DOCUMENT RESUME

ED 395 955 TM 025 032

AUTHOR	Lawrence, Ida M.; Dorans, Neil J.
TITLE	A Comparison of Observed Score and True Score Equating Methods for Representative Samples Matched on an Anchor Test.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-88-23
PUB DATE	Apr 88
NOTE	27p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; *College Entrance Examinations; Comparative Analysis; *Equated Scores; *Item Response Theory; *Sampling
IDENTIFIERS	*Anchor Tests; Equipercentile Equating; Levine Equating Method; Linear Fquating Method; *Scholastic Aptitude Test; Three Parameter Model; Tucker Common Item Equating Method

ABSTRACT

This paper addresses the sample invariant properties of four equating methods (Tucker and Levine linear equating, equipercentile equating through an anchor test, and three-parameter item response theory equating). Data from several national administrations of the Scholastic Aptitude Test served as the source of data for the study. Equating results across two sampling conditions, "representative" sample and "matched" sample, were compared to determine which equating procedures produced the most consistent results. In the representative sample condition, equatings were based on old-form and new-form samples that differed in ability; in the matched sample condition, the old-form sample was selected to match the anchor test score distribution of the new-form sample. Results for the item response theory equating method differed for representative and matched samples, as did the equating results for the Levine and equipercentile methods. Results based on the Tucker observed score equating method were found to be essentially the same across representative and matched sample conditions. Results for the four equating methods tended to converge under the matched sample condition. The last section of this paper offers tentative explanations for the findings. An appendix contains formulas for the equating methods discussed. There are several series of unreadable numbers in the appendix. (Contains 2 figures, 3 tables, and 11 references.) (Author/SLD)

RR-88-23

RESEARCH REPORT

A COMPARISON OF OBSERVED SCORE AND TRUE SCORE EQUATING METHODS FOR REPRESENTATIVE SAMPLES AND SAMPLES MATCHED ON AN ANCHOR TEST

Ida M. Lawrence
Neil J. Dorans

Educational Testing Service
Princeton, New Jersey
April 1988

A COMPARISON OF OBSERVED SCORE AND TRUE SCORE EQUATING METHODS FOR REPRESENTATIVE SAMPLES AND SAMPLES MATCHED ON AN ANCHOR TEST[1]

Ida M. Lawrence

Neil J. Dorans

College Board Statistical Analysis

Educational Testing Service

April 1988

[1]The authors thank W. Angoff, D. Eignor, P. Holland, C. Lewis, S. Livingston, G. Marco, N. Petersen, M. Stocking, and D. Wright for reviewing earlier drafts of this manuscript, and Miriam Feigenbaum for providing statistical support.

Abstract

This paper addresses the sample invariant properties of four equating methods (Tucker and Levine linear equating, Equipercentile equating through an anchor test, and three-parameter item response theory equating). Equating results across two sampling conditions, "representative" sample and "matched" sample, were compared to determine which equating procedures produced the most consistent results. In the representative sample condition, equatings were based on old-form and new-form samples that differed in ability; in the matched sample condition, the old-form sample was selected to match the anchor test score distribution of the new-form sample.

Results for the item response theory equating method differed for representative and matched samples, as did the equating results for Levine and Equipercentile methods. Results based on the Tucker observed-score equating method were found to be essentially the same across representative and matched sample conditions. Results for the four equating methods tended to converge under the matched sample condition. The last section of this paper offers tentative explanations for the findings.

A COMPARISON OF OBSERVED SCORE AND TRUE SCORE EQUATING METHODS FOR REPRESENTATIVE SAMPLES AND SAMPLES MATCHED ON AN ANCHOR TEST

A desired outcome of any score equating procedure is that the equating transformation for a test be the same regardless of the candidate group from which it is derived. Under such circumstances, the equating transformation is said to be population invariant, or population independent. Lack of invariance, or population dependence, implies that a score on test X may be equivalent to one score on test Y in population P and to another score on test Y in population Q. The purpose of this study was to assess the invariance assumption, by comparing the results of Tucker, Levine, IRT and Equipercentile equating methods under two sampling conditions: representative samples and samples matched on an anchor test.

The issue of population invariance has been addressed in other educational testing programs. Several studies designed to evaluate the population invariance property of different equating methods were described by Cook and Petersen (1987) and this discussion draws heavily from that paper. In one study (Angoff & Cowell, 1986), two randomly equivalent base populations of examinees, each taking a different form of the Graduate Record Examination General Test, were used to develop a population conversion in which scores on one form were equated to scores on the other. Eleven variously defined and characteristically different subgroups were then selected from the base population. To evaluate population invariance, the researchers examined the degree of agreement or disagreement between each subpopulation conversion and the base population conversion. In that study the equating samples within each subpopulation were assumed to be similar in ability (that is, to the extent that spiralling, i.e., alternation of test booklets within the same

administration, had the desired effect), and random-group equating methods (linear and equipercentile without an anchor test) were used. In a second study (Kingston, Leary, & Wightman, 1985), Item Response Theory (IRT) true-score equating was used to equate scores on the Graduate Management Admissions Test in a variety of populations. As in the Angoff and Cowell study, the equating samples within each population were similar in ability. For the most part, both of these studies found negligible differences among conversions based on different subpopulations.

