Effect of Examinee Ability on Test Equating Invariance

Gary Skaggs
Fairfax County, Virginia Public Schools

Robert W. Lissitz
University of Maryland

Previous research on the application of IRT methodology to vertical test equating has demonstrated conflicting results about the degree of invariance shown by these methods with respect to examinee ability. The purpose of this study was to examine IRT equating invariance by simulating the vertical equating of two tests under varying conditions. Rasch, three-parameter, and equipercentile equating methods were compared. Six equating cases, using different sets of item parameters, were replicated based on examinee samples of low, medium, or high ability or where ability was matched to the difficulty level of the test. The results showed that all three methods were reasonably invariant to examinee ability level under all conditions imposed. This suggests that multidimensionality is likely to be the cause of the lack of invariance found in real datasets. *Index terms: Examinee ability; Invariance in item response theory; Item response theory, equating; Item response theory, invariance; Test equating; Vertical equating.*

The equating of scores on two or more tests is fundamental to large testing programs. Scores on new forms of tests must be equated to scores on previous versions of the same test; this type of equating is typically referred to as "horizontal" because new test forms are designed to be as similar as possible to previous forms in terms of their psychometric properties. The need for a different kind of equating, called "vertical" equating, arises when it is desired to link scores across levels of achievement tests. These levels are likely to be targeted to grades in school, and tests corresponding to different levels differ intentionally both in content and difficulty. Obviously, from theoretical and operational perspectives, vertical equating is much more difficult to accomplish than horizontal equating.

In recent years, a considerable quantity of research has investigated the application of item response theory (IRT) methodology to both types of test equating. Although a large number of IRT models have been proposed, the research on test equating has been primarily limited to two models—the one-parameter logistic (Rasch) model and the three-parameter model.

According to the Rasch model, the probability of an examinee with ability θ (defined generically) answering item i correctly is solely a function of the item's difficulty (b_i) and the examinee's ability (θ):

$$P_i(\theta) = [1+\exp(\theta - b_i)]^{-1} \quad . \tag{1}$$

The three-parameter model incorporates two additional item parameters:

$$P_i(\theta) = c_i + (1-c_i)\{1+\exp[-Da_i(\theta - b_i)]\}^{-1} \quad , \tag{2}$$

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 1, March 1988, pp. 69–82

0146-6216/88/010069-14$1.95

where a_i is the item's discrimination and c_i is the item's lower asymptote (pseudo-guessing parameter). D is a scaling constant, usually set to 1.7.

The major difference between these two models concerns the Rasch model's assumptions of equal discriminations and lower asymptotes equal to 0. Although multiple-choice tests rarely meet these assumptions, estimation and equating are far simpler, computer costs are considerably less, and smaller samples may be used for the Rasch model. For further discussions of these two models, the reader is referred to Hambleton and Swaminathan (1985), Lord (1980), Wright (1977), and Wright and Stone (1979).

Although research investigating the use of these two models for equating has come to conflicting conclusions about their effectiveness, several trends have emerged. The Rasch model appears not to work well for vertical equating of multiple-choice tests (Holmes, 1982; Kolen, 1981; Loyd & Hoover, 1980; Marco, Petersen, & Stewart, 1983; Slinde & Linn, 1978, 1979). Skaggs and Lissitz (1986) demonstrated that for vertical equating, the Rasch model was less robust to violations of its assumptions than for horizontal equating. In horizontal equating, for example, the Rasch model appears to be relatively immune to deviations from equal item discriminations within a test (Curry, Bashaw, & Rentz, 1978), but not from differences in *mean* discrimination between tests (Divgi, 1981; Forsyth, Saisangjan, & Gilmer, 1981; Skaggs & Lissitz, 1986).

The three-parameter model has generally demonstrated better results for vertical equating than the Rasch model (Kolen, 1981; Marco et al., 1983). However, the effectiveness of three-parameter model equating has been shown to vary considerably according to test length, sample size, estimation and linking strategies, and test content (Cook & Eignor, 1983; Holmes & Doody-Bogan, 1983; Kolen, 1981; Kolen & Whitney, 1982; Petersen, Cook, & Stocking, 1983).

Many questions therefore remain concerning the use of IRT equating methods, particularly in the case of vertical equating. One of the major theoretical advantages of IRT is that the equating results should be independent of the examinees used to develop the equating function. This is a necessary characteristic for successful equating (see Angoff, 1984), but IRT methods promise to be more invariant than conventional methods of equating (Lord, 1980, chap. 13). However, if the assumptions of the models are violated, this may not be true. The invariance property of IRT is, in theory, valid for any subgroup from a given population. The question, from an operational point of view, is whether different groups of examinees are samples from the same population or are samples from different populations altogether (see Cook & Petersen, 1987).

