An Exploration of the Robustness of Four Test Equating Models

Gary Skaggs and Robert W. Lissitz
University of Maryland

This monte carlo study explored how four commonly used test equating methods (linear, equipercentile, and item response theory methods based on the Rasch and three-parameter models) responded to tests of different psychometric properties. The four methods were applied to generated data sets where mean item difficulty and discrimination as well as level of chance scoring were manipulated. In all cases, examinee ability was matched to the level of difficulty of the tests. The results showed the Rasch model not to be very robust to violations of the equal discrimination and non-chance scoring assumptions. There were also problems with the three-parameter model, but these were due primarily to estimation and linking problems. The recommended procedure for tests similar to those studied is the equipercentile method.

The application of item response theory (IRT) to many measurement problems has been one of the major psychometric breakthroughs of the past 20 years. IRT methodology is currently being used in many large standardized testing programs, and at the same time a great deal of research is being done to evaluate the robustness of the procedures under a variety of conditions. One of the most important applications of this methodology is in the area of test equating.

The purpose of test equating is to determine the relationship between scores on two tests that purport to measure the same trait. Equating can be horizontal (between tests of equivalent difficulty and content) or vertical (between tests of intentionally different difficulties).

Most recent equating research has divided equating techniques into IRT and conventional methods. IRT methods have been developed by, among others, Lord (1980) and Wright and Stone (1979). In these methods, a mathematical relationship between scores on tests is modeled. This relationship is based on estimates of item parameters from two tests and the placement of these estimates on the same metric.

So-called conventional equating methods have been well documented and described by Angoff (1984). The two most common approaches are linear and equipercentile equating. In linear equating, equated raw scores from two tests are based on equal standard scores. For equipercentile equating, equated raw scores are based on equal percentile ranks from two tests. Both of these procedures depend upon samples of equal ability.

In addition to offering sample-invariant equating results, IRT equating also has the potential of adding a great deal of flexibility to the development of alternate forms of tests. The workload comes mainly in

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 3, September 1986, pp. 303–317

0146-6216/86/030303-15$2.00

the estimation of item parameters and their placement on a common metric. Once this is done, alternate forms can be created by selecting items in some fashion from a pool of calibrated items. The equating itself is simply the process of computing test characteristic curves for each form. Equated raw scores are based on the same degree of underlying ability. This means that IRT equating is based on the properties of the items that comprise a test rather than the distributions of total test scores, which is the case with conventional equating.

A number of equating studies using IRT methods have appeared in recent years; Holland and Rubin (1982) is an excellent resource. A majority of these have dealt with the one-parameter logistic, or Rasch, model. This model is the simplest of the IRT models but also the most demanding in terms of its assumptions. Unfortunately, research on Rasch equating has been inconclusive. Several studies have found the model to be useful and appropriate for item calibration and linking (Forsyth, Saisangjan, & Gilmer, 1981; Guskey, 1981; Rentz & Bashaw, 1977; Tinsley & Dawis, 1975). On the other hand, a number of researchers have noticed problems with the Rasch model for vertical equating (Holmes, 1982; Loyd & Hoover, 1980; Slinde & Linn, 1978, 1979; Whitely & Dawis, 1974).

Some of the inconsistency can be attributed to different equating designs, different sources of data, and different procedures used to analyze the results. Aside from problems in generalizability, however, there are some fundamental concerns about the Rasch model. The most frequently postulated arguments concern the failure of the model to account for chance scoring, unequal discriminations, and multidimensionality. The last concern applies to more complex IRT models as well.

Research using the three-parameter logistic and other models has not been as plentiful as with the Rasch model, but it has addressed the same interpretive difficulties. Most of the studies have examined the three-parameter model in the context of horizontal equating of general ability tests. This work has generally supported the use of the three-parameter model (Kolen, 1981; Marco, Petersen, & Stewart, 1983; Petersen, Cook, & Stocking, 1983). For vertical equating, results have been mixed, with some studies finding the three-parameter model to be more effective than the Rasch model (Kolen, 1981; Marco et al., 1983) for some tests. Moreover, the comparison between the models has been shown to depend largely on the content of the tests being equated (Holmes & Doody-Bogan, 1983; Kolen, 1981; Petersen et al., 1983).

In the midst of this conflicting research, it is very difficult to make decisions about how to use IRT equating or whether to use it at all. The purpose of the present study was to explore (independent of test content) how test equating results can be affected by the parameters of the items that make up the tests being equated.

Four methods of equating were chosen representing popular versions of linear, equipercentile, Rasch model, and three-parameter model techniques. Data for two tests plus an external anchor test were generated from a three-parameter model in which mean test differences in difficulty, discrimination, and lower asymptote were manipulated. For Rasch model and linear equating, this study is an exploration of robustness when the model's assumptions are violated. For the three-parameter model, this study amounts to an examination of the parameter estimation strategy. For equipercentile equating, this study explores its effectiveness under a variety of test conditions.

Method

Data

Data in this study were generated from the three-parameter logistic model:

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

where P_{ij} is the probability of correct response (u_{ij}) to item i by person j. The parameters for item i are discrimination (a_i), difficulty (b_i), and lower asymptote (c_i), while θ is the ability of person j.

