Quantifying Equating Errors with Item Response Theory Methods

S. E. Phillips
Michigan State University

The purpose of this paper was to examine alternative techniques for quantifying the errors associated with the criterion of equating a test to itself. Data for the study came from the national standardization of the 3-R's Achievement Test. The reading and mathematics subtests were analyzed using random samples from the Grade 4 norming group. Errors for two item response theory (IRT; three-parameter and Rasch) methods and the equipercentile equating method were investigated. A total of 45 error estimates from the sampling distribution were obtained for each combination of equating method and content area. Analysis of variance procedures were also used to estimate the average error across methods for each content area. In addition, the results of the Phillips (1983a, 1983b) studies were reevaluated using the mean of the sampling distribution of equating errors for each of the methods from the present study and from the corresponding ANOVA error estimates. The results of this study suggest that single-replication error estimates may provide misleading assessments of the errors associated with equating a test to itself. The analysis of variance mean squares appeared somewhat promising as alternatives to error estimates by replication. Finally, the results of this study together with those of the Phillips (1983a) study suggest that the Rasch model may be more reliable than other IRT models for equating, but in some applications it is less valid.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 1, March 1985, pp. 59–71

0146-6216/85/010059-13$1.90

A major concern in comparative equating studies is the choice of a criterion measure for judging the results. In the same measurement sense in which an examinee's true score on an examination can never be known, Cook and Eignor (1983) indicated that the true equating for a given situation is impossible to determine. Thus, there is no standard by which to judge the "better" equating method, and it is difficult to determine how large an equating difference between methods is significant. Several different methods have been used in the research literature to evaluate the results of actual equatings. These methods involve (1) comparison to well-established traditional equating procedures, (2) stability of equatings across equal-ability or cross-validation groups, (3) assessment of scale drift, and (4) equating a test to itself.

The purpose of this paper was to examine alternative techniques for quantifying the errors associated with the criterion of equating a test to itself. To put this method in perspective, a brief survey of the other three methods and their applications is presented first. This background material is followed by a summary of the literature and the difficulties associated with the method of equating a test to itself.

Equating Criteria

Cook and Eignor (1983) suggested that when a traditional equating procedure has been used successfully with the same test over a period of time, comparison to that procedure may provide an appropriate criterion against which to evaluate other equating methods. Several research studies have utilized this approach in comparing traditional and item response theory (IRT) equating methods. Lord (1975, 1977) found that the estimated observed-score, estimated true-score, and traditional equipercentile equating methods produced slightly different results. Rentz and Bashaw (1977) compared equipercentile and Rasch equating methods using the Anchor Test Study data. The Rasch and linear equating methods have been studied by Beard and Pettie (1979) and Bell (1979). Studies by Marco (1977), Guskey (1981), Woods and Wiley (1978) and Golub-Smith (1980) have also utilized this approach to compare IRT and traditional equating methods. In a more recent study, Phillips (1983a) compared IRT and equipercentile equating methods when the scaling test approach was applied to a multilevel achievement battery.

Rather than attempting to assess the accuracy of an equating method relative to a standard based on a more traditional procedure, some researchers have focused on the stability of the equating results obtained with a given method. Kolen (1981) used random samples of examinees to cross-validate his comparisons of several IRT equating methods with traditional equipercentile and linear equating methods using overlapping forms of the Iowa Tests of Educational Development (ITED). Kolen's criterion was the mean squared difference between scores with equivalent percentile ranks in the cross-validation and equated distributions. Loyd and Hoover (1980) studied vertical equating of mathematics subtests of the Iowa Tests of Basic Skills (ITBS) with the Rasch model across unequal ability groups. They estimated sampling differences in the equatings using equating results from the same test forms with groups of similar ability. Other studies that have employed the cross-validation technique include Holmes (1982) and Kolen and Whitney (1982).

Scale drift can be used to evaluate equating results by examining differences in the results obtained with direct and indirect equating chains. This equating criterion has been utilized with IRT and traditional equating methods comparisons with achievement (Cook & Eignor, 1983) and aptitude (Petersen, Cook, & Stocking, 1983) test data. Using this criterion, the equating method judged most suitable is the one producing the least differences in comparative results between equating Text X to Test Z directly, and equating the two tests through intermediate Tests $Y_1, \ldots, Y_n$.