The major criticism of vertical equating with the Rasch model is that samples of examinees at different ability levels produce substantially different equating functions (Holmes, 1982; Loyd & Hoover, 1980; Slinde & Linn, 1978, 1979). Recently, several investigators have examined equating invariance using the three-parameter model with real test data for both horizontal and vertical equating.

Cook, Eignor, and Taft (1984) examined the stability of item parameter estimates for items administered as parts of biology achievement tests but at two different times. The differences in administration time reflected differences in recency of instruction for examinees. Their results suggested that the same items measured different attributes when given at different times, thus implying that the two samples represented different populations. Anderson (1985) obtained similar results for a battery of achievement tests administered to Head Start children. Two test forms with overlapping items were administered in consecutive years. Results indicated high correlations between difficulty estimates but large differences between the actual estimates for common items.

Harris and Kolen (1986) compared equating functions based on linear, equipercentile, and three-parameter model methods for alternate forms of a mathematics achievement test. They studied invariance by computing the equating separately for samples of high and low ability in a manner similar to the Slinde

and Linn, Loyd and Hoover, and Holmes studies mentioned above. Their results showed that all three methods produced very similar equatings. There was a small difference between high- and low-ability equatings, but this bias was replicated for all three methods.

On the other hand, Harris and Hoover (1987) examined vertical equating between levels of a mathematics achievement test. They replicated the equating of tests at adjacent levels using students at different grade levels. The results indicated that "the method used to establish the vertical equating profoundly influences the results obtained" (p. 158). In general, an examinee would receive a higher score if the test had been calibrated on less able examinees, a result consistent with Rasch model equating research.

The above studies taken together suggest that vertical equating presents more of a problem in terms of invariance than does horizontal equating. This seems to hold true for the three-parameter model as well as for the Rasch model. Such a finding in turn suggests that multidimensionality across test levels is a major factor in the lack of invariance.

The purpose of this study was to examine in a different way the effect on vertical equating of the ability level of the examinees on which the equating is based. Because the previous studies were empirical in nature and employed actual test data, it was impossible to manipulate the extent of data fit to the IRT models. In this study, equipercentile, Rasch, and three-parameter model equating methods were compared using data generated to simulate the vertical equating of two tests under varying conditions.

Method

Equating Methods

For all cases in this study, an external anchor test design was used. Under this design, one sample of simulated examinees provided responses to one test (Test A) plus another, shorter test called an anchor test. A second sample provided responses to a second test (Test B) and to the same anchor test. The anchor test was considered external to the other tests, that is, these items were not included in the scoring of either Test A or Test B.

Three equating methods were compared in this study—equipercentile, Rasch model, and three-parameter model equating. These methods differ in an important way. Rasch and three-parameter model methods are based on IRT, which uses individual item responses and assumes that the item responses conform to a specific model. Equipercentile equating is an empirical method based on observed scores. With this method, individual item information is not used, nor is any specific model for the data assumed. The equating cases in this study involved violations of the assumptions of the Rasch model, but not of equipercentile equating.

Equipercentile equating. For an equipercentile equating of two tests based on groups of equal ability, equivalent raw scores are derived through the computation of the cumulative raw score distributions of each test. Equal percentile ranks denote equal ability, and the two tests are equated according to raw scores that correspond to the same percentile ranks. This study employed a method referred to as frequency estimation (Design IVB in Angoff, 1984; originally developed by Levine, 1958). This method is designed for use in an anchor test design with samples which are not necessarily equal in ability. The method has not been used or examined extensively because it has been thought to require very large samples for even modestly accurate results. However, Jarjoura and Kolen (1985) and Skaggs and Lissitz (1986) have provided evidence of positive results with this method.

The basis for frequency estimation equipercentile equating is the estimation of raw score distributions for each test from a hypothetical combined group of examinees. In this study, these distributions were smoothed using the Cureton-Tukey (1951) rolling weighted average algorithm. Once this was accom-

plished, a set of equivalent raw scores was developed whereby each pair of scores was linked to the same percentile rank.

Rasch model equating. Although true-score equating can be applied to both the Rasch and three-parameter models, there are several operational differences between procedures based on the two models. The equating procedures described by Wright and Stone (1979) were used for this study for the Rasch model.

Item and ability parameters were estimated separately for each test/anchor combination using the BICAL program (Wright, Mead, & Bell, 1980). Two sets of item difficulties were thus produced for the anchor test items, and these were used to place the item difficulties from one test on the same scale as the other. According to Rasch model theory, the two sets of estimates differ only by a constant, which can be computed as the mean difference in difficulty across the items on the anchor test. Because Test B was equated to Test A, this constant was added to each of the item difficulties of Test B to place them on the same scale as the item difficulties from Test A.