All of the data in this study were, therefore, generated from a unidimensional model. This allows a determination of whether differences in test item parameters could affect equating results even when the data were unidimensional. This choice of model was also convenient in that a criterion equating function could be developed easily for each case.

The response to item i by person j, a 0 or 1, was determined by comparing the probability defined by Equation 1 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 by the GGUBS generator (IMSL, 1980).

In all simulation cases, an external anchor test design was used. In each case, two data sets were generated. Each data set consisted of the responses of 2,000 examinees to a 35-item test plus an anchor test of 15 items. Each test equating case was comprised, therefore, of 4,000 examinees and 85 items. This size was chosen to be large enough to provide stable parameter estimates, especially for the three-parameter model (Hulin, Lissak, & Drasgow, 1982).

Item and Ability Parameters

The item parameters used to generate the data were determined by manipulating lower asymptotes and mean test difficulty and discrimination of the tests being equated (Tests A and B).

Item difficulty was studied at three levels: (1) $\bar{b}_A = \bar{b}_B = 0$; (2) $\bar{b}_A = -.5$, $\bar{b}_B = .5$; and (3) $\bar{b}_A = -1.0$, $\bar{b}_B = 1.0$. Within each test, difficulties were uniformly distributed across a range of ± 2 logits. The first level represents a horizontal equating situation while the other two are vertical equatings. The latter two could be thought of as moderate and extreme differences in test difficulty.

Item discrimination was also examined at three levels: (1) $\bar{a}_A = \bar{a}_B = .8$; (2) $\bar{a}_A = .5$, $\bar{a}_B = 1.1$; and (3) $\bar{a}_A = 1.1$, $\bar{a}_B = .5$. Within each test, discriminations were uniformly distributed across a range of $\pm .1$. Although a range of discrimination greater than this might be expected in practice, these values were chosen to increase the differentiation of tests with different mean discriminations. Difficulties and discriminations were randomly paired.

In terms of violation of the equal discrimination assumption of the Rasch model, Divgi (1981) has pointed out that ability estimates are systematically affected by unequal mean discrimination on the two tests to be equated. The mean discriminations selected here represent weak, moderate, and high levels of test discrimination.

Lower asymptote values were manipulated in four ways: (1) $c_A = c_B = 0$; (2) $c_A = c_B = .2$; (3) $c_A = 0$, $c_B = .2$, and (4) $c_A = .2$, $c_B = 0$. In each case, lower asymptote values were the same for all items within a test.

For five-option multiple choice tests, a random response would result in a .2 probability of a correct response. While a value of .2 is unlikely to hold for all items of an actual test, the interest in this study was the effect of zero versus non-zero lower asymptote values. Therefore, all items had c values of either 0 or .2.

All anchor test items had a mean difficulty of zero and a mean discrimination of .8. For vertical equating, the anchor items represented an overlap in difficulty between Tests A and B. Lower asymptote values were all zero except in the case where the values were .2 for both tests. In these cases, lower asymptotes were .2 for the anchor test items. In this study, a complete crossing of all levels produced a $3 \times 3 \times 4$ design consisting of 36 cells, or cases, of pairs of tests to be equated. These item parameters were chosen to reflect a typical test equating between tests of either similar or dissimilar item parameters.

The abilities of the examinees were chosen to match each test's difficulty. Each sample of 2,000 examinees was selected from a normal distribution with a mean equal to the mean difficulty of the test and a standard deviation of 1. The GGNML generator (IMSL, 1980) was used to generate ability parameters

for each sample. Ability was not an independent variable in this study. These distributions were chosen to reflect an ideal match of tests with examinees and data with parameter estimation programs.

Equating Design

One linear, one equipercentile, and two IRT equating methods were chosen for this study on the basis of their popularity in recent IRT equating studies. In all cases, an external anchor test design was used, and Test B was equated to Test A. That is, for each raw score on Test B, an equivalent was found on the raw score scale of Test A. For vertical equating, Test B was always the more difficult test.

Conventional Equating Methods

Linear equating. The basic concept behind linear equating is that equal standard score deviates (Z_a and Z_b) denote equivalent raw scores. The objective is to compute a linear equation that will convert scores from one test to another. For this approach to be valid, it must be assumed that the shapes of the raw score distributions on the two tests to be equated are the same for groups of equal ability.

The linear function used in this study was derived by Levine (1955) and described by Angoff (1984) as Design IVC–1. This method was developed for the data collection design described above and is appropriate for groups that are widely different in ability and for tests that are equally reliable.

Equipercentile equating. The basic principle of frequency estimation equipercentile equating is that, in two groups of equal ability, one taking each test, equal percentile ranks indicate equal ability, whatever the corresponding raw scores may be. Unlike the linear method, this method represents a curvilinear approach when the raw score distributions on the two tests are different in shape.

The equipercentile portion of this study used a method developed by Levine (1958) and described by Angoff (1982, 1984) as Design IVB. The raw score distribution for each test was estimated for a hypothetical combined group of examinees. These distributions were then smoothed using the Cureton-Tukey (1951) algorithm (Angoff, 1982, 1984) and converted to cumulative proportion distributions. Once this was accomplished, a set of equivalent raw scores was developed whereby each pair of scores was matched to the same percentile rank.