Results from the aforementioned studies indicated that, in general, equatings were consistent across subpopulations of different ability. These studies, however, involved equating samples from old form and new form populations of approximate equal ability. Neither study addressed the issue of the population independence of equatings where old-form and new-form populations differ in ability. Studies by Cook, Eignor & Taft (1985) and by Cook (1984) have attempted to study this issue in the context of achievement test equating. While the assumption of invariance did not hold for conventional and IRT equating methods in these studies, the authors speculated that the discrepant equatings might be attributed to an interaction between training and test content, such that the same tests were measuring different constructs in different populations.

The Cook et al. research seemed to suggest that the validity of the population independence assumption can be questioned under certain circumstances. The circumstance addressed in the present study is one where essentially parallel tests are equated in subpopulations differing in ability.

Traditionally, SAT equatings have been based on what will be referred to as "representative samples". The old form sample is a representative sample of juniors and seniors who took both the old form of the SAT and the equating

(anchor) test linking it to the new form. Similarly, the new form sample is a representative sample of juniors and seniors who took both the new form of the SAT and the equating test linking it to the old form.

Several SAT equatings have also involved the use of what will be referred to as "matched samples". The new form sample is a representative sample, as described above. The old-form sample is selected from a subpopulation of juniors and seniors who have taken the old form and the anchor test, by using the students' scores on the anchor test as a stratifying variable. A separate sample is selected at each score level on the anchor test, so that the old form sample includes the same number of students at each anchor test score level as there are in the new form sample. That is, the distribution of anchor test scores is made to be the same in the old form and new form samples, even though this distribution may be quite different from the representative sample drawn from the old form population.

Comparisons of the results of representative sample equatings and matched sample equatings have shown that several equating methods produce systematically different results under the two sampling plans. In particular, equipercentile equating, Levine true-score linear equating, and IRT true-score equating have demonstrated a sensitivity to population differences that has not been observed with linear Tucker equating. This paper illustrates these differential sensitivities and offers some explanation for why certain equating methods do not produce equivalent results under the two sampling conditions.

Procedure

Data Source

Equating data from several national administrations of the SAT served as the source of data for the study. Equating results under matched sample and

representative sample conditions were compared for nine forms of SAT-Mathematical and six forms of SAT-Verbal and (see Tables 1 and 2).

Matching

The equating data collection design for this study (see Figure 1) was as follows: one test form (X) is administered to one group of examinees, a second form (Y) is administered to a second group of examinees, and a third form (Z) is administered to a third group of examinees. The populations taking forms X and Y represent populations of similar ability, and the group of examinees taking form Z often represents either a more able or less able candidate population. Form X (the new form) is linked to Form Y via one equating test (W) and to Form Z via another equating test (V). For typical equatings of the SAT, the average of anchor equatings to the old forms is taken as the operational conversion for the new form.

The focus of this study was on the equating of X to Z. Under representative sampling (i.e. samples drawn to represent their parent population), the equating samples for Forms X and Z often represent populations of differing ability. Under matched sampling, the equating samples for forms X and Z are constructed to represent equal ability populations. This is because scores on equating test V have been used to match the raw-score distribution for the old-form sample (examinees who took Form Z) to the raw-score distribution for the new-form sample (examinees who took Form X).

Thus, while the distribution for the new-form sample is held fixed, the distribution for the old-form sample is altered under matched sample conditions to be similar to that of the new-form sample. Depending on the direction of ability differences between the new-form and old-form samples, the old-form sample is changed to reflect either a more able or a less able candidate population. This matching procedure is a means of controlling for

the effects on score equating of differences in the abilities of the populations.

Equating Methods

This study involved the evaluation of equating results from four anchor equating methods used routinely to equate scores on the SAT. Two linear methods (Tucker observed-score and Levine true-score) and two curvilinear methods (equipercentile and IRT true-score) were evaluated. With the anchor test data collection design, one group of examinees (group a) takes the old form and an anchor test, another group (group b) takes the new form and the same anchor test.

The two linear methods used in this study were the Tucker observed-score model and the Levine method for Equally Reliable tests (Angoff, 1971). Under linear equating, scores are said to be equated if they correspond to the same number of standard deviation units from the mean in some population of examinees. Scores on the anchor test are used to estimate the performance of a combined group ($c = a + b$) of examinees on both the old and new forms of the test. Both the Tucker and the Levine models produce an equating transformation of the form

$$e_Z(X) = AX + B, \tag{1}$$

where $e_Z(X)$ is the function equating test X to test Z, and A and B are the parameters of the equating transformation.