Once this was done, equivalent scores on the two tests were computed as a function of equal ability in the following manner:

$$\varepsilon = \sum_i P_i(\theta) \tag{3}$$

$$\xi = \sum_j P_j(\theta) \quad , \tag{4}$$

where ε and ξ are the number-correct true scores on Tests A and B, respectively, corresponding to the same ability θ. $P_i(\theta)$ and $P_j(\theta)$ are the estimated probabilities for correct responses to items i and j under the Rasch model. The summations are across the items of the two tests.

For each possible raw score on Test B (except 0 and perfect scores), a Test A equivalent was computed in the following manner. First, Equation 4 was solved for θ using Newton's iterative method. This value was then substituted into Equation 3 to obtain the Test A equivalent score. In other words, Test B was equated to Test A.

Three-parameter model equating. For the three-parameter model, all item and ability parameters were estimated using the LOGIST 5 program (Wingersky, Barton, & Lord, 1982). The version used in this study was adapted for use on a UNIVAC 1100 with all assembly language routines rewritten in FORTRAN 77. For each equating, only one LOGIST run was required because items not answered by one of the samples were coded as "not reached." As a result, item parameters from the two tests were automatically placed on the same scale.

Computing equivalent scores proceeded in a manner similar to the Rasch model by using Equations 3 and 4, with one modification. For any θ, it is not possible to compute an expected true score less than the sum of the c_is, or below chance level. In order to account for below-chance raw scores (except 0), a linear extrapolation developed by Lord (1980, pp. 210–211) was used to obtain raw score equivalents.

Ability Parameters

Sample sizes for each test/anchor combination were 2,000. This sample size was intended to be large enough to provide accurate estimation for all three equating methods (see Hulin, Lissak, & Drasgow, 1982; Jarjoura & Kolen, 1985). The manipulation of examinee ability level was the major independent variable of this study. This was done in four ways for each equating case:

1. Mean ability for both samples was −.5 logits;
2. Mean ability for both samples was 0.0 logits;
3. Mean ability for both samples was .5 logits;

4. Mean ability was −.5 logits for the sample taking Test A, and .5 logits for the sample taking Test B.

In the first three cases, equating was based on samples of high, medium, and low ability. These cases replicate those found in the Slinde and Linn, Loyd and Hoover, and Holmes studies described above. In the fourth case, the ability of each sample was matched to the difficulty level of the test, probably the most common situation for vertical equating.

For all four ability levels, initial abilities for each of the two groups were sampled from a normal distribution with a standard deviation of 1 using the GGNML generator (IMSL, 1980). It is commonly held that the variance of scale scores increases with ability, and this perhaps should have been reflected in the standard deviation of initial abilities used in this study. However, Yen (1986) made a strong argument that such an increase in variance may be a function of scaling techniques. In her study, an IRT scaling procedure resulted in a *decrease* in the variance of scale scores as grade level increased.

Item Parameters

Difficulty. This study simulated the vertical equating of two tests through an anchor test design. Test B, a more difficult test, was equated to Test A, an easier test. That is, for each raw score on Test B (except 0 and perfect scores), an equivalent was calculated on the Test A raw score scale using one of the three equating methods. For the item parameters used to generate response data, the mean difficulty was −.5 logits for Test A and .5 logits for Test B. Each of these tests was 35 items in length, and the difficulties were equally spaced in an interval of ±2 logits.

The level and spacing of difficulties was designed to approximate an equating of adjacent levels of an achievement test battery, where item difficulty is spread across a range sufficient to measure most examinees. These levels of item difficulty are consistent with other vertical equating studies using established achievement test batteries. While many published tests are considerably longer than 35 items, this length is typical of many subtests within an achievement test battery. Preliminary work for this study indicated that 35 items were sufficient to provide enough score points to assess equating accuracy.

Discrimination. In contrast to ability and item difficulty parameters, item discrimination and lower asymptote parameters were manipulated in ways not usually found in published tests. This was done in order to violate systematically the assumptions of the Rasch model. Item discrimination parameters were generated in two ways. In the first case, both Tests A and B had a mean discrimination of .8, a typical value found in the literature. In the second case, the items for Test A had a mean discrimination of .5, and Test B had a mean discrimination of 1.1. This case was a clear violation of the equal discrimination assumption of the Rasch model for equating. It was thought that this would provide a severe test for the equipercentile and three-parameter methods as well.