IRT Equating Methods

Rasch model estimation. Item and ability parameters were estimated using the BICAL program of Wright, Mead, and Bell (1980). For the Rasch model, the raw score serves as a sufficient statistic for ability estimation, and the item score (frequency of correct response) is a sufficient statistic for estimating item difficulty. The BICAL program therefore produces an ability estimate for each raw score category, except for zero and perfect scores for which there are no estimates of ability. The program also establishes a metric that insures that the average item difficulty equals zero.

For each test equating, two BICAL runs were necessary, one for each test/sample combination. In each run, the anchor test items were estimated along with the main test items. Two sets of estimates were therefore computed for the anchor test items, and these were used to place the item difficulty estimates 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 from the anchor test items as follows:

$$T = \sum_{i}^{m} (B_{ai} - B_{bi})/m \quad , \tag{2}$$

where B_{ai} and B_{bi} are the difficulties of anchor test item i when estimated with Tests A and B, respectively,

and m is equal to the anchor test length, in this case 15 items. By adding T to the item difficulties of Test B, all item parameters were placed on the same scale.

Three-parameter model estimation. For the three-parameter model, all item and ability parameters were estimated using the LOGIST 5 program of Wingersky, Barton, and Lord (1982). Unlike BICAL, there are no sufficient statistics for any of the parameters in the three-parameter model. Each ability estimate from LOGIST corresponds to a pattern of item responses instead of an overall raw score. As before, abilities for zero and perfect scores cannot be estimated, and the program deletes such cases. Another difference between LOGIST and BICAL is the resulting metric. LOGIST scales the ability estimates to a mean of zero and standard deviation of one (instead of centering item difficulties).

Because of its likelihood function, the program can handle omitted items. This means that only one LOGIST run was required for each pair of tests to be equated. This was done for each sample by coding items from the other test as "not reached". Because the anchor test items were answered by both samples, estimation proceeded as if there were a single data matrix to be evaluated. As a result of using a single LOGIST run, item parameters from the two tests were simultaneously placed on the same scale.

Equating Technique

This study employed the estimated true score equating model of Lord (1980, ch. 13). Under this method, estimated true scores for two tests are related to ability in the following manner:

$$\xi = \sum_{i}^{k} P_i(\theta) \tag{3}$$

and

$$\eta = \sum_{j}^{m} P_j(\theta) \quad , \tag{4}$$

where $P_i(\theta)$ and $P_j(\theta)$ are the probabilities for correct responses to items i and j on the two tests. The summations are across items. If the item parameters of the two tests are on the same scale, the ξ and η may be considered equivalent true scores if they correspond to the same ability θ. In practice, estimated item and ability parameters are used, and the $P(\theta)$ can represent any of the latent trait models. Equating using either the Rasch model or the three-parameter model differs only in the manner in which item and ability parameters are estimated and placed on the same scale. In both cases, the above relationship forms the basis for the equating.

Once item parameters are estimated and placed on the same scale, equating is a relatively simple process. For each raw score on Test B, a Test A equivalent was computed. First, Equation 4 was solved for θ using Newton's iterative method. This was then substituted into Equation 3 to obtain the Test A equivalent score. This was done for every raw score category except for zero and perfect scores.

Both of these procedures were performed for both the Rasch and three-parameter models and for the criterion equatings. For the three-parameter model, however, one modification was made. For any θ, it is not possible to find an expected true score less than $\sum c_i$, or below chance level. Yet there were raw scores in that range that needed to be accounted for. In this situation, a linear extrapolation developed by Lord (1980, pp. 210–211) was used.

Analysis Procedures

Since the data were generated from a known three-parameter model, these initial item parameters were used to develop a criterion for the test equating cases. This criterion was simply a pairing of raw

scores corresponding to the same ability estimates, using the relationship in Equations 3 and 4. This equating function was then compared to the equating functions produced by the four equating methods. Two summary statistics were used to interpret the results. These statistics are very similar to mean square error statistics used in other equating studies (e.g., Marco et al., 1983; Petersen et al., 1983). These indices are referred to here as the weighted and unweighted mean square error (MSE) and are defined as follows:

$$\text{Weighted MSE} = \sum_{i}^{k-1} f_i (X_E - X_{crit})^2 / \sum_{i}^{k} f_i S_B^2 \quad (5)$$

$$\text{Unweighted MSE} = \sum_{i=1}^{k-1} (X_E - X_{crit})^2 / S_B^2 \quad (6)$$

where k is the number of items on Test B,
S_B^2 is the raw score variance for Test B,
X_{crit} is the criterion test score equivalent on Test A for raw score i on Test B,
X_E is the Test A equivalent for raw score i that is produced by one of the equating methods, and
f_i is the frequency of raw score i on Test B.

The summation is over raw score values. For the weighted MSE, however, the summation is only across that part of the scale where extrapolation was not necessary, in this case from raw scores of 7 to 34.

Results

Raw score means and standard deviations for all data sets are shown in Table 1. Raw score means ranged from approximately 17.5 to 21.3 and standard deviations ranged from 5.0 to 7.7. By comparing data sets generated under similar item parameters, it is clear that the generation procedure produced very consistent results. An examination of the frequency distributions for each data set (not shown) also revealed a high degree of consistency in the shapes of the distributions. This in turn suggested that there was a high degree of stability in the equatings.