The Test Equated to Itself as a Criterion

The remaining equating criterion that has received attention in recent research literature is the procedure of equating a test to itself using random samples of examinees. Application of this procedure provides an estimate of the errors associated with the replication of methods. Except for errors of measurement and sampling errors in the selection of examinees, the pairs of random samples should produce identical equating results. The degree to which any score in one sample is not equated to the identical score in the other sample provides one estimate of the errors associated with the equating procedure. Studies utilizing the equating of a test to itself as a criterion have included Lord (1975, 1977), Marco, Petersen, and Stewart (1979) and Phillips (1983a). Lord (1982) provided a formula for the asymptotic standard error of an IRT true-score equating using an anchor test, but the formula is not applicable in situations like the Phillips (1983a) study when the two tests to be equated do not overlap and an anchor test has not been administered.

A serious limitation in previous studies that used the test equated to itself as a criterion was that only one replication from the sampling distribution of equating errors was obtained. With only one data point for estimating the errors associated with a given equating method, it is possible that any particular error estimate may have been unusually large or small (Blommers & Forsyth, 1977). A more accurate procedure would have been to estimate a measure of central tendency for the sampling dis-

tribution of errors obtained when a test is equated to itself.

In practice, multiple replications for estimating equating error will be costly and time-consuming. Obtaining an overall average estimate of equating error from the equating tables alone for the equating methods of interest in a particular application would be more cost-effective and might provide a reasonable alternative to single or multiple replications of equating a test to itself for each equating method.

Clearly, the error associated with the sampling distribution of replications of equating a test to itself derives from a different equating design than an estimate based on the application of several different equating methods to the same data. The former error statistic is a measure of differences between results for a single equating method across random samples of examinees. In contrast, the latter error statistic is a measure of differences among equating methods in a single sample of examinees. Thus, there was no reason, a priori, to believe that the two techniques would yield similar results. However, the following heuristic argument provides a rationale for examining the two error statistics together in the same study.

Ideally, if the method of equating a test to itself is to be used as an error criterion, repeated replications are desirable. The mean of the sampling distribution of the error statistics over the repeated replications can be estimated for each equating method of interest. By comparing the differences between results for any pair of equating methods with the average replication error statistics that have been estimated for each method separately, an indication of the "significance" of the observed differences between two methods can be obtained. However, this process represents a significant investment of research time and computer analysis. Thus, a technique that does not require replications but estimates the "average" of the values of the error statistics from repeated replications for both equating methods being compared would be cost-effective, attractive to test users, and would provide a single standard of comparison. This is analogous to the procedure of estimating a single error term by pooling the errors from different treatment groups. If such a procedure were effective in the present situation, then it would provide a technique for quantifying equating error that could be estimated from a single small data sample, such as that which is usually obtained when a publisher administers a separate scaling test for developing a common score scale across grades.

Purpose

The purpose of this study was to provide empirical data for estimating the characteristics of the sampling distribution of chance equating errors with IRT and equipercentile methods. For a single grade, a reading and a mathematics achievement subtest each was repeatedly equated to itself using different random samples of examinees from the national norming group. These replications provided an empirical estimate of the sampling distribution of such errors in the given situation and also provided some evidence of the accuracy of a single replication in estimating equating error.

The study also explored the possibility of estimating the average error across methods using an analysis of variance (ANOVA) term based on the data from an equating table of all the equating methods of interest. Although such a procedure is likely not to conform to all of the assumptions of the ANOVA model, if its results were reasonably close to the averages obtained over multiple replications, it would have the advantage of requiring no additional data beyond that of the original equating study.

Method

The three major objectives of the study were as follows:

1. To estimate the sampling distribution of IRT and equipercentile equating method chance errors using the procedure of equating a test to itself;
2. To explore the use of ANOVA techniques to obtain an overall average estimate of equating error across methods; and

3. To obtain empirical evidence related to the accuracy of a single replication (equating a test to itself) in estimating error.

Data for the study came from the national standardization of the 3-R's Achievement Test (Cole, Trent, & Wadel, 1982). The reading and mathematics subtests were analyzed using random samples from the Grade 4 norming group.