The Tucker model assumes that the regression of total test X on the anchor test V is linear and homoscedastic, and that this regression, which is observed in the sample that took test X with test V, also holds in the sample that took test Z with test V. A similar set of assumptions is made about the regression of test Z on anchor test V. Formulae for obtaining the A and B parameters for the Tucker model are presented in the Appendix.

The Levine method for Equally Reliable tests assumes that the ratio of the standard deviation of true scores on X to the standard deviation of true scores on V is the same in groups a and c. In addition, it assumes that the intercept of the regression line relating true scores on X to true scores on V is the same for groups a and c. A similar set of assumptions are made about true scores on Z and V in groups b and c. Formulae for obtaining the A and B parameters for the Levine model are presented in the Appendix.

For both linear models, scores on the external anchor are used to estimate performance of the combined groups of examinees on both the old and new forms of the test (actually taken by two different groups). According to Angoff (1971), the Tucker model is assumed to be an appropriate linear model for groups not widely different in ability and the Levine model, which makes use of true-score relationships, is recommended for samples of different ability (Angoff, 1971).

The equipercentile method employed in this study is what Angoff (1971) refers to as Design V, and what Braun and Holland (1982, pp. 39-42) call equating two tests through a third test. In group a, test X is equated to the anchor test V such that equated scores correspond to the same percentile rank of examinees in group a. Similarly, scores on test Z earned by group b are equated to scores on V such that equated scores correspond to the same percentile rank of examinees in group b. Raw scores on test X and test Z are said to be equivalent if they correspond to the same raw score on the anchor test V. Note that scores on X and Z are never actually equated in the same population, be it real or synthetic.

The item response theory true-score equating model is based on assumptions germane to item response theory. In particular, it assumes that there is a mathematical function that describes the relationship between an

examinee's ability and the probability that the examinee will answer the item correctly (Lord, 1980). The model used for SAT equating is the three-parameter logistic model in which the probability of a correct response is

(2) $P_i(\theta) = c_i + [(1-c_i)/1+e^{-1.702a_i(\theta-b_i)}]$

In (2), θ represents examinee ability theta, a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the lower asymptote and e is the exponential function. The computer program LOGIST V (Wingersky, Barton & Lord, 1982) is used to obtain item parameter estimates for tests X, Z, and V together in one large concurrent calibration (which also includes samples S_{11} and S_{12} from the X to Y part of the equating data collection design, see Figure 1). True scores on X and Z are said to be IRT true-score equated if they correspond to the same value of θ. Clearly, the item parameters play a central role in IRT true-score equating. IRT true-score equating uses item parameter estimates to obtain the IRT equating function. Further details regarding IRT true-score equating are presented in the Appendix.

Comparison of Equating Results

In order to compare equating results for the four methods within and across representative sample and matched sample conditions, scaled score means and standard deviations were projected for each test form. These projected summary statistics were computed by applying the respective conversion to raw score distributions for the candidate population.

Results

Scaled score means and standard deviations for the various equating procedures under the conditions of representative and matched samples are displayed in Table 1 (SAT-Mathematical) and Table 2 (SAT-Verbal). The third and fourth columns of each table show the effects of matching on subpopulation

ability differences. Prior to matching, the absolute value of the mean difference in ability (as measured by raw scores on the equating test) between the new-form sample and the old-form sample for SAT-Mathematical ranged between .192 and .390 standard deviations. Prior to matching, the new-form and old-form samples also differ in terms of variablility, as indicated by ratios of standard deviations (given in parentheses in the third and fourth columns). The next column in the table shows the degree to which it was possible to match new-form and old-form samples with respect to raw score distributions on the equating test. Note that, as expected, under matched sample conditions the standardized mean differences are close to zero and the ratios of standard deviations are close to one. Perfect matching of the old form sample to the new form sample (i.e., mean difference equal to zero and ratio of standard deviations equal to 1.00) was achieved only twice: for equatings involving SAT-Mathematical Form 9M and SAT-Verbal Form 3V. Due to insufficient examinees in the tails of the distributions for some of the old-forms, it was not always possibly to attain perfectly matched samples.

Examination of the data contained in Table 1 indicates that, for the representative sample equatings, the means for the Tucker method appear to be most extreme; they differ considerably from the means produced by the other methods. When the new-form sample is less able than the old-form sample (as was the case in January 1986 and December 1986), the Tucker mean is higher than the means produced by the other three equating methods. Conversely, when the new-form sample is more able than the old-form sample, the Tucker mean is lower relative to the means produced by the other methods. The Levine and IRT means appear to agree fairly well under representative sample conditions.