As 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. Non-zero lower asymptotes ($c = .2$) produced negatively skewed raw score distributions and slightly higher means.

The results of the test equating cases are presented in Tables 2, 3, and 4. Table 2 shows MSE values for horizontal equating. Tables 3 and 4 give MSE values for vertical equating under conditions of moderate and extreme differences in difficulty, respectively. The first case in these tables, where test difficulties and discriminations were equal and lower asymptotes were zero, was a situation where the data fit the Rasch model for both tests. From a psychometric point of view, this represented an ideal (easy) equating situation. The best results for all four methods (i.e., almost perfect equating) were found in this case. The worst results for all methods occurred where mean test difficulties and discriminations were unequal and where levels of chance scoring were unequal. Cases where low discrimination was paired with non-zero lower asymptotes proved especially troublesome for all methods.

In general, across all cases, the error indices were lowest for equipercentile and three-parameter model equating. The values for linear and Rasch model equating fluctuated considerably between cases. As expected, equatings for the latter procedures were best in situations where their assumptions were not violated.

One other general result that is apparent from Tables 2, 3, and 4 is that the weighted MSE values are usually considerably smaller than the unweighted MSE values. There seem to be two reasons for this. First, the weighted index excluded the below-chance raw score region where the largest equating errors

Table 1
Raw Score Means and Standard Deviations for Generated Data

Test A					Test B				
$\bar{b}$	$\bar{a}$	c	Mean	S.D.	$\bar{b}$	$\bar{a}$	c	Mean	S.D.
0	.8	.0	17.54	7.23	0	.8	.0	17.58	7.27
0	.8	.2	21.08	5.89	0	.8	.2	21.15	6.10
0	.8	.0	17.79	7.12	0	.8	.2	20.95	6.00
0	.8	.2	21.25	5.98	0	.8	.0	17.67	6.99
0	.5	.0	17.31	6.13	0	1.1	.0	17.16	7.72
0	.5	.2	20.99	4.98	0	1.1	.2	21.04	6.24
0	.5	.0	17.79	6.00	0	1.1	.2	21.08	6.38
0	.5	.2	20.85	5.09	0	1.1	.0	17.58	7.50
0	1.1	.0	17.48	7.55	0	.5	.0	17.66	6.03
0	1.1	.2	20.83	6.36	0	.5	.2	21.14	5.09
0	1.1	.0	17.62	7.62	0	.5	.2	20.99	5.02
0	1.1	.2	21.29	6.36	0	.5	.0	17.63	5.90
-.5	.8	.0	17.54	7.08	.5	.8	.0	17.42	7.05
-.5	.8	.2	20.92	5.94	.5	.8	.2	20.99	5.96
-.5	.8	.0	17.57	7.21	.5	.8	.2	21.01	5.84
-.5	.8	.2	21.04	5.87	.5	.8	.0	17.39	7.17
-.5	.5	.0	17.63	5.93	.5	1.1	.0	17.41	7.60
-.5	.5	.2	21.16	5.07	.5	1.1	.2	20.98	6.37
-.5	.5	.0	17.64	6.04	.5	1.1	.2	21.16	6.27
-.5	.5	.2	20.97	5.10	.5	1.1	.0	17.55	7.70
-.5	1.1	.0	17.51	7.40	.5	.5	.0	17.58	6.07
-.5	1.1	.2	21.03	6.32	.5	.5	.2	21.07	4.98
-.5	1.1	.0	17.49	7.60	.5	.5	.2	20.63	5.15
-.5	1.1	.2	20.88	6.42	.5	.5	.0	17.39	6.09
-1.0	.8	.0	17.41	7.00	1.0	.8	.0	17.37	7.19
-1.0	.8	.2	20.84	5.90	1.0	.8	.2	20.81	6.02
-1.0	.8	.0	17.26	7.06	1.0	.8	.2	21.04	5.95
-1.0	.8	.2	21.11	5.88	1.0	.8	.0	17.63	7.08
-1.0	.5	.0	17.47	6.12	1.0	1.1	.0	17.68	7.49
-1.0	.5	.2	21.00	5.14	1.0	1.1	.2	20.79	6.32
-1.0	.5	.0	17.43	6.05	1.0	1.1	.2	20.85	6.41
-1.0	.5	.2	21.14	5.14	1.0	1.1	.0	17.60	7.41
-1.0	1.1	.0	17.69	7.65	1.0	.5	.0	17.63	5.95
-1.0	1.1	.2	20.98	6.46	1.0	.5	.2	20.97	5.15
-1.0	1.1	.0	17.55	7.46	1.0	.5	.2	20.94	5.01
-1.0	1.1	.2	20.88	6.32	1.0	.5	.0	17.40	6.00

tended to occur. Second, all four methods tended to operate most accurately where there were the greatest concentrations of examinees, thus producing smaller error values for the weighted index.