Description of the Test

The 3-R's Achievement Test (Cole et al., 1982) is a basic skills achievement battery designed for use in Grades Kindergarten through 12. There are 11 different levels of the test: Levels 6 through 14 correspond to Grades Kindergarten through 8, Level 15/16 to Grades 9 and 10, and Level 17/18 to Grades 11 and 12.

The test consists of three subtests that assess proficiency in the basic skills of reading, language, and mathematics. The three subtests consist of 50, 40, and 35 multiple-choice items, respectively, in Grades 3–12. These grades are allowed 40, 30, and 28 minutes for completing the respective subtests with responses recorded on separate machine-scorable answer sheets. Further description of the specific skills tested is provided in the test manual (Cole et al., 1982).

The national standardization of the 3-R's Test was conducted in the Spring and Fall of 1980. Approximately 8,000 students per grade in Grades Kindergarten through 12 were tested in the main standardization during April. Participating school districts were randomly sampled from listings compiled from census data based on geographic location and district size. Both public and parochial school districts were included in the sample. Within each selected school district, buildings were randomly sampled. In general, all students attending a sampled school building were tested in the standardization. A more detailed description of the national sample can be found in the test manual (Cole et al., 1982).

Equating Design

Ten random samples of 1,000 students each were drawn from the fourth grade standardization group. Item parameters for the reading and mathematics subtests were estimated separately for each sample using the LOGIST4 (Wood, Wingersky, & Lord, 1976) and BICAL (Wright, Mead, & Bell, 1979) computer programs. For each random sample and content area, parameter estimates were obtained for two IRT models, the three-parameter and the Rasch models. To check the unidimensionality assumption required for these models, linear factor analyses were run on the first three random samples in each content area. Model fit was examined for each of the 10 random samples by calculating the Yen (1981) $Q1$ statistic (three-parameter model) and the BICAL weighted total fit statistic (Rasch model) for each item in each content area.

For each of the two IRT equating methods, Lord's (1980) estimated true-score equating procedure was used to equate each subtest to itself for every possible pair of random samples. (See Phillips, 1983a, for a more detailed description of this equating procedure.)

The average absolute raw score and grade development score differences across the full score scale were computed for each equating. A total of 45 error estimates was obtained for each combination of equating method and content area. These 45 points were used to estimate the empirical sampling distribution of chance equating errors for each content area and equating method. Average absolute differences in equating results for every pair of random samples in each content area (45 total) were also computed using the traditional equipercentile method. The mean, standard deviation, largest and smallest average absolute differences of raw and grade development scores were calculated for each sampling distribution. Average differences were also computed to check for consistent bias in any equating method. Since all average differences were close to zero, no consistent method bias appeared evident and average absolute differences were reported to describe the magnitude of equating error.

ANOVA Error Estimates

Analysis of variance procedures were used to estimate the average error across methods for each content area. A grade development score equating

table was prepared for each content area (reading and mathematics) with the columns representing equating methods (three-parameter, Rasch, and equipercentile) and the rows representing fractional grade development scores equated (by method) to each possible integer raw score. Using these data, a pooled error estimate was obtained across equating methods by computing the interaction mean square of a treatments by replication ($A \times R$) analysis of variance design (Lindquist, 1956), with the equating methods as treatments and the score levels as replications. This analysis was *exploratory* in nature since there was no reason, strictly speaking, to believe that the data met the assumptions of the ANOVA model.

In particular, the proposed ANOVA model assumes that the equating errors between pairs of equating methods are independent. This assumption ignores the correlation (over replications) between sampling errors for a pair of equating methods. However, because the assumption of independence results in an overestimate of the actual error, and because equating differences between pairs of equating methods are required to exceed this overestimate of error in order to be considered "significantly" different, the test is conservative.

In addition to the problems of using a different equating design and violating the assumption of independence, the use of the error term from the ANOVA of the equating table also has the problem of using a different metric. The average absolute error statistic is a measure of absolute error on the grade development scale, whereas the mean squared error statistic is a measure of squared error. To place the ANOVA error estimate on the metric of the grade development scale, the square root of the mean squared error could have been calculated. However, preliminary analyses suggested that the values obtained by this adjustment would have substantially overestimated the average mean replication errors for the three equating methods being studied.