In both cases, the item discriminations within each test were uniformly distributed across a range of ± .1. Item discriminations are usually thought to follow a skewed or peaked rather than a uniform distribution. Nevertheless, Skaggs and Stevenson (1986) showed through simulation that item discrimination estimates will be skewed even when the initial discriminations are generated from a uniform distribution. The range used here is admittedly more narrow than would normally be found. Divgi (1981) pointed out that unequal mean discrimination is more likely to bias Rasch equating than unequal or large variances of item discrimination. The purpose here was to have one case that approximated fit to the Rasch model and one case that distinctively violated the model.

Lower asymptote. The lower asymptote (c_i) parameter was manipulated in three ways. First, all cs for both tests were set equal to 0. This gave an equating case consistent with the Rasch model. Second, all cs for both tests were set equal to .2, introducing an equal level of chance scoring into both tests.

Third, all *c*s for Test A (the easier test) were set to 0 and all *c*s for Test B (the more difficult test) were set to .2. Both of the latter cases involved a violation of the Rasch model. In the second case, both tests violated this assumption, while in the third case, chance responding occurred only on the more difficult test, where more guessing would be expected to occur.

By crossing the two levels of item discrimination parameters with the three levels of lower asymptote parameters, the study comprised a total of six equating cases. Each of these cases was replicated four times according to different ability levels, as described above.

Anchor test items. In all cases, the anchor test consisted of 15 items. Item difficulty was uniformly distributed between -1.0 and 1.0 logits. The decision to make the anchor test external to Tests A and B was a matter of convenience. The range of anchor item difficulty fits within the range of item difficulty on each test, and would be similar to the range of item difficulties if the anchor test were internal (i.e., overlapped) to both tests. Item discrimination was uniformly distributed between .7 and .9. All lower asymptotes were 0 on the anchor test when they were 0 on the two tests to be equated. Otherwise, all *c*s on the anchor tests were equal to .2.

Generation of Item Responses

Data for this study were generated from the following three-parameter function:

$$P(u_{ij} = 1|\theta_j,a_i,b_i,c_i) = c_i+(1-c_i)\{1+\exp[-1.702a_i(\theta_j-b_i)]\}^{-1} \quad , \tag{5}$$

where P is the probability of a correct response (u_{ij}) to item i by person j conditional on item and ability parameters. θ_j is the ability of person j, and a_i, b_i, and c_i are the item parameters, as specified earlier.

The response to item i by person j, a 0 or 1, was determined by comparing the probability defined above to a random number drawn from a (0,1) uniform distribution. If the probability of a correct response exceeded the random number, the item was scored as correct. Otherwise, the item was scored as incorrect. The random numbers were produced from the GGUBS generator (IMSL, 1980).

Evaluation of Equating Results

The initial item parameters were used to develop the criterion equating function against which the three methods were compared. This criterion was simply a pairing of raw scores corresponding to the same abilities, using the relationship in Equations 3 and 4. Two summary statistics were used to interpret the results. These statistics were used by Skaggs and Lissitz (1986) and are very similar to total error indices found in other equating studies. They are referred to here as the unweighted mean square error (UMSE) and weighted mean square error (WMSE) and are defined as follows:

$$\text{UMSE} = \frac{\sum_{i=1}^{k-1}(X_E-X_C)^2}{S_B^2} \tag{6}$$

$$\text{WMSE} = \frac{\sum_{i=7}^{k-1}f_i(X_E-X_C)^2}{S_B^2\sum_{i=7}f_i} \quad , \tag{7}$$

where k is the number of items on Test B (35),
S_B^2 is the raw score variance for Test B,
X_C is the criterion test score equivalent on Test A for raw score i on Test B,

X_E is the Test A equivalent produced by one of the equating methods, and
f_i is the frequency of raw score i on Test B.

When Test A equivalents were computed using one of the equating methods, fractional parts of score units usually resulted. In the above formulas, X_C and X_E were therefore in decimal rather than integer form.

The two indices represent two different ways of evaluating equating accuracy. The UMSE provides an indication of how well equating worked across the entire raw score scale, regardless of where on the scale most examinees fell. The summation was across the possible range of raw score values, excluding 0 and perfect scores, which in this study corresponded to raw scores from 1 to 34. Skaggs and Lissitz (1986) found that the equating methods studied here tend to minimize error in the regions where examinees are concentrated, frequently resulting in large equating errors in the unpopulated extremes of the raw score scale. If the equating functions were to be applied to a different ability group, serious errors could occur. The purpose of the unweighted index was to provide a measure of overall equating error independent of examinee ability.

The WMSE shows how well equating worked for each examinee who did not score at a below-chance level. This index is identical to that found in several previous equating studies (e.g., Marco et al., 1983; Petersen et al., 1983). For the WMSE, the summation was only across that part of the scale where extrapolation was not necessary—in this study, raw scores from 7 to 34.