Discussion

Linear Equating

The assumptions of the linear equating model were violated whenever the shapes of the raw score distributions differed for the two tests being equated. This occurred when the mean test difficulty, mean test discrimination, and/or level of lower asymptotes differed between the two tests. The appropriateness

Table 2
Unweighted Mean Square Error (UMSE) and Weighted Mean Square Error (WMSE) for Horizontal Equating Cases ($\bar{b}_A = \bar{b}_B = 0$)

Mean Item Discrimination and Equating Method	$c_A = c_B = .0$		$c_A = c_B = .2$		$c_A = .0$ $c_B = .0$		$c_A = .2$	
	UMSE	WMSE	UMSE	WMSE	UMSE	WMSE	UMSE	WMSE
$\bar{a}_A = \bar{a}_B = .8$								
Lin	.000	.000	.003	.001	.086	.004	.003	.002
Eq%ile	.000	.000	.004	.002	.010	.004	.016	.001
Rasch	.000	.000	.000	.000	.226	.127	.172	.062
3P	.000	.000	.001	.001	.045	.019	.037	.006
$\bar{a}_A = .5$ $\bar{a}_B = 1.1$								
Lin	.005	.003	.061	.004	.022	.002	.004	.002
Eq%ile	.008	.002	.014	.002	.022	.001	.010	.002
Rasch	.340	.237	.195	.154	.068	.047	.704	.367
3P	.069	.043	.063	.055	.034	.020	.123	.046
$\bar{a}_A = 1.1$ $\bar{a}_B = .5$								
Lin	.030	.008	.146	.012	.604	.041	.031	.002
Eq%ile	.010	.008	.036	.006	.043	.032	.016	.001
Rasch	.589	.346	.274	.203	1.593	.978	.078	.040
3P	.123	.091	.103	.100	.353	.289	.038	.024

of linear equating could be gauged by the degree of curvilinearity in the criterion equating function. The total error for linear equating was minimized in cases of horizontal equating with equally discriminating tests. Chance scoring did not affect the equating in these cases because the criterion equating function was still linear.

In all vertical equating cases, the criterion equating function was curvilinear to some extent. The assumptions of linear equating were therefore violated for all cases of vertical equating. For horizontal equating, where mean test discriminations were unequal, the criterion equating resembled a monotonic cubic function, again violating the assumptions of the linear model. Under conditions of equal test discrimination, the introduction of chance scoring did not increase the degree of nonlinearity in the criterion, and linear equating results were quite accurate. For unequal test discrimination, chance scoring affected the degree of curvilinearity in the criterion and thus produced some fluctuation in the values of the error indices.

Equipercentile Equating

An examination of the values in Tables 2, 3, and 4 shows that equipercentile equating produced results as good as or better than three-parameter model equating in most of the equating cases. In general, equipercentile equating showed the least amount of sensitivity to changes in the independent variables of this study. Unweighted MSE values ranged from 0.0 to .215, and weighted MSEs ranged from 0.0 to .220.

Table 3
Unweighted Mean Square Error (UMSE)
and Weighted Mean Square Error (WMSE)
for Horizontal Equating Cases ($\bar{b}_A$ = -.5 and $\bar{b}_B$ = .5)

Mean Item Discrimination and Equating Method	$c_A = c_B = .0$		$c_A = c_B = .2$		$c_A = .0$, $c_B = .2$		$c_A = .2$, $c_B = .0$	
	UMSE	WMSE	UMSE	WMSE	UMSE	WMSE	UMSE	WMSE
$\bar{a}_A = \bar{a}_B = .8$								
Lin	.122	.035	.284	.040	.257	.071	.107	.040
Eq%ile	.018	.016	.083	.045	.029	.029	.020	.009
Rasch	.000	.000	.119	.017	.658	.147	.085	.043
3P	.005	.001	.001	.001	.034	.011	.040	.004
$\bar{a}_A = .5$, $\bar{a}_B = 1.1$								
Lin	.049	.025	.407	.020	.266	.044	.043	.026
Eq%ile	.021	.012	.063	.020	.066	.023	.020	.006
Rasch	.314	.189	.087	.093	.195	.102	.460	.195
3P	.083	.044	.044	.042	.050	.037	.122	.046
$\bar{a}_A = 1.1$, $\bar{a}_B = .5$								
Lin	.391	.055	.265	.057	.536	.127	.314	.050
Eq%ile	.038	.053	.101	.084	.058	.067	.029	.032
Rasch	.546	.276	.662	.259	2.247	.884	.108	.069
3P	.064	.063	.072	.056	.205	.201	.032	.011

In all horizontal equating cases, equipercentile equating was as good as or better than three-parameter model equating. Equipercentile equating was the best of the four methods in cases where mean test discriminations were unequal. In cases where test discriminations were unequal and either or both tests contained non-zero lower asymptotes, equipercentile equating was the only one of the four methods to produce what would be considered acceptable results.

For vertical equating, as the difference in test difficulty increased, the MSE values associated with equipercentile equating increased as well. This result did not hold for three-parameter model equating. In Tables 3 and 4, equipercentile results were generally not as good as they were for the three-parameter model, although they were better than both linear and Rasch model equating.

Because standard errors are not available for the MSE statistics, it is difficult to assess what would be a statistically significant difference between MSE values. However, a difference of .05 or more (i.e., 5% of the raw score variance) might be considered practically significant.