Vertical Equating Comparisons

In the Phillips studies (1983a, 1983b) the procedure of equating a test to itself was used as a criterion for estimating equating error for a vertical equating application. In these comparative equating studies, regular battery on-grade-level tests were equated to a multilevel scaling test administered to the same sample of students in the given grade. For each grade and content area, a table of grade development scores equated to each possible raw test score and average absolute differences between pairs of methods were available for the equipercentile, Rasch, and three-parameter models.

In these studies Phillips (1983a, 1983b) noted that use of the procedure of equating the test to itself may have underestimated the errors associated with each method, because (1) it was based on equating tests of equal difficulty rather than unequal difficulty, (2) the equating was not done through a scaling test, and (3) the use of random samples did not provide the strictly equal ability groups required for the true-score equating procedure and for comparison with the experimental groups in which the same persons took both tests. Despite these limitations, it was of interest in the present study to determine whether the same conclusions would have been reached had the criterion equatings been based on different pairs of random samples or on more than one replication of equating the test to itself. To examine this issue, the results of the Phillips studies were reevaluated using the mean of the sampling distribution of equating errors for each of the methods from the present study and from the corresponding ANOVA error estimates. In addition, the single replication error estimates from the original study were compared to the appropriate sampling distributions from the present study to identify possible outliers in these error estimates.

Results

The results of the linear factor analyses for the first three random samples are presented in Table 1 by content area. In each case the eigenvalues and corresponding percentages of variance are reported for the first five factors. Results across samples within each content area were very similar and showed large first factors accounting for about 57% of the variance in mathematics and 64% to 68% of the variance in reading. First factor eigenvalues

Table 1
Factor Analyses For Three Random Samples By Content Area

Test and Factor	Sample 1 Eigenvalue	Sample 1 Percent of Variance	Sample 2 Eigenvalue	Sample 2 Percent of Variance	Sample 3 Eigenvalue	Sample 3 Percent of Variance
Reading						
1	18.15	68.2	17.44	68.8	16.32	64.2
2	1.50	5.6	1.72	6.8	1.90	7.5
3	1.25	4.7	1.47	5.8	1.52	6.0
4	1.04	3.9	0.94	3.7	1.04	4.1
5	0.86	3.2	0.75	3.0	0.87	3.4
Mathematics						
1	9.06	57.8	9.31	57.0	9.18	57.4
2	1.67	10.6	1.65	10.1	1.58	9.8
3	0.94	6.0	1.06	6.5	1.10	6.9
4	0.74	4.7	0.76	4.7	0.77	4.8
5	0.64	4.1	0.74	4.6	0.69	4.3

were large relative to approximately equal subsequent factor eigenvalues. Using the scree technique (as described by Green, 1983), the data appeared to be sufficiently unidimensional for use of the BICAL and LOGIST programs. It also met the Reckase (1979) criterion that 20% or more of the variance should be accounted for by the first factor to ensure stable parameter estimates.[1]

Fit statistics for both the Rasch and the three-parameter IRT models are reported in Table 2 by content area. Two types of statistics are given: (1) the $Q1$ or total fit statistic averaged over the 10 random samples for each item, and (2) the percentage of times out of the 10 replications for each item that the fit statistic ($Q1$ or total) was significant. Due to the large number of statistics being examined (50 or 35 items by 10 random samples each), significance was defined conservatively as $p < .01$ for the Yen $Q1$ statistic and $t > 2.00$ for the BICAL total fit statistic. None of the average $Q1$ fit statistics were significant in reading or math; two reading items and three math items had significant $Q1$ statistics in three or four samples; 17 reading and 11 math items had significant $Q1$ fit statistics in one or two of the random samples; and 31 reading and 21 math items had no significant $Q1$ fit statistics in any of the random samples. Since the majority of items showed no significant misfit and almost all misfit occurred in three or fewer samples for each item, it appeared that the three-parameter model fit the data reasonably well.