The interesting finding in these data concerns the effect of matched sample equating. Of the four equating methods, only the means for the Tucker

conversion are invariant across representative and matched sample conditions. Thus, despite the shift in ability for half of the population (i.e., the old-form sample is made either more able or less able), the Tucker conversion is not seriously affected by sampling. The results presented in Table 1 suggest that while matching on observed scores does not affect Tucker equatings, the conversions for the other methods are different under the two sampling conditions. Moreover, under matched samples, the means for the other methods were brought closer to the invariant Tucker mean.

When the matched sample is more able than the original representative sample, the mean scaled score based on the matched sample equating is lower than the mean scaled score based on the representative sample equating. The opposite relationship is found when the matched sample is less able than the original representative sample. The results for the Levine and Equipercentile equatings also follow this pattern.

Similar patterns emerge for the SAT-Verbal data (see Table 2), although to a lesser extent. While the results for the Tucker conversion agree most closely across representative sample and matched sample conditions, means based on this equating method are not as stable across forms as those observed for the SAT-Mathematical data. Still, results from the four methods appear to converge under matched sample conditions.

Findings from Tables 1 and 2 are shown graphically in Figure 2, where the mean difference between representative and matched sample equatings is plotted against the standardized mean difference in ability between the new-form sample and the old-form sample. The values along the vertical axis indicate, for each method, the difference in mean scaled scores for representative sample and matched sample equatings. The values along the horizontal axis represent the magnitude of shift in ability needed to match the old-form

sample to the new-form sample. Larger values correspond to more extreme differences in ability between matched sample equating versus random sample equating. Therefore, administrations with larger differences reflect a more stringent evaluation of the degree to which the equating function holds up across different ability subpopulations.

Figure 2 (panel a) reveals the tendency for the Levine and IRT methods to perform similarly for the SAT-Mathematical data. The lack of invariance for these two methods is related in a somewhat linear fashion to the shift in ability across sampling conditions. For projections based on the Tucker conversion line, differences between representative samples and matched samples are close to zero.

Figure 2 (panel b) shows that the Tucker equatings are less invariant for the SAT-Verbal data, relative to the SAT-Mathematical data. Again, however, there is a tendency for the Levine and IRT methods to demonstrate instability across samples (as evidenced by non-zero differences in scaled score means for these methods).

Some Possible Explanations for the Observations

Under the matched sample condition, both the Tucker and Levine equating models reduce to the simple mean/sigma equating method in which means and standard deviations are set equal for the new-form and old-form samples. This convergence to the mean/sigma equating model occurs because the new form sample and the matched old form sample are equivalent with respect to score distributions on the anchor test. Note, if the assumptions of the Tucker model hold, equatings based on that model should be invariant with respect to sampling. This is because matched samples are produced via direct selection on the anchor test (observed scores), and the regression of the total test on the anchor test is invariant with respect to that type of selection.

In contrast to the Tucker model, the true score regressions of the Levine model are subject to selection effects when samples are matched on an observed equating score. This indirect, rather than direct, selection on the anchor test true scores results in non-zero mean error scores in the matched sample, and consequently violates an assumption of classical test theory that is made in Levine equating, i.e., the assumption of mean error scores equal to zero. The positive or negative direction of the non-zero mean error scores will be systematically related to the direction in which the matched sample mean will change from the representative sample mean.

For IRT equating, the sample dependency can be traced to shifts in item parameter estimates (primarily the item difficulty parameter, although the discrimination parameter also shifts slightly) between the matched sample item calibration and the representative sample item calibration. The shift in the b-parameter estimates is limited, for the most part, to the old form that was taken by the matched sample. Data pertaining to this issue are presented in Table 3. This table presents summary statistics for item parameters resulting from representative sample and matched sample concurrent calibrations involving the X to Z part of the equating data collection design (see Figure 1). In order to make the data comparable across separate calibrations, item parameter estimates were placed on the same scale by means of an item parameter transformation method (Stocking and Lord, 1983), which used item parameters from the remaining components of the calibration matrix (i.e., excluding tests Z and V) as common items.

While item parameter estimates based on IRT are supposed to be invariant with respect to differences in the samples upon which they are based, the crucial point to note from Table 3 is that the mean item difficulty value for old form Z is higher in the matched sample than in the representative sample

(i.e. items in the matched sample appear to be more difficult). A possible explanation for this phenomenon has been offered by C. Lewis (personal communication, October 21, 1987), and it is as follows. In selecting the matched sample (which took form Z) to have the same distribution of observed scores on the anchor test (V) as the new form sample (which took form X), an old form sample has been obtained that is characterized by somewhat higher than average observed scores relative to the representative sample from which it was selected. This selection on observed scores produces a sample with higher than average item-ability regressions (percent correct given true score theta) for the items composing the anchor test used for matching. This is analogous to positive mean error scores in classical test theory. The item-ability regressions for the anchor test items in the new form sample are not affected by the selection process and the percent correct given true score should reflect population values.