In this version of equipercentile equating, a total group cumulative distribution was estimated for each test, based on the responses of the combined sample to the anchor test items. That this estimation, in conjunction with a smoothing routine, worked so well was somewhat surprising. One possible explanation is that equipercentile equating is the only one of the four approaches not based on a model; it is simply the best fit of the data at hand. The issue of cross-validation might be raised, but the preliminary work for this study (not reported here) shows that these results were very stable across replications with these item and sample sizes.

Table 4
Unweighted Mean Square Error (UMSE)
and Weighted Mean Square Error (WMSE)
for Equating Cases ($\bar{b}_A$ = -1.0 and $\bar{b}_B$ = 1.0)

Mean Item Discrimination and Equating Method	$c_A = c_B = .0$		$c_A = c_B = .2$		$c_A = .0$, $c_B = .2$		$c_A = .2$, $c_B = .0$	
	UMSE	WMSE	UMSE	WMSE	UMSE	WMSE	UMSE	WMSE
$\bar{a}_A = \bar{a}_B = .8$								
Lin	.848	.449	1.345	.399	1.749	.799	.753	.436
Eq%ile	.108	.067	.166	.142	.158	.117	.072	.030
Rasch	.000	.000	.477	.036	1.317	.146	.040	.034
3P	.051	.000	.010	.002	.027	.018	.054	.004
$\bar{a}_A = .5$, $\bar{a}_B = 1.1$								
Lin	.440	.300	1.084	.199	1.274	.397	.392	.270
Eq%ile	.088	.031	.215	.072	.124	.068	.079	.018
Rasch	.218	.134	.227	.066	.752	.215	.284	.141
3P	.064	.053	.040	.043	.109	.051	.078	.044
$\bar{a}_A = 1.1$, $\bar{a}_B = .5$								
Lin	2.279	.798	1.837	.507	2.964	1.240	1.669	.612
Eq%ile	.086	.110	.152	.194	.179	.220	.116	.096
Rasch	.486	.157	1.362	.208	3.406	.575	.204	.060
3P	.171	.036	.061	.032	.201	.110	.213	.007

The anchor test in this study comprised items whose difficulties overlapped those of the two tests to be equated. This meant that, for vertical equating, the anchor items were easy for one group of examinees and difficult for the other. When test difficulties differed widely, distributions on the anchor test were severely skewed, resulting in less accurate estimation of a combined group cumulative distribution. Better results for equipercentile equating in these situations might have been observed if the anchor items spanned a wider range of difficulty. This result has already been observed with Rasch model equating (Loyd, 1983).

Rasch Model Equating

An examination of the results in Tables 2, 3, and 4 suggests that the Rasch model was not very robust to the violations of its assumptions that were introduced in this study. The first case in Table 2 shows a situation where the data fit the Rasch model across both tests. The Rasch model, as expected, performed extremely well, as did the other three methods. In the second case in Table 2, all items had a lower asymptote of .2, yet the equating was still quite good for all methods, including the Rasch model.

In subsequent cases in Table 2, where the level of chance scoring was unequal in the two tests and where mean test discriminations were unequal, the Rasch model performed very poorly. An explanation for these results can be found in the estimation and linking procedures. When the BICAL program is faced with estimating parameters from a data set, a metric is chosen so that mean difficulty equals 0 and all

discriminations equal 1. When BICAL runs are done for two tests with different properties, the resulting metrics are different, and estimated item difficulties for one test are compressed to a greater or lesser extent than they should be. The use of an equating constant does not alter the underlying metric, and a bias is introduced into the equating.

That this can result in severe equating bias can be seen in cases where low discrimination and chance scoring were both present in one test while the other test had a higher level of discrimination and no chance scoring. For example, in Table 2, such a case produced the largest MSE values, 1.593 and .978. The BICAL estimates for this case revealed a range of difficulty of -4.0 to 3.1 for Test A, but for Test B the range was -1.5 to 1.2. Both tests were generated with a range of ± 2 logits. Obviously the metrics were quite different, and a substantial bias resulted.

For vertical equating, where the data for each test fit the Rasch model, Rasch equating again produced excellent results. In fact, these results were slightly better than those produced by three-parameter model equating. This suggests that, if the data are known to fit the Rasch model, BICAL-based equating is preferable to the more general approach of using LOGIST.

The error introduced by unequal mean discrimination and chance scoring was even more pronounced for vertical equating. Even where the same degree of chance scoring occurred on the two tests, Rasch equating was clearly inadequate. These results therefore corroborate, from a different methodological perspective, empirical results that suggest that the Rasch model should not be used to equate vertically whenever chance scoring is a possibility. These results also suggest that the Rasch model not be used in any situation where mean test discriminations are unequal.

These problems would be especially difficult to overcome when alternate forms are being constructed from an item bank. For example, just to ensure that all tests formed from the bank had the same mean discrimination, a complicated algorithm for item selection would have to be developed if all discriminations were not equal. Similarly, chance scoring would not be a problem only if all items had the same degree of chance scoring *and* the forms to be equated were of comparable difficulty.