Results

Table 1 shows means and standard deviations of the raw score distributions for Tests A and B. The means ranged from approximately 10.6 to 26.2 and standard deviations ranged from 5.0 to 7.7, depending on how well the test was matched to the ability of the sample and on the level of item discrimination and chance scoring. Where mean ability was equal to mean item difficulty, and when no chance scoring was present, raw score means were close to the expected value of 17.5 (out of 35). Raw score distributions generated under these conditions showed a high degree of consistency in their first two moments.

As would be expected, a higher degree of item discrimination in the generating item parameters produced more dispersion in the raw score distributions. The reverse was true for low discrimination. Nonzero lower asymptotes ($c = .2$) produced negatively skewed raw score distributions and slightly higher means. The results in Table 1 show values which were within expectations given the generating item and ability parameters, as well as a high degree of consistency between raw score distributions that were based on similar item and ability parameters.

Table 2 shows the results of the equatings for the three cases where mean test discriminations were equal across Tests A and B. Because the distributional properties of the MSE statistics are not known, it is difficult to know precisely how large a difference in MSE values is statistically significant. These indices were standardized by the raw score variance on Test B; consequently, they are expressed on a score variance scale. A value greater than .05 (i.e., 5% of the variance) on either statistic might be considered a crude indication of poor equating results, but clearly this is a value judgment to be made ultimately by the user.

Case 1 was a situation where the data reasonably conformed to the Rasch model across both tests. All MSE indices were less than .02. The Rasch model indices were all less than .001, indicating that the model worked extremely well when the data fit the model. All of the methods worked quite well no matter which level of examinee ability was used to perform the equating.

Table 1
Raw Score Means and Standard Deviations for Equating Cases

a	c	θ	Mean	S.D.	a	c	θ	Mean	S.D.
.8	.0	-.5	17.66	7.21	.8	.0	-.5	11.36	6.64
		0	20.85	6.98			0	14.08	6.84
		.5	24.15	6.36			.5	17.79	7.07
		-.5	17.54	7.08			.5	17.42	7.05
.8	.2	-.5	21.14	5.99	.8	.2	-.5	15.93	5.61
		0	23.82	5.80			0	18.21	5.87
		.5	26.17	5.26			.5	20.87	5.90
		-.5	20.92	5.94			.5	20.99	5.96
.8	.0	-.5	17.76	7.07	.8	.2	-.5	15.95	5.52
		0	20.90	7.00			0	18.41	5.92
		.5	24.21	6.53			.5	21.03	5.91
		-.5	17.57	7.21			.5	21.01	5.84
.5	.0	-.5	17.39	6.17	1.1	.0	-.5	10.62	6.74
		0	20.06	5.92			0	13.80	7.28
		.5	22.97	5.67			.5	17.64	7.72
		-.5	17.63	5.93			.5	17.41	7.60
.5	.2	-.5	20.99	5.16	1.1	.2	-.5	15.21	6.02
		0	22.96	5.02			0	18.22	6.19
		.5	25.28	4.66			.5	21.09	6.35
		-.5	21.16	5.07			.5	20.98	6.37
.5	.0	-.5	17.54	6.05	1.1	.2	-.5	15.23	5.93
		0	20.30	5.83			0	18.06	6.33
		.5	22.87	5.49			.5	21.06	6.16
		-.5	17.64	6.04			.5	21.16	6.27

For Case 2, an equal degree of chance scoring was introduced in both tests. Here, three-parameter model equating produced the best results, with all MSE values less than .01. Equipercentile equating MSEs were not as good as those for the three-parameter model. WMSEs were less than .05, but ranged from .04 to .08 for the UMSE. Rasch model equating was clearly inadequate, with UMSEs ranging from .12 to .15 and WMSEs ranging from .02 to .06. However, as in Case 1, the results for all three methods appeared to be uninfluenced by the ability level of the samples.

In Case 3, where chance scoring was introduced only in the more difficult test, results from all three methods were worse than in the previous cases, especially for the Rasch model. All three-parameter model MSE values were less than .05 but greater than they were for Cases 1 and 2. Equipercentile equating MSEs were very similar to those for Case 2. WMSEs were less than .03, but UMSEs ranged from .03 to .08. As in previous cases, the results seemed relatively unaffected by ability level.

The results for cases where mean test discriminations were unequal are presented in Table 3. In all of these cases, the more difficult test was more highly discriminating than the easier test. As was the case in Table 2, the MSE values varied considerably between equating methods but were similar across different levels of ability.

In Table 3, the values for the UMSEs are, for the most part, larger than their corresponding WMSEs, more so than in Table 2. This occurred for two reasons. First, all three methods were fairly successful in minimizing equating error in the regions where most Test B examinees fell, and this is reflected in the WMSE statistic. Second, all three methods produced greater equating error in the lower region of the raw score scale where relatively few examinees scored.