In order to reconcile these differences in item ability regressions for the same anchor items given to equal ability samples, LOGIST could fit the data by shifting the regressions from the matched sample until they coincide with the regressions from the new form sample. The result of this shift would be to estimate the matched sample examinees as more able than they actually are. Put differently, for the matched sample the items in the anchor test V will appear to be easier relative to the items in old form Z, and the old form will consequently appear more <u>difficult</u>. This phenomenon is observed in Table 3, where we see that the data for Form Z show a mean item difficulty from the matched sample analysis which is .08 greater (more difficult) than the corresponding mean for the representative sample analysis, where selection on the anchor test has not taken place. As a consequence of estimating the old form to be more difficult than it actually is, the matched sample equating for

the new form results in lower converted scores and lower means, compared to the representative sample. On the other hand, when the old form sample is matched to a less able group (i.e., as in 12/86), the IRT conversion for the matched sample is higher that the corresponding conversion for the representative sample.

With respect to the findings for the equipercentile equating method, no explanation is available to explain why this observed score method is not invariant in the same way that the Tucker observed score method is invariant. One possibility is that this particular type of equipercentile method involves equating a more reliable test (the total test) to a less reliable test (the anchor test), and this difference in reliabilities is affected by changes in ability differences across representative and matched sample conditions.

Discussion

This study looked at the sensitivity of equating results to differences in population ability, using matched samples. Some tentative explanations for the findings were offered.

The results of this study have demonstrated that matched sampling may be a useful technique for controlling for population differences prior to equating, and avoids the problems associated with disagreement among equating methods often observed in the representative sample situation. In effect, using the equating test as a direct selection variable produces a convergence of equating results across different equating methods. Of the four equating methods studied, the Tucker model produced scaled score means and standard deviations that were essentially the same across representative and matched sample conditions. The means for the other methods tended to converge under matched sampling. These results should not be taken to mean that equatings based on the Tucker model are closer to the "true" equating results than are

the results from the other methods; they simply say that the Tucker equating method is more nearly population independent than the other methods. Further research is needed to evaluate matched-sample results when there is a known criterion.

One possible weakness in the method followed in this study is the use of the same variable (the equating test) for matching and for equating. For true-score methods, the magnitude of bias resulting from matching examinees on the same measure used to adjust for ability differences in equating is presently unknown. Another approach to the study of population invariance would involve the use of an external measure of ability (other than the equating test) as the basis for selecting subpopulations of different ability levels. One possibility is to use data from a large SAT administration and perform an equating study on subpopulations selected (on the basis of SAT scores) to be similar to the low, medium, and high ability administrations typically observed in a given testing year. Equating results for the various subpopulations could then be compared. In addition to using real data, this kind of study could also be conducted with simulated data designed to look like typical SAT data.

References

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement (pp.508-600). Washington, DC: American Council on Education. (Reprinted by Educational Testing Service, Princeton, NJ, 1984).

Angoff, W.H. & Cowell, W.R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. Journal of Educational Measurement, 23, 327-345.

Braun, H.I., & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland and D.B. Rubin (Eds.), Test equating (pp. 9-49). New York, NY: Academic press.

Cook, L.L. (1984). Equating refurbished achievement tests (unpublished statistical report). Princeton, NJ: Educational Testing Service.

Cook, L.L., Eignor, D.R., & Taft, H.L. (1985). A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates (RR-85-38). Princeton, NJ: Educational Testing Service.

Cook, L.L. & Petersen, N.S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied Psychological Measurement, 11, 225-244.

Kingston, N., Leary, L., & Wightman, L. (1985). An exploratory study of the applicability of item response theory methods to the Graduate Management Admissions Test (RR-85-34). Princeton, NJ: Educational Testing Service.

Lewis, C. (1987). Personal communication.

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

Stocking, M.L. & Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). LOGIST V user's guide. Princeton, NJ: Educational Testing Service.

Appendix

Tucker Linear Conversion Parameters

$$A = [S^2_{zb} + C^2_{zvb}(S^2_{vc} - S^2_{vb})/S^4_{vb}]^{1/2} \;/\; [S^2_{xa} + C^2_{xva}(S^2_{vc} - S^2_{va})/S^4_{va}]^{1/2}$$

$$B = M_{zb} + C_{zvb}(M_{vc} - M_{vb})/S^2_{vb} - AM_{xa} - AC_{xva}(M_{vc} - M_{va})/S^2_{va}$$

where M, S, and C refer to mean, standard deviation, and covariance, respectively; group a takes new form X and anchor test V, group b takes old form Z and anchor test V, and group c is the composite of group a and group b.