Misfit for the Rasch model appeared much more consistent and seemed to be affected less by random fluctuation. Eight reading and six math items had significant t values in half or more of the random samples; an additional two reading and three math items showed misfit in less than half the samples. A majority of items in both content areas (39 reading and 26 math) showed no misfit in any of the 10 random samples. The average t fit statistic was significant for eight reading and six math items. Overall, the Rasch model on average fit more than 80% of the items and was judged to fit the data reasonably well enough to proceed with the analysis.[2]

[1] The Reckase conclusion was based on factor analyses of phi rather than tetrachoric correlations. A larger criterion may be more appropriate when factor analyzing tetrachoric correlations.

[2] Although the consistently misfitting items could have been removed and the Rasch data reanalyzed, such a procedure was judged inappropriate in this application in which equating error was to be estimated for a standardized test of fixed length and content.

Table 2
Fit Statistics For the Rasch and Three-parameter Item
Response Theory Models by Content Area Averaged Over Ten Random Samples of 1000 Cases

	Reading				Mathematics			
Item Number	$\bar{Q}_i$	Percent of Times Significant[a]	Av. BICAL Total Fit Statistic	Percent of Times Significant[aa]	$\bar{Q}_i$	Percent of Times Significant[a]	Av. BICAL Total Fit Statistic	Percent of Times Significant[aa]
1	9.42	20	-0.58	0	10.45	30	-0.01	0
2	7.37	0	-1.31	0	9.15	0	-0.16	0
3	6.32	0	-1.27	0	8.25	0	-0.60	0
4	12.76	0	-1.34	0	9.34	10	-0.93	0
5	11.18	10	-1.88	0	8.53	0	-0.87	0
6	6.49	0	-1.74	0	8.64	10	-1.89	0
7	8.04	0	2.66[aa]	80	9.87	10	-1.53	0
8	6.59	0	-0.80	0	7.27	0	-1.53	0
9	6.74	0	0.66	0	9.91	0	-0.95	0
10	8.61	10	-1.61	0	10.03	0	-0.13	0
11	7.01	0	-1.17	0	8.51	0	-1.08	0
12	8.55	0	-3.71	0	8.35	0	-1.66	0
13	8.89	10	-2.08	0	10.09	20	-1.94	40
14	9.74	10	-2.71	0	11.92	20	-2.99	0
15	8.00	0	-2.46	0	12.24	30	-3.21	0
16	11.72	10	0.94	10	10.09	10	-3.78	0
17	9.32	0	-0.56	0	12.71	10	-1.65	0
18	4.57	0	-3.67	0	11.44	0	-1.74	0
19	11.05	10	-0.43	0	9.49	0	-0.95	0
20	8.72	0	2.44[aa]	70	8.19	0	-1.15	0
21	7.69	0	-1.84	0	11.84	10	-3.29	0
22	5.88	0	-2.89	0	11.41	0	-2.51	0
23	7.88	0	3.07[aa]	80	9.38	0	3.05[aa]	90
24	10.85	10	-1.61	0	5.41	0	0.22	10
25	13.51	20	-0.28	0	8.00	0	-0.78	0
26	10.68	0	-2.26	0	15.35	40	2.16[aa]	50
27	12.60	0	-0.08	0	9.58	0	4.13[aa]	100
28	9.16	10	11.26[aa]	100	10.50	10	-3.62	0
29	10.24	0	-2.15	0	12.09	10	-0.62	0
30	6.91	0	-2.57	0	7.66	0	4.58[aa]	100
31	12.12	0	-0.65	0	10.58	0	1.05	10
32	6.73	0	5.97[aa]	100	9.52	0	2.63[aa]	90
33	8.51	0	2.26[aa]	60	7.38	0	-3.56	0
34	9.01	0	.01	0	12.30	20	2.60[aa]	80
35	8.53	10	-2.42	0	9.39	0	-1.97	0
36	7.97	0	-1.73	0				
37	10.88	10	5.35[aa]	100				
38	8.01	0	1.24	10				
39	10.08	0	0.56	0				
40	7.79	0	0.79	10				
41	12.02	20	-0.39	0				
42	13.50	20	-2.43	0				
43	14.02	10	-2.32	0				
44	8.11	0	-0.38	0				
45	14.30	30	-0.49	0				
46	9.16	0	-1.34	0				
47	11.40	10	-0.87	0				
48	14.97	30	-3.67	0				
49	14.05	10	-1.95	0				
50	7.50	0	5.04[aa]	100				