Table 2
Unweighted Mean Square Error (UMSE)
and Weighted Mean Square Error (WMSE) for
Vertical Equating When Test Discriminations Are Equal
($\bar{a}_A=\bar{a}_B=.8$; $\bar{b}_A=-.5$; $\bar{b}_B=.5$)

			Equi-percentile		Rasch		Three-Parameter	
Case	θ_α	θ_β	UMSE	WMSE	UMSE	WMSE	UMSE	WMSE
1: $c_A=0$, $c_B=0$								
	-.5	-.5	.001	.000	.000	.000	.008	.000
	0	0	.003	.001	.000	.000	.006	.001
	.5	.5	.001	.001	.000	.000	.008	.001
	-.5	.5	.018	.016	.000	.000	.005	.001
		S.D.	.007	.007	.000	.000	.001	.000
2: $c_A=.2$, $c_B=.2$								
	-.5	-.5	.041	.014	.136	.058	.001	.001
	0	0	.030	.009	.132	.037	.002	.001
	.5	.5	.079	.008	.150	.021	.007	.002
	-.5	.5	.083	.045	.119	.017	.001	.001
		S.D.	.023	.015	.011	.016	.002	.000
3: $c_A=0$, $c_B=.2$								
	-.5	-.5	.053	.026	.662	.363	.047	.036
	0	0	.077	.026	.549	.231	.037	.020
	.5	.5	.080	.010	.607	.141	.090	.018
	-.5	.5	.029	.029	.658	.147	.034	.011
		S.D.	.021	.007	.046	.090	.022	.009

In Case 4, where there was no chance scoring present in either test, and where mean discriminations were unequal between the two tests, equipercentile equating was the only method to produce acceptable results. Rasch equating produced very large errors, indicating little robustness to this violation of the model's assumptions. Three-parameter model equating also produced questionable results, despite the fact that the data were generated from a three-parameter model.

For a constant level of chance scoring across both tests, as in Case 5, none of the methods performed well. Three-parameter equating provided the best results, while both equipercentile and Rasch equating produced poor results. In Case 6, where chance scoring was present only in the more difficult test, only three-parameter model equating was marginally adequate. Cases 5 and 6 show that chance scoring and unequal discrimination interact with each other to produce somewhat unpredictable results for each of the methods. All three methods had difficulty equating tests of extremely different properties. Tests with the properties found in these two cases are rarely seen in practice, but these results indicate that chance scoring and discrimination must be taken into account in any test development effort.

These results generally confirm those reported by Skaggs and Lissitz (1986), under similar conditions, concerning the adequacy of equatings produced by the three methods studied here. Unequal discrimination and chance scoring interacted with each other to affect equating results in complex ways. For vertical equating, the previous study focused only on the situation where both tests were matched to the level of examinee ability. In both studies, Rasch equating produced acceptable results only when mean discriminations were equal and all lower asymptotes were 0. Three-parameter model equating gave the best overall performance, which was to be expected because the data were generated from a three-parameter

Table 3
Unweighted Mean Square Error (UMSE)
and Weighted Mean Square Error (WMSE) for
Equating When Test Discriminations Are Unequal
($\bar{a}_A$=.5, $\bar{a}_B$=1.1; $\bar{b}_A$=-.5; $\bar{b}_B$=.5)

Case	θ_α	θ_β	Equi-percentile UMSE	Equi-percentile WMSE	Rasch UMSE	Rasch WMSE	Three-Parameter UMSE	Three-Parameter WMSE
4: c_A=0, c_B=0								
	-.5	-.5	.028	.002	.381	.150	.118	.030
	0	0	.009	.002	.343	.164	.096	.034
	.5	.5	.005	.002	.343	.170	.087	.040
	-.5	.5	.021	.012	.314	.189	.083	.044
	S.D.		.009	.004	.024	.014	.016	.005
5: c_A=.2, c_B=.2								
	-.5	-.5	.123	.026	.082	.036	.044	.040
	0	0	.104	.013	.067	.054	.058	.056
	.5	.5	.139	.008	.090	.070	.059	.050
	-.5	.5	.063	.020	.087	.093	.044	.042
	S.D.		.028	.007	.009	.021	.007	.006
6: c_A=0, c_B=.2								
	-.5	-.5	.123	.040	.174	.051	.080	.047
	0	0	.145	.030	.131	.031	.061	.036
	.5	.5	.128	.006	.115	.034	.067	.028
	-.5	.5	.066	.023	.195	.102	.050	.037
	S.D.		.030	.012	.032	.028	.011	.007

model. Equipercentile equating showed surprisingly good robustness to the conditions simulated in both the earlier study and this one. The results for this method were better than those for the three-parameter model for one of the cases in this study, and were reasonably close on all others.