Levine Linear Conversion Parameters

$$A = [S^2_{zb} + (S^2_{zb} - S^2_{z''b})(S^2_{vc} - S^2_{vb})/(S^2_{vb} - S^2_{v''b})]^{1/2} \;/$$
$$[S^2_{va} + (S^2_{xa} - S^2_{x''a})(S^2_{vc} - S^2_{va})/(S^2_{va} - S^2_{v''a})]^{1/2}$$

$$B = M_{zb} + (M_{vc} - M_{vb})[(S^2_{zb} - S^2_{z''b})/(S^2_{vb} - S^2_{v''b})]^{1/2} - AM_{xa}$$

where x", z", and v" refer to the errors of measurement on the new form, old form, and anchor test, respectively.

Item Response Theory Equating

For test X, one obtains the expected value of the examinee's formula true score (τ_x) via a transformation of the test characteristic curve $\Sigma P_i(\theta)$, i.e.,

$$\tau_x = \sum_{i=1}^{n} [(k_i+1)P_i(\theta)/k_i - k_i^{-1}].$$

Likewise, for test Z, one can obtain

$$\tau_z = \sum_{j=1}^{n} [(k_j+1)P_j(\theta)/k_j - k_j^{-1}],$$

the expected formula true score as a function of θ; (k_i+1) and (k_j+1) are the number of choices per item on X and Z, respectively. To obtain equated scores, paired values of τ_x and τ_z are computed at each value of θ.

Table 1

Projected Scaled Score Means and Standard Deviations for SAT-Math Equatings

Admin.	Form	Stz. Diff. in Ability (new-old) and Ratio of Standard Devs. (new/old): Repr. Sample	Matched Sample	Equating Method	Means: Repr. Sample	Matched Sample	Diff. (R-M)	Standard Deviations: Repr. Sample	Matched Sample	Ratio R/M
10/85	1M	.390	.024	Tucker	482.5	481.3	1.2	118.0	117.1	1.0077
		(1.077)	(.984)	Levine	492.9	482.1	10.8	118.9	116.7	1.0189
				IRT	496.2	487.1	9.1	114.0	113.6	1.0035
				Equip.	489.6	482.5	7.1	115.2	116.8	.9863
10/85	2M	.371	.007	Tucker	482.2	481.5	0.7	120.2	119.9	1.0025
		(1.108)	(1.013)	Levine	493.6	480.9	12.7	123.3	119.2	1.0344
				IRT	494.3	485.5	8.8	117.6	117.6	1.0000
				Equip.	488.3	482.5	5.8	117.2	118.9	.9857
11/85	3M	.367	.001	Tucker	482.4	483.0	-0.6	119.6	119.3	1.0024
		(.971)	(.999)	Levine	491.4	483.0	8.4	119.7	119.3	1.0033
				IRT	492.4	487.1	5.3	115.4	116.0	.9948
				Equip.	489.4	484.8	4.6	116.4	120.4	.9668
1/86	4M	-.192	-.001	Tucker	460.8	462.8	-2.0	118.1	118.8	.9941
		(1.052)	(1.003)	Levine	457.2	462.8	-5.6	118.7	118.8	.9992
				IRT	456.2	460.5	-4.3	120.0	120.3	.9975
				Equip.	457.3	462.9	-5.6	119.1	118.5	1.0051
5/86	5M	.288	.032	Tucker	483.1	483.4	-0.3	116.2	116.4	.9983
		(1.125)	(1.092)	Levine	488.9	483.9	5.0	118.5	118.1	1.0034
				IRT	490.8	484.9	5.9	117.2	116.8	1.0034
				Equip.	487.7	483.6	4.1	117.4	117.1	1.0026
6/86	6M	.297	.001	Tucker	484.0	483.8	0.2	115.7	116.1	.9966
		(.889)	(1.004)	Levine	488.8	483.8	5.0	113.0	116.1	.9733
				IRT	487.5	482.9	4.6	114.4	115.4	.9913
				Equip.	487.0	483.3	3.7	113.0	115.6	.9775
10/86	7M	.198	-.019	Tucker	484.9	484.5	0.4	119.4	116.8	1.0223
		(.906)	(.998)	Levine	489.5	484.5	5.0	117.9	116.8	1.0094
				IRT	488.6	483.4	5.2	113.5	112.5	1.0089
				Equip.	489.0	485.6	3.4	117.0	117.4	.9966
10/86	8M	.233	.016	Tucker	487.1	487.1	0.0	121.1	118.9	1.0185
		(.918)	(1.011)	Levine	492.2	487.3	4.9	119.6	118.7	1.0076
				IRT	491.7	486.5	5.2	114.9	114.1	1.0070
				Equip.	490.4	488.0	2.4	118.3	119.0	.9941
12/86	9M	-.259	.000	Tucker	454.9	454.9	0.0	112.8	113.0	.9982
		(.979)	(1.000)	Levine	449.5	454.9	-5.4	112.0	113.0	.9912
				IRT	447.4	451.5	-4.1	113.9	114.3	.9965
				Equip.	450.5	454.8	-4.3	112.0	112.9	.9920

Note. Projected means and standard deviations were obtained by applying rounded conversion lines to preliminary raw score distributions for the total group of examinees (except Form 4M, where distributions for juniors and seniors were used instead).