[a] $p<.01$
[aa] $t>2.00$

The distributions of three-parameter model, Rasch model, and equipercentile equating method error estimates obtained by equating the reading or math test to itself by way of every pair of the 10 random samples were computed in both raw score and grade development score terms.[3] The statistic calculated was the absolute difference between raw or grade development scores averaged across the entire test score range. For each content area, 45 average absolute raw score differences and 45 average absolute grade development score differences were calculated and their distributions tabulated for each equating method.[4]

The empirical relative cumulative frequencies of average absolute differences for each equating method were compared to the corresponding theoretical uniform and normal distribution functions using the Kolmogorov-Smirnov goodness-of-fit test (Conover, 1980). The three-parameter empirical distributions did not differ significantly from the uniform or normal distributions for either content area; the Rasch distributions of average absolute differences differed significantly from the theoretical uniform distribution in reading but not math and did not differ significantly from the normal distribution in either content area. For the equipercentile method, the empirical distributions did not differ significantly from the theoretical uniform or normal distributions in either reading or mathematics. All significance tests were two-tailed with results reported as significant when $p < .05$. Clearly, the nonparametric goodness-of-fit tests were not very powerful. Comparing the empirical minus theoretical distribution differences, it appeared that the three-parameter, Rasch, and equipercentile empirical distributions of average absolute differences fit the normal distribution better than the uniform distribution, with the three-parameter results exhibiting the best overall fit.

Descriptive statistics for the distributions are reported in Table 3. The mean, standard deviation, maximum, and minimum average absolute differences by score type, content area, and equating method are reported. For both content areas, the mean of the average absolute differences across the 45 sample pairs was smaller for the Rasch model than for both the three-parameter model and the equipercentile method, both in raw score and grade development score terms. The equipercentile means of the average absolute differences were slightly smaller than the corresponding values for the three-parameter model across content areas and score types.

The mean of the distribution of average absolute differences provided one type of overall summary of the errors associated with equating a test to itself by way of each equating method. A much more conservative approach considered the maximum average absolute difference in raw scores or grade development scores obtained across the 45 replications.

A third approach was based on the mean square error term from the exploratory ANOVA results. These mean squares were .0352 and .0075, respectively, for the reading and mathematics tests. These mean square errors required only the experimental data for computation and were obtained for comparison to the error estimates of the original studies (one replication of equating a test to itself) and for comparison to the mean average absolute differences and maximum average absolute differences reported in the present study.

The results of the original studies for Grade 4 reading and math together with the additional error estimates obtained in the present study are reported in Table 4. In vertical order under each method, the values reported are the error estimates from the original study based on a single pair of random samples, the mean average absolute differences (mean of the empirical sampling distribution of errors of equating a test to itself), and the maximum average absolute differences from the present study. The numbers in parentheses in the far right column are the ANOVA mean square error terms. All other values are experimental results from the original studies. Each represents the average absolute grade

[3] The grade development score is similar to a grade equivalent score but is designed to be interpretable as an approximate instructional level indicator. For a more complete description of this score and its advantages and disadvantages see Cole (1982).

[4] See Phillips (1984) for complete frequency distribution tables.

Table 3
Descriptive Statistics For Average Absolute Equating Differences By Content Area

Model and Score	Reading	Mathematics
Three-Parameter Model		
Raw Score		
Mean	.695	.320
Standard Deviation	.342	.119
Maximum	1.446	.615
Minimum	0.071	.121
Grade Development Score		
Mean	.056	.020
Standard Deviation	.028	.009
Maximum	.118	.037
Minimum	.008	.005
Rasch Model		
Raw Score		
Mean	.279	.162
Standard Deviation	.028	.117
Maximum	.945	.477
Minimum	.008	.022
Grade Development Score		
Mean	.025	.012
Standard Deviation	.022	.009
Maximum	.085	.038
Minimum	.001	.002
Equipercentile Method		
Raw Score		
Mean	.513	.306
Standard Deviation	.187	.107
Maximum	1.092	.470
Minimum	.273	.106
Grade Development Score		
Mean	.042	.019
Standard Deviation	.015	.008
Maximum	.084	.045
Minimum	.021	.008

development score difference between the equating results for the methods heading its row and column.