All three methods produced their worst results when mean test discriminations were unequal combined with some degree of chance scoring in the design. The differences in mean discrimination in this study were quite severe, more than would be expected in practice. How well the methods would perform under less stringent conditions is not known.

Discussion

The primary focus of this study was the effect of examinee ability on equating results. All three methods seemed to be relatively robust to differences in ability level of the sample taking each test. Within each case, for a given method, the differences between smallest and largest MSE values across levels of ability ranged from .02 to .08 for equipercentile equating, from .00 to .11 for Rasch equating, and from .00 to .06 for three-parameter model equating. The majority of these differences were less than .05. Differences between MSEs within a case were largest in those instances where the overall equating errors were largest. In general, MSE values were more similar within cases than across different methods and cases, indicating that, from a practical standpoint, all three methods were reasonably invariant to the conditions imposed by this simulation.

To illustrate this point further, Figures 1 and 2 show the actual equating functions for high- and low-ability sample equating. Both figures are based on Case 2 from Table 2, where test discriminations

were equal and where all cs on both tests were .2. The straight-line part of the criterion equating function reflects the use of Lord's extrapolation method for below-chance scores.

In Figure 1, equating functions are shown for equipercentile equating based on high- and low-ability samples. Above a Test B raw score of 12, both equating functions were very close to the criterion. Below

Figure 1
Equipercentile Equating: $a_A = a_B = .8$; $c_A = c_B = .2$

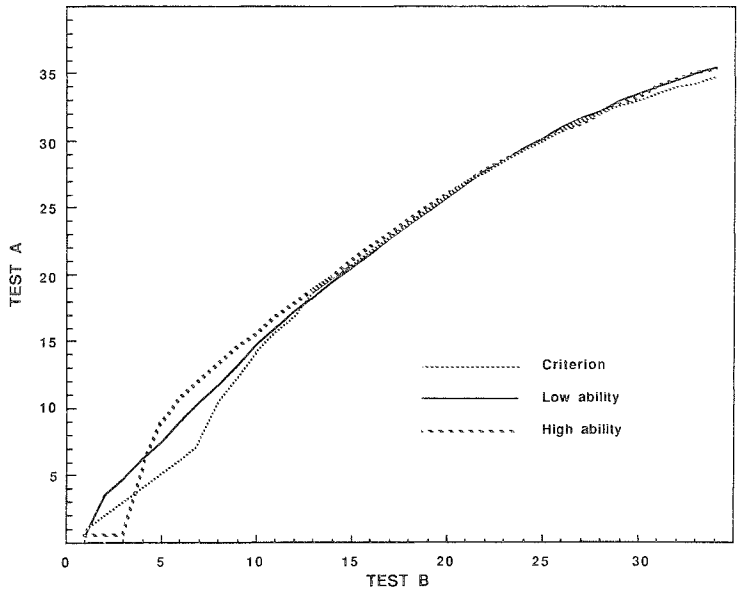
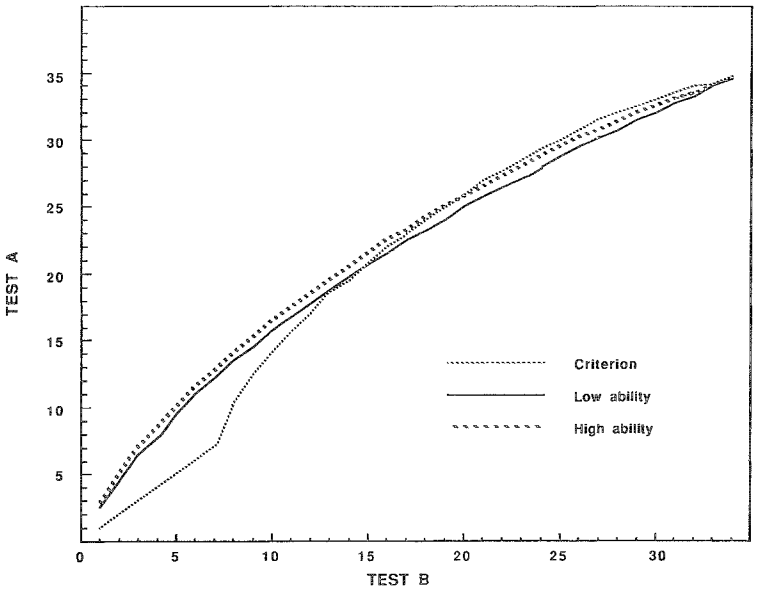


Figure 2
Rasch Model Equating: $a_A = a_B = .8$; $c_A = c_B = .2$



a raw score of 12, both equatings deviated considerably from the criterion and from each other. The UMSEs from Table 2 for high- and low-ability groups were .079 and .041, respectively, a difference of .038. WMSEs were .008 and .014, a difference of only .006, indicating that invariant equatings were produced by the equipercentile method in the region where most of the examinees fell.