Table 2

Projected Scaled Score Means and Standard Deviations for SAT-Verbal Equatings

Admin.	Form	Stz. Diff. in Ability (new-old) and Ratio of Standard Devs. (new/old) Repr. Sample	Matched Sample	Equating Method	Means Repr. Sample	Means Matched Sample	Diff. (R-M)	Standard Deviations Repr. Sample	Standard Deviations Matched Sample	Ratio R/M
1/86	1V	-.250	-.006	Tucker	409.3	411.3	-2.0	105.2	103.7	1.0145
		(1.039)	(1.007)	Levine	405.3	411.1	-5.8	105.9	103.5	1.0232
				IRT	405.5	409.5	-4.0	102.9	102.6	1.0029
				Equip.	406.3	410.7	-4.4	105.4	103.3	1.0203
5/86	2V	.277	.010	Tucker	429.2	428.5	0.7	107.3	107.7	.9963
		(1.030)	(1.037)	Levine	435.0	428.9	6.1	108.0	108.3	.9972
				IRT	435.5	430.4	5.1	109.5	109.1	1.0037
				Equip.	433.2	428.2	5.0	107.6	108.6	.9908
6/86	3V	.226	.000	Tucker	425.3	427.0	-1.7	102.5	103.4	.9913
		(.906)	(1.000)	Levine	430.0	427.0	3.0	100.8	103.4	.9749
				IRT	430.4	427.9	2.5	103.7	104.5	.9923
				Equip.	428.9	427.0	1.9	100.6	102.9	.9776
10/86	4V	.158	.005	Tucker	433.4	431.4	2.0	105.8	107.2	.9869
		(.968)	(1.003)	Levine	435.8	431.5	4.3	105.3	107.3	.9814
				IRT	435.8	432.3	3.5	107.2	107.6	.9963
				Equip.	435.5	431.4	4.1	104.9	106.7	.9831
10/86	5V	.144	-.009	Tucker	435.7	433.4	2.3	105.0	106.3	.9878
		(.973)	(1.008)	Levine	438.4	433.5	4.9	104.8	106.3	.9859
				IRT	440.3	436.1	4.2	107.6	108.2	.9943
				Equip.	437.5	433.9	3.6	104.5	106.8	.9785
12/86	6V	-.224	-.001	Tucker	403.0	402.7	0.3	100.9	100.2	1.0070
		(.983)	(1.004)	Levine	398.6	402.7	-4.1	100.3	100.3	1.0000
				IRT	398.0	401.4	-3.4	100.3	101.1	.9921
				Equip.	399.8	402.7	-2.9	100.0	99.9	1.0010

Note. Projected means and standard deviations were obtained by applying rounded conversion lines to preliminary raw score distributions for the total group of examinees (except Form 1V, where distributions for juniors and seniors were used instead).

Table 3

Means and Standard Deviations of LOGIST Parameter Estimates for Representative and Matched* Samples using SAT Mathematics data from the October 1985 Administration

		a-parameter			b-parameter			c-parameter		
Test		Repr.	Matched	Diff.	Repr.	Matched	Diff.	Repr.	Matched	Diff
Z	Mean	.951	.924	.027	-.109	-.030	-.080	.126	.127	-.001
	SD	.3020	.2883		1.219	1.187		.116	.128	
V	Mean	.886	.880	.006	.374	.354	.020	.115	.108	.007
	SD	.341	.349		1.248	1.291		.083	.086	
X	Mean	.908	.906	.002	-.099	-.113	.015	.124	.118	.007
	SD	.290	.291		1.424	1.443		.101	.105	

*Old-form sample that took Test Z was matched to the new form sample that took Test X on the basis of scores on anchor test V.

