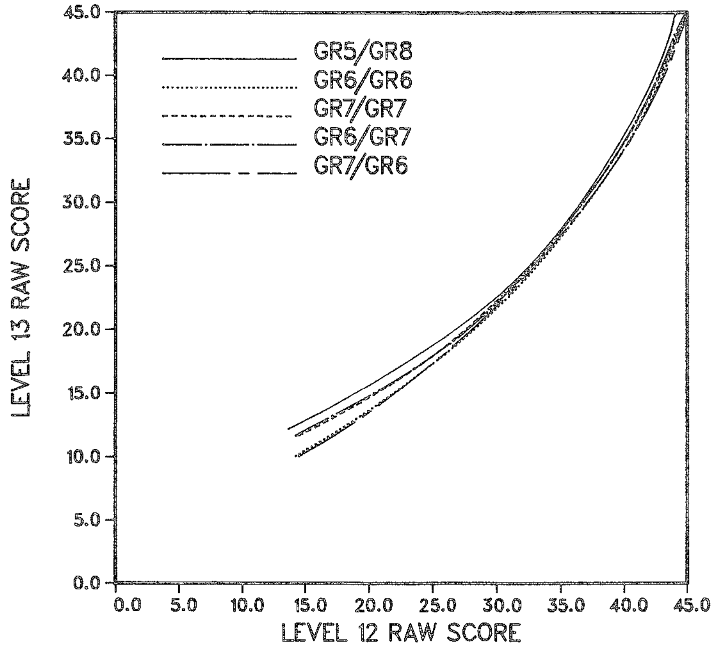
Figure 1 illustrates that the equating relationship does vary with the groups used to establish it. Although the effect on a given examinee may be relatively minor, the effect on class and school averages would be more substantial.

Discussion

Loyd and Hoover (1980) studied vertical equating with the Rasch model using a subset of the data employed here, and using common-item equating (see Loyd & Hoover, 1980, for a description of procedures used). Their study concentrated on Test Levels 12, 13, and 14 and sixth, seventh, and eighth grade examinees. Because the data involved in both the present study and the Loyd and Hoover study are essentially the same, it is interesting to compare the results, though it should be remembered that differences exist in the latent trait model used and the equating method followed. The comparison of results between these two studies, shown in Table 4, concentrates on seventh grade students and the vertical equating of Levels 12 and 14.

The ''direct'' equating procedure involves vertically equating Level 12 to Level 14 through 15 common items. The ''chaining'' procedure involves equating Level 12 to Level 13 through the 30 common items they share, and then equating Level 13 to Level 14 through their 30 common

Figure 1
Equatings of Level 12 to Level 13 Using Different Ability Groups



items. Although this was the equating procedure followed in the Loyd and Hoover study, in the present study Level 12 and Level 14 are directly equated by virtue of the fixed θs used in the estimation procedure. However, in order to provide a better basis for comparison between results of the two studies, direct and chaining equating procedures were also conducted for the present study (see Hambleton & Swaminathan, 1984, chap. 10, for a description of the procedures used). The equipercentile method, perhaps the most common traditional vertical equating method, is also shown in Table 4. While the equipercentile method should not be viewed as the correct standard, the equipercentile method results were given in the Loyd and Hoover study to enhance comparison between IRT (using the Rasch model) and traditional equating procedures. Examination of Table 4 shows the results of the equipercentile and three-parameter latent trait methods to be more similar than the results between the Rasch and three-parameter methods, a finding shared by Kolen (1981) and Phillips (1983).

Both the present results and those of Loyd and Hoover (1980) support the conclusions of Slinde and Linn (1978, 1979), Goulet, Linn, and Tatsuoka (1975), Patience (1981), and others: IRT models, at least those examined in the contexts of the above studies, do not yield person-free calibration. It was found that the equatings were dependent on the ability of the examinee group. The present study supports the findings of Loyd and Hoover (1980): "For pupils who take an easier (lower) level of the test and have their scores equated to a more difficult level, the resulting scores will be more favorable, i.e., higher, when the equating is based on the higher ability group. For pupils who take a more difficult (higher) level of the test and then have their scores equated to an easier (lower) level, the resulting scores will be more favorable, i.e., higher, when the equating is based on the lower ability group" (p. 188).

Possible explanations for the discrepancy of results could include questions as to (1) the dimensionality of the dataset and (2) the goodness of fit of the IRT parameter estimates. Because the intent of the present study was to examine the usefulness of IRT methodology to vertically equate achievement test data in its current state (regardless of the violation of IRT assumptions), factor analysis of the data was not performed. However, a factor analysis performed by Loyd and Hoover (1980) on Level 13 for the combined seventh and eighth grade examinees may provide some insight into the results obtained here. Using a principal axis method, Loyd and Hoover found the percentage of variance accounted for by the first four factors to be 34.7, 9.0, 6.8, and 3.9, respectively, and concluded that the unidimensionality assumption was not strictly met (see Loyd & Hoover, 1980, pp. 188–190).

Yen's (1981) Q_1 goodness-of-fit statistic was computed to examine the fit of the three-parameter IRT model to the data used. (See Yen, 1981, for a description of the statistics.) From the overall master run of 102 items, 36 items were found to have Q_1 values significant at the .01 level. Of these, 4 items achieved their high Q_1 value primarily due to misfit in the first decile, 17 due to misfit in the 10th decile, and 3 due to misfit in both the first and last decile, suggesting (as was perhaps expected) lack of fit, primarily in the tails.

The majority of items with large Q_1 values occurred where the greatest overlap between levels occurred, and therefore, because of the way this study was structured, where the most diversity among examinees occurred. For the first and last item sections (which appeared only on Level 10 and only on Level 14, respectively), there were 1 and 0 items with significant Q_1 values, respectively; the middle sections (which appeared on three test levels) had up to 10.