Three-Parameter Model Equating

Because the data were generated from a three-parameter model, it would have been expected that three-parameter model estimation and equating would be quite accurate. An examination of the values in Tables 2, 3, and 4 indicates that this was not always the case.

Paradoxically, the values of the indices decreased, overall, as the differences in test difficulty increased. As this difference increased, the functions became more sharply curvilinear. The variations introduced by manipulating discrimination and lower asymptotes were more pronounced for horizontal equating than for vertical equating. The situation is analogous to restriction of range problems in correlational research. The shrinkage of error indices from Table 2 to Table 4 is an artifact of the equating situation. This will be further discussed below. In terms of the three-parameter model, these results indicate that equating was relatively unaffected by differences in test difficulty.

On the other hand, the equating was adversely affected by unequal discrimination. While different levels of chance scoring did not by themselves seem to affect equating results, the largest equating errors were observed in cases where low discrimination was paired with non-zero lower asymptotes.

Because the data actually fit the model, LOGIST's estimation algorithms are responsible for the success of the equating. In this study, simultaneous estimation was used. A single LOGIST run was used for each test equating by employing the "not reached" option. In every case for this study, the LOGIST estimation procedure converged. However, when test discriminations and/or degrees of chance scoring were unequal between the two tests, the program typically took at least 35 stages to converge. (For practical reasons, it was decided to extend the limit on the number of stages rather than produce continuation runs.)

That LOGIST was unable to recover the initial metric can be illustrated by examining the parameter estimates from one of the cases. For the situation where the initial parameters for Tests A and B were $\bar{b}_A = -.5, \bar{a}_A = 1.1, c_A = 0.0$ and $\bar{b}_B = .5, \bar{a}_B = .5, c_B = .2$ (Table 3), the unweighted and weighted MSEs were .205 and .201, respectively. For Test A, the LOGIST difficulty estimates ranged from -3.14 to 1.72, while the original difficulties ranged from -2.5 to 1.5. However, by linearly transforming the LOGIST estimates to the original metric (i.e., equalizing the means and standard deviations of the difficulty estimates), the estimates ranged from -4.03 to 2.52. The LOGIST discriminations for Test A ranged from .8 to 1.3, compared to the original 1.0 to 1.2. After transformation, the range became .6 to .9. For Test B, the LOGIST difficulties after transformation ranged from -1.37 to 2.55; the original span was from -1.5 to 2.5. However, the difficulties were poorly estimated for the easiest half of the test. The LOGIST discriminations after transformation ranged from .5 to 1.0, the original range having been .4 to .6.

Ironically, the lower asymptotes were estimated reasonably well. The default options were used, and default values were obtained for the easiest items on both tests. Yet no item had a $\hat{c}$ value greater than .1 on Test A, and only six items on Test B had $\hat{c}$ values less than .1. It seemed, in general, that the default c values did not impair the equating results.

Another peculiarity was observed in the LOGIST results across all cases. On each test, parameters for a few items (1 to 3 out of 35) were estimated extremely poorly. These tended to occur more frequently on tests with weaker discriminations, and on items with lower difficulty. No apparent reason for these outliers could be found, as all item responses were generated from the same function. Still, an erroneous decision on the quality of an item could be made from these results.

Clearly, in cases of unequal discrimination, LOGIST was unable to reproduce the original metric, and equating was therefore biased. The differences in discrimination were quite severe in this study, and it is not known how well LOGIST would respond to milder differences. On the other hand, the parameters for this study—sample size, test length, and ability distribution—were chosen so as to yield stable, reproducible estimates. The results therefore suggest that the use of the simultaneous estimation procedure of LOGIST is questionable in circumstances such as these. Some other method for transforming estimates to the same scale should be considered. One possibility would be to conduct separate calibration runs for each test/anchor combination and follow this with a transformation procedure (e.g., Stocking & Lord, 1983).

Analysis Procedures

A review of published equating studies reveals a wide variety of evaluation procedures and summary statistics. The degree to which methodology affected the conclusions of these studies is not known. In this study, the weighted MSE statistic was chosen because it had appeared frequently in the literature (e.g., Marco et al., 1983; Petersen et al., 1983). When the results from these statistics were compared to graphs of the equating functions, the weighted MSE values did not seem to represent some of the cases accurately. This was because there were relatively few persons in the raw score ranges where the greatest equating errors occurred, at the lower end of the raw score scale.

Because of this, the unweighted MSE was also computed. As discussed above, the values tended to be higher than for the unweighted statistic. In many of the cases in this study, the value for the weighted statistic appeared quite acceptable while the value for the unweighted statistic was relatively large. This indicates that the equating procedure minimized error in the ranges where most of the examinees scored, but that errors did result at the extremes of the raw score scale. This would be unacceptable if the equating function were to be applied to a new sample of examinees with a different level of ability than the one on which the equating is based. In other words, for equating to be accurate across the entire raw score scale, the abilities of the equating samples must be fairly well spread out across the range of difficulty

of the test. In cases where the values of both indices are quite low, the equating was reasonably accurate across the entire score range.