Figure 2 shows the high- and low-ability equating functions for the Rasch model for the same case. For both Rasch equatings, the functions were above the criterion equating at the lower end of the raw score scale and below the criterion at the higher end. As was demonstrated for equipercentile equating, they were most discrepant from the criterion at the lower end. The equivalent scores based on high- and low-ability examinees were never more than one raw score unit apart (on the Test A scale), with the equating based on high-ability examinees providing higher Test A equivalents. UMSEs differed by .014 (.150 and .136), and WMSEs differed by .037 (.021 and .058). It is clear that while both equatings were severely biased, they were quite similar to each other. In other words, the equating results were invariant with respect to examinee ability.

These results conflict with those reported by Slinde and Linn (1978, 1979), Loyd and Hoover (1980), and Holmes (1982), who reported differences between equatings from high- and low-ability groups that were considerably larger than those found here. Slinde and Linn and Holmes suggested that the Rasch model's failure to account for guessing was the reason for the lack of invariance that they found. The results of the present study advise against using the Rasch model for vertical equating, but not for the reasons they suggested.

Data for this study were generated from a unidimensional model. The Rasch model was shown not to be robust to violations of its assumptions of equal discrimination and no chance scoring, but at least Rasch equating was invariant with respect to examinee ability. Loyd and Hoover suggested that Rasch vertical equating was suspect, due to multidimensionality across the levels of an achievement test battery. The results reported here indicate that invariance can be demonstrated for unidimensional models.

This conclusion is supported by research described above by Harris and Kolen (1986) and Harris and Hoover (1987). These studies reported invariance with respect to examinee ability for three-parameter model horizontal equating but not for vertical equating. Furthermore, Kolen (1981) and Petersen et al. (1983) both reported that the equating methods they used responded very differently to verbal and mathematical tests for horizontal equating. Taken as a whole, the current and previous research suggests that multidimensionality is probably the primary reason for the lack of invariance shown in vertical equating studies.

That multidimensionality might account for a lack of test equating invariance has profound implications for vertical equating. It may not be meaningful to vertically equate certain kinds of tests. The Harris and Hoover, Harris and Kolen, and Loyd and Hoover studies all used mathematics tests, the content of which is known to vary considerably across grade levels. Reading and vocabulary tests, on the other hand, might be more unidimensional across grades and may provide more invariant equating results. The issue of dimensionality of test content across test levels appears to be a critical one and deserves further exploration.

References

Anderson, D. O. (1985). *Scale stability during test equating*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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.]

Cook, L. L., & Eignor, D. R. (1983). *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., & Taft, H. L. (1984). *A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

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.

Cureton, E. E., & Tukey, J. W. (1951). Smoothing frequency distributions, equating tests, and preparing norms. (Abstract of presented paper.) *American Psychologist, 6*, 404.

Curry, A. R., Bashaw, W. L., & Rentz, R. R. (1978). *Invariance of Rasch model ability parameter estimates over different collections of items*. Paper presented at the annual meeting of the American Educational Research Association, Toronto.

Divgi, D. R. (1981). *Does the Rasch model really work? Not if you look closely*. Paper presented at the annual meeting of the American Educational Research Association, 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.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Harris, D. J., & Hoover, H. D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11*, 151–159.

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10*, 35–43.

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

Holmes, S. E., & Doody-Bogan, E. N. (1983). *An empirical study of vertical equating methods using the three-parameter logistic model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement, 6*, 249–260.

IMSL, Inc. (1980). *International Mathematical and Statistical Libraries reference manual*. Houston TX: Author.

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

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.

Levine, R. S. (1958). *Estimated national norms for the Scholastic Aptitude Test* (Statistical Report No. 1). Princeton NJ: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

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.

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.

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

Skaggs, G., & Stevenson, J. (1986). *A comparison of ASCAL and LOGIST parameter estimation programs*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

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.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116.

Wright, B. D., Mead, R. J., & Bell, S. R. (1980). *BICAL: Calibrating items with the Rasch model* (Research Memorandum No. 23C). Chicago: University of Chi-

cago, Department of Education, Statistical Laboratory.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*, 299–325.

Author's Address

Send requests for reprints or further information to Gary Skaggs, Office of Research and Evaluation, Fairfax County Public Schools, 7423 Camp Alger Avenue, Falls Church VA 22042, U.S.A.