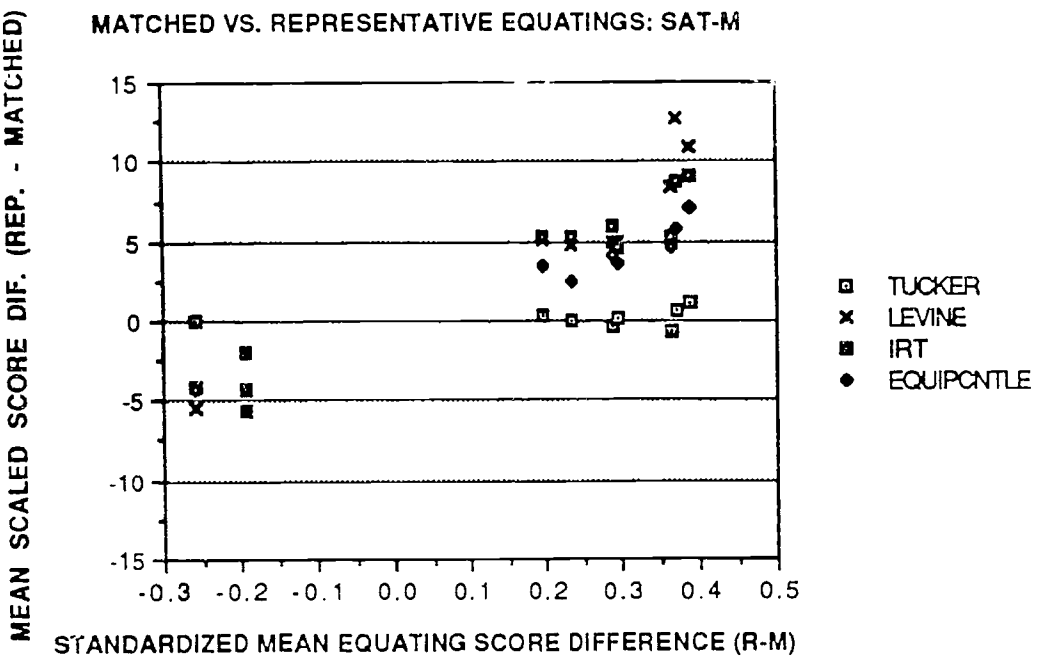
	TOTAL TESTS			ANCHOR TESTS	
	X	Y	Z	W	V
$Sample_{11}$	+			+	
$Sample_{21}$	+				+
$Sample_{12}$		+		+	
$Sample_{13}$			+		+

$Sample_{ij}$ = sample i from population j.

Note: $Sample_{11}$ and $Sample_{21}$ are random samples from the same population.
$Sample_{11}$ and $Sample_{12}$ are similar in ability.
$Sample_{21}$ and $Sample_{13}$ are dissimilar in ability.

Figure 1. Data collection design for equating the SAT

(a)



(b)

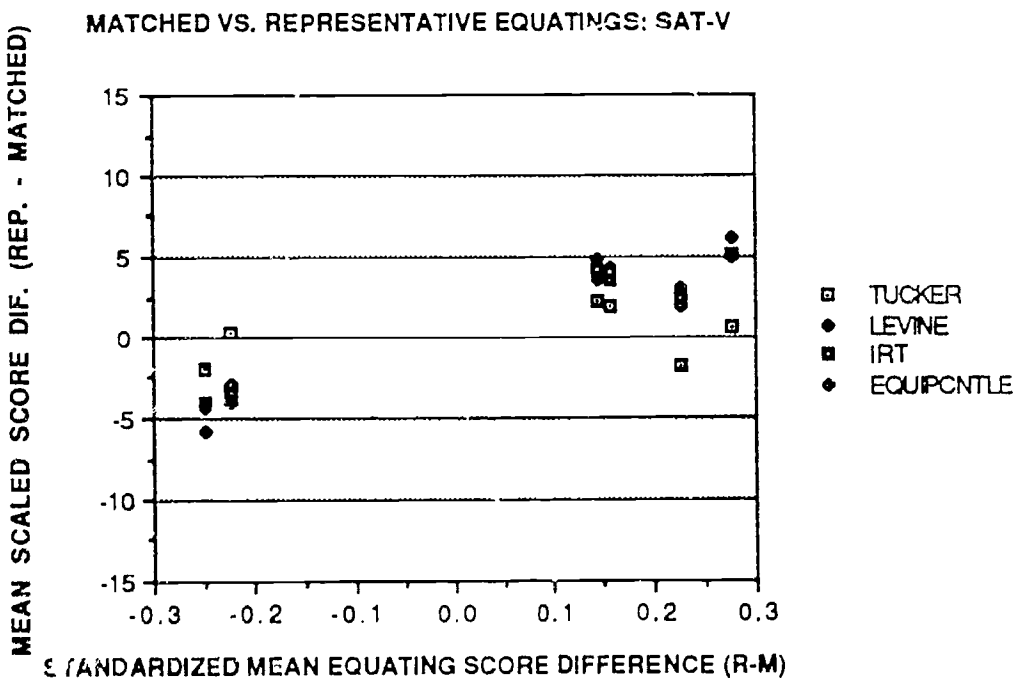


Figure 2: Matched versus representative sample equatings.