Other differences appear when examining the values in Tables 2, 3, and 4. For example, the symmetry in the study's design does not appear in the MSE values for levels of the discrimination and lower asymptote independent variables. Tests A and B alternate between mean discriminations of .5 and 1.1 and lower asymptotes of 0 and .2. Yet the higher MSE values occur when Test A has the higher discrimination. The same thing occurs when there is more chance scoring on Test B than on Test A. Equating cases that appear to be mirror images of one another result in different MSE values.

The paradox can be explained by the fact that the MSE statistic uses vertical distances between the equating functions. If horizontal distances were used (i.e., Test A equated to Test B), the pattern of results would be reversed. This analysis highlights the limitations in the use of MSE statistics for this purpose. A great deal of theoretical statistical work is needed in the area of proper error indexing.

Conclusions

The purpose of this study was to examine how four commonly used test equating procedures would respond to situations in which the properties of the two tests being equated were different. The results indicate that equipercentile equating provides stable results that were as good as or better than any of the other three methods across all horizontal equating cases. This occurred even though the data were generated from a three-parameter IRT model and should therefore have favored that method of equating. Linear and Rasch model equating were very sensitive to violations of their assumptions. Rasch model equating showed robustness only for horizontal equating where the degree of chance scoring was the same for both tests.

For vertical equating, the three-parameter model produced overall the best results of the four methods. Linear equating was clearly inadequate due to the curvilinear nature of the criterion. Rasch model equating produced adequate results only when the data for both tests strictly fit the model, that is, when test discriminations were equal and there was no chance scoring present in the item responses. Equipercentile results were generally good when the tests differed by only one logit in difficulty. When they differed by two logits, the procedure tended to produce considerable equating errors. However, better results might have been achieved with a wider range of difficulty on the anchor test items.

When data fit the Rasch model, equating results were slightly better than three-parameter model equating. In all other cases, three-parameter equating was far better than the Rasch model. Simultaneous estimation using LOGIST seemed unable to recover the original metric when mean test discriminations were unequal.

All of the equating methods were affected similarly in some of the cases. Where the tests being equated differed in mean difficulty and mean discrimination as well as in their degree of chance scoring, the equating error was the largest for all four methods. This suggests that equating tests should not be attempted under such extreme conditions. None of the equating methods could completely overcome the effect of such divergence in item type.

The use of the MSE statistics produced several paradoxes in the results. These could be resolved by examining the equating functions themselves. Certainly, more statistical work needs to be done in the comparison of test characteristic curves.

Finally, all of the data for this study were generated from a unidimensional three-parameter model. Real data do not conform exactly to this model, although its use seems reasonable in a wide variety of situations. How these methods would respond to multidimensional data is not known, but problems for both IRT methods were uncovered in the unidimensional case, and even greater problems would certainly be expected for multidimensional data sets. Because this study did not explore this issue and was restricted

to these four commonly used equating procedures, the generalizability of these results is necessarily limited. It is also limited by the equating procedures themselves. All four methods have a considerable number of variations in their procedures; the use of these variations might have enhanced the results.

This study supports other research findings that have found the Rasch model inappropriate for use in vertical equating situations unless the data are known to fit the model. The three-parameter model procedure used here also did not generally produce acceptable results in more complex situations where it might have been expected to do so. The best advice at this point would seem to be to pay very close attention to the properties of the test items.

In this study, the best overall results were obtained for frequency estimation equipercentile equating and three-parameter model equating. If equating must be done between tests with large differences in their psychometric properties (or item characteristic curves), then equipercentile equating is suggested.

References

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

Angoff, W. H. (1984). *Scales, norms and equivalent scores*. Princeton NJ: Educational Testing Service. (Reprint of chapter in R. L. Thorndike (Ed.), *Educational measurement* [2nd ed.]. Washington DC: American Council on Education, 1971.)

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

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 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.

Guskey, T. R. (1981). Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. *Applied Psychological Measurement*, *5*, 187–201.

Holland, P. W., & Rubin, D. B. (Eds.) (1982). *Test equating*. New York: Academic Press.

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.

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

Levine, R. S. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (Research Bulletin No. 23). Princeton NJ: Educational Testing Service.

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: Lawrence Erlbaum.

Loyd, B. H. (1983). *A comparison of the stability of selected vertical equating methods*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

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*. 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.

Rentz, R. R., & Bashaw, W. L. (1971). The National Reference Scale for reading: An application of the Rasch model. *Journal of Educational Measurement*, *14*, 161–179.

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.

Stocking, M. L., & Lord, F. M. (1983). Bayesian estimation in the Rasch model. *Applied Psychological Measurement*, *7*, 201–210.

Tinsley, H. E., & Dawis, R. V. (1975). An investigation of the Rasch simple logistic model: Sample-free item and test calibration. *Educational and Psychological Measurement*, *35*, 325–339.

Whitely, S. E., & Dawis, R. V. (1974). The nature of objectivity with the Rasch model. *Journal of Educational Measurement*, *11*, 163–178.

Wingersky, M. S., Barton, M. H., & 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., Mead, R. J., & Bell, S. R. (1980). *BICAL: Calibrating items with the Rasch model* (Research Memorandum No. 23c). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

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

Author's Address

Send requests for reprints or further information to Robert W. Lissitz, EDMS/Education, Benjamin Building, University of Maryland, College Park MD 20742, U.S.A.