Although there may be some question as to the relative accuracy of the Q_1 statistic (see, e.g., McKinley & Mills, 1985), such a large proportion of misfitting items may be a partial explanation for the behavior of the IRT model examined here. Again it should be stressed that the purpose of this study was to examine the results of IRT methodology in vertically equating an existing mathematics achievement test, and not to determine how well the existing data fit the rather strong IRT assumptions.

Practical Implications

Achievement tests designed for elementary school examinees contain different levels for the express purpose of providing accurate measurement of examinees throughout a developmental range. For scores to be useful, a single score scale must be developed that spans this entire range; the function of vertical equating is to establish this underlying scale. Once this scale is established, out-of-level testing becomes possible for those examinees either too advanced or too retarded in their development to be correctly assessed with the "on-level" test. Despite the fact that these examinees are not given the on-level test, it is desirable to assign on-level percentile ranks to their test results, and to include these examinees in the computation of classroom and school averages. Such applications require a common scale.

Loyd and Hoover used the example of an eighth grade examinee testing two levels below on-level, a common practice in many schools. This examinee receives a raw score on Level 12, which is then converted to an "equivalent" score on Level 14, the on-level test for eighth graders. Table 4 illustrates Level 14 equivalents for several Level 12 raw scores, using various equating methods. It is perhaps most instructive to use the national percentile ranks to provide a basis for interpreting the results. For example, an out-of-level examinee earning a raw score of 30 on Level 12 would receive a Grade 8 percentile rank of 13 using the Rasch direct method, a rank of 10 using the Rasch chaining method, a rank of 42 using the three-parameter model (equating through fixed θs), and a rank of 37 using the equipercentile method. Clearly, the difference between percentile ranks of 13 and 42 is nontrivial, and the method of vertical equating used to establish the score scale must be considered in interpreting out-of-level score results.

It is interesting to note the differences across the 3 three-parameter methods shown. Because the θs for the item parameter estimation runs were fixed from a single calibration run, Levels 12 and 14 are directly equated; differences between the θ column in Table 4 and the direct and chaining methods

Table 4

A Comparison of Rasch, Three-Parameter IRT and Equpercentile Equating of Level 12 to Level 14 with Seventh Grade Examinees

	EQUATING METHOD									
	Rasch Direct*		3-Parameter Direct		Rasch Chaining*		3-Parameter Chaining		3-Parameter θ		Equipercentile*	
Level 12	Lv 14	Gr 8	Lv 14	Gr 8	Lv 14	Gr 8	Lv 14	Gr 8	Lv 14	Gr 8	Lv 14	Gr 8
Raw Score	RS	%ile	RS	%ile	RS	%ile	RS	%ile	RS	%ile	RS	%ile
20	6.0	1	11.8	11	5.4	1	10.8	8	11.9	11	12.7	14
25	8.8	4	14.6	23	8.1	2	14.0	19	15.0	25	15.1	25
30	12.3	13	18.2	39	11.4	10	18.4	40	18.9	42	17.7	37
35	17.3	35	22.5	58	16.2	30	24.1	63	23.8	63	21.8	56
40	24.7	66	28.2	78	23.8	63	31.9	87	30.1	83	27.4	75

*Obtained from Loyd & Hoover (1980).

should be minimal. While the discrepancies between the fixed θ method and the direct and chaining methods appear relatively small in terms of raw scores, the differences when interpreted in terms of percentile ranks may not be small, especially when used in computing classroom or school averages. It should also be pointed out that because the same θs were used for all item parameter estimations, equating errors usually associated with problems in estimating the θs were minimized. The magnitude of such errors was recently explicated by Skaggs and Lissitz (1986), who found it to be substantial.

Conclusions

This study examined the three-parameter IRT model in the context of vertically equating five levels of a mathematics computation test. The results show that the model's properties of item-free measurement and person-free calibration do not hold for these data. Findings from this study were compared to findings from Loyd and Hoover (1980), a study which used a subset of the same data and employed the Rasch model to vertically equate three levels of the mathematics computation test. Jointly, these results indicate that the method used to establish the vertical equating profoundly influences the results obtained, and that caution should be employed in using IRT models in situations such as those described here, although better model fit, and perhaps consistency of results, could be obtained if the test involved were constructed under the assumption of viewing model fit, rather than test content, as sacrosanct.

It should also be pointed out that these results have direct implications to other applications that rely on the item-free or person-free properties of item response theory. One is adaptive testing, the limiting case of out-of-level or individualized testing. Another is the combined use of IRT and matrix sampling proposed for many large-scale assessment programs, such as the National Assessment of Educational Progress.

References

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. (1982). *Score equating and item response theory: Some practical considerations*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Divgi, D. R. (1981). *Does the Rasch model really work? Not if you look closely*. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles.

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement, 5*, 175–186.

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., & Swaminathan, H. (1984). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hieronymus, A. N., Lindquist, E. F., & Hoover, H. D. (1977). *Iowa Tests of Basic Skills*. Boston: Houghton Mifflin.

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement, 19*, 139–147.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*, 1–11.

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

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

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 D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 147–177). New York: Academic Press.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49–57.

Patience, W. M. (1981). *A comparison of latent trait and equipercentile methods of vertically equating tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137–156.

Phillips, S. E. (1983). Comparison of equipercentile and item response theory equating when the scaling test method is applied to a multilevel achievement battery. *Applied Psychological Measurement, 7,* 267–281.

Skaggs, G., & Lissitz, R. W. (1986). An exploration of the robustness of four test equating models. *Applied Psychological Measurement, 10,* 303–317.

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. (1979). 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.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (LOGIST 5, version 1).* Princeton NJ: Educational Testing Service.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245–262.

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, U.S.A.