In the original studies, significance was determined by comparing each experimental result with the error estimates for the two equating methods being compared. If the experimental value was greater than both error estimates, it was considered significant. All of the differences between pairs of methods were significant in the original studies. Similar comparisons have also been made in Table 4 using the error estimates from the present study. Except for the three-parameter/Rasch comparison in reading, which was not significant using the maximum average absolute difference to estimate error, all experimental comparisons remained significant no matter which value was used to quantify error.

Table 4
Original and Average Error Mean Absolute Differences Compared to Experimental Results in GDS Units By Content Area

	Equiper	Rasch	Log3p	Anova
Reading				
Equiper				
original error estimates	.0404	.2911**	.0615*	
mean replication error	.0415			
max. avg. error	.0843			
Rasch				
original error estimates		.0001	.2769**	
mean replication error		.0252		(MS=.0352)
max. avg. error		.0850		
Log3p				
original error estimates			.0154	
mean replication error			.0556	
max. avg. error			.1175	
Mathematics				
Equiper				
original error estimates	.0094	.1100**	.0889**	
mean replication error	.0194			
max. avg. error	.0448			
Rasch				
original error estimates		.0188	.1333**	
mean replication error		.0118		(MS=.0075)
max. avg. error		.0380		
Log3p				
original error estimates			.0148	
mean replication error			.0199	
max. avg. error			.0371	

** "significant" = greater than original error estimates, mean replication error estimates and maximum average error estimates for both methods being compared and greater than the ANOVA mean squared error estimate.

* "significant" = greater than all of the above (**) except maximum average error estimates for both methods being compared.

Compared to the original single-replication error estimates, the mean average absolute differences were larger (except for Rasch math). The single-replication Rasch error estimates fell approximately at the 1st and 84th percentile ranks in the empirical sampling distributions for reading and math, respectively. The corresponding percentile ranks for the single-replication three-parameter error estimates were 9 and 31. The single-replication equipercentile error estimates from the original study fell at approximately the 56th and 1st percentile ranks in the equipercentile empirical sampling distributions for reading and mathematics, respectively. If the mean average absolute differences are

averaged for the three equating methods, the result is approximately equal to the ANOVA mean square for reading and about double the ANOVA mean square for math. Similar comparisons with the averages of the single replication errors for all three methods indicate a mean square about twice as large in reading and of approximately the same magnitude in math.

Discussion

The results of this study suggest that single-replication error estimates may provide misleading assessments of the errors associated with equating a test to itself. In four out of six cases these estimates were in the lower third of the empirical sampling distribution, and in one case the estimate fell in the upper third of the empirical sampling distribution. However, obtaining the sampling distribution for every new application may not be possible due to limited resources.

In relating the experimental results of the original study to the error estimates obtained in the present study, the results remained essentially the same. However, it cannot be concluded that similar results would be obtained in other applications. In addition, it may be argued that the error estimates of the present study are of inadequate magnitude for the experimental comparisons, because equating a test to itself does not include errors related to unequal item difficulties and equating through a scaling test as was done in the original studies. However, the empirical sampling distributions do suggest the magnitude of chance sampling errors that might be expected in this particular application.

The analysis of variance mean squares appeared somewhat promising as alternatives to error estimates by replication. The mean squares tended to be a bit conservative for the three-parameter empirical sampling distribution (33rd and 11th percentile ranks, respectively, for reading and math), and tended to overestimate the error for the Rasch model (80th and 56th percentile ranks, respectively, in the empirical sampling distributions for reading and math). For the equipercentile method, the reading mean square was slightly above the median of the empirical sampling distribution but the math mean square was very conservative (56th and 1st percentile ranks, respectively).

The form of the empirical sampling distributions to be expected when equating an achievement test to itself is still indeterminate. The uniform and normal distributions were selected for comparison in the present study due to their familiarity to researchers and practitioners. Particularly for the Rasch error sampling distributions, there are probably other theoretical distributions that fit the data better.

Finally, the results of this study together with those of the Phillips (1983a) study suggest that the Rasch model may be more reliable than other IRT models for equating but in some applications less valid. In the present study the errors for the Rasch model were smaller than those for the three-parameter model, suggesting greater consistency of estimation for the Rasch model. If the equipercentile method is used as a criterion (and this is debatable among researchers), the Rasch method appears less valid for equating achievement tests than other IRT models. Using other criteria, other researchers have also arrived at the same conclusion.

There is still a need for better methods for establishing a criterion against which to judge the results of various equating methods. The present study provides data relevant to the errors to be expected when the criterion is an achievement test equated to itself. Another technique that might be considered as a more general procedure involves basing the statistical analysis on equating the pair of tests of interest over repeated examinee samples, including the study of the correlation between sampling errors for pairs of equating methods over replications. However, such a procedure may not be feasible in practical applications because it would double the IRT estimation costs.

References

Beard, J. G., & Pettie, A. L. (1979). *A comparison of linear and Rasch equating methods results for basic skills assessment tests*. Paper presented at the American Educational Research Association, San Francisco.

Bell, A. (1979). *A comparison of three equating procedures on the certifying examination for primary care*

physicians assistants. Paper presented at the American Educational Research Association Annual Meeting, San Francisco.

Blommers, P. J., & Forsyth, R. A. (1977). *Elementary statistical methods in psychology and education* (2nd ed.). Boston: Houghton-Mifflin.

Cole, N. S. (1982). *Grade equivalent scores: To GE or not to GE*. AERA Division D Vice Presidential Address, American Educational Research Association Annual Meeting, New York City.

Cole, N. S., Trent, E. R., & Wadel, D. C. (1982). *The 3-R's technical manual achievement edition* (Grades K-12). Chicago IL: The Riverside Publishing Company.

Conover, W. J. (1980). *Practical nonparametric statistics*. New York: Wiley.

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 American Educational Research Association Annual Meeting, Montreal.

Golub-Smith, M. (1980). *The application of Rasch model equating techniques to the problem of interpreting longitudinal performance on minimum competency tests*. Paper presented at the American Educational Research Association Annual Meeting, Boston.

Green, S. B. (1983). Identifiability of spurious factors using linear factor analysis with binary items. *Applied Psychological Measurement, 7,* 137–147.

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

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

Kolen, M. J. (1981). Comparison of traditional and latent trait 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,* 297–308.

Lindquist, E. F. (1956). *Design and analysis of experiments in psychology and education*. Boston: Houghton-Mifflin.

Lord, F. M. (1975). *A survey of equating methods based on item characteristic theory* (Research Bulletin 75–13). Princeton NJ: Educational Testing Service.

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

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

Lord, F. M. (1982). Standard error of an equating by item response theory. *Applied Psychological Measurement, 5,* 463–472.

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

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139–160.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1979). *Applicability of two logistic models for equating test scores when tests and samples are varied*. Paper presented at the American Educational Research Association Annual Meeting, San Francisco.

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

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

Phillips, S. E. (1983b). *Logistic achievement test scaling and equating with fixed versus estimated lower asymptotes*. Paper presented at the National Council on Measurement in Education Annual Meeting, Montreal.

Phillips, S. E. (1984). *Quantifying errors in IRT equating methods*. Paper presented at the American Educational Research Association Annual Meeting, New Orleans.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Rentz, R. R., & Bashaw, W. L. (1977). The National Reference Scale for Reading: An application of the Rasch model. *Journal of Educational Measurement, 14,* 161–180.

Woods, E. M., & Wiley, D. E. (1978). *An application of item characteristic curve equating to item sampling packages for multiform tests*. Paper presented at the American Educational Research Association Annual Meeting, Toronto.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum 76–6). Princeton NJ: Educational Testing Service.

Wright, B. D., Mead, R. J., & Bell, S. R. (1979). *BICAL: Calibrating items with the Rasch model* (Statistical Laboratory Research Memorandum No. 23B). Chicago: University of Chicago, Department of Education.

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

Acknowledgments

The author thanks Arlin E. Anderson for computer programming assistance. An earlier version of this paper was presented at the 1984 American Educational Research Association annual meeting in New Orleans.

Author's Address

Send requests for reprints or further information to S. E. Phillips, 458 Erickson Hall, College of Education, Michigan State University, East Lansing MI 48824, U.